Exploring the development of frailty in older adults

Raghul Sekar
201484484

Supervisors: Professor Robert West, Silviya Nikolova, and Farag Shuweihdi

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

January 2021

Abstract

The care of elderly persons who are frail consumes a significant amount of social and health-care resources. That is, persons who are at a high risk of negative consequences including failure and hospitalization. As a result, it's essential to consider the development of frailty. Frailty development may be investigated using everyday activities and a person's lifestyle. To better understand this, the English Longitudinal Study of Aging, or ELSA, was established in England. This contains personal information about persons in England. This poll was started in 2002 and is still going strong today. This information will aid us in understanding the evolution of the frailty of the English people.

Survival analysis is a field of statistics that studies how long it will take for an event to occur, such as death in biological organisms. Medical researchers and data analysts mostly utilise this study to calculate the lifespans of a population. This study resolves some of the most important problems concerning frailty development, such as whether frailty worsens with age. Is there a difference between men and women when it comes to frailty? Is there a link between riches and the development of frailty? And a deeper understanding of the English people's frailty.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Frailty continues to be a serious worldwide health issue. The care of frail old people takes a large amount of social and health-care resources. That is, individuals who are at a high risk of unfavorable outcomes such as failure and hospitalization. As a result, it's critical to think about how frailty develops. In this report, Frailty development is investigated using Survival Analysis. Survival analysis has been used to data about the time to a certain event, such as death, the onset of a disease, or the relapse of a condition. The evolution of survival analysis in the 17th century began with John Graunt's creation of the first life table in 1662[Cam19]. Throughout history, survival analysis has been exclusively associated with the study of mortality rates; however, in recent decades, applications of statistical methods for survival data analysis have been expanded beyond biomedical research to include criminology, sociology, marketing, institutional research, and health insurance practise[Cam19]. 1669, an early example of survival analysis. The 1669 curve created by Christiaan Huygens shows how many people out of 100 live to be 86 years old [Wai01].

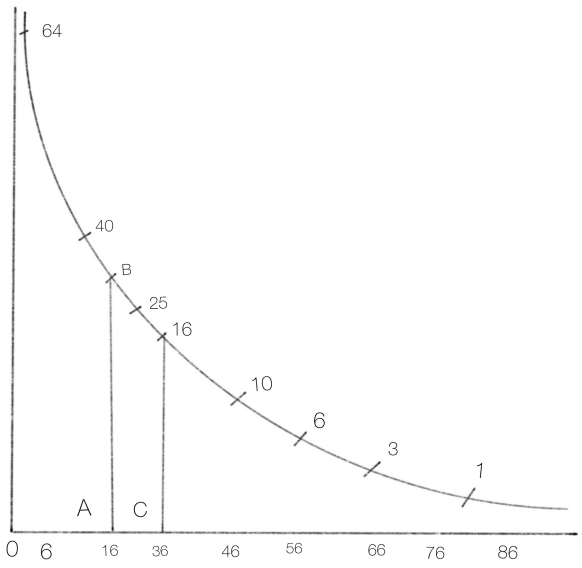

Figure 1.1: Example of a survival curve [Wai01]

From the figure

- What is the function's rough shape?

- What were the chances of the person surviving the last 20 years? passed the age of 36?
- This is what we call survival analysis. Outcome may be any binary event, not only death, as we are attempting to predict the curve.

Kaplan and Meier's contributions to calculating survival probability and hazard rates in 1958 resulted in ground-breaking advances in survival analysis. A Nelson Aalen estimator is a non-parametric approach for calculating the cumulative hazard rate. Another important addition to survival analysis was Cox's proportional hazard model, which he introduced in 1972.[Cam19] There are two elements to this semi-parametric model. The baseline hazard, which is a function of time and represents how risk evolves over time, is the first component. The second component is a time-independent exponential function of a linear combination of predictors.[Cam19]

1.1 Frailty

The most problematic indicator of population ageing is frailty. It's a condition of vulnerability to poor homeostasis resolution after a stressful event, and it's the result of a lifetime of decrease in numerous physiological systems [CAct]. Frailty is a condition of greater sensitivity to poor homeostasis resolution after a stressful event, which raises the likelihood of negative consequences such as falls, delirium, and disability [CAct]. There has been a rise in comorbid chronic illness, functional dependency, disability, lower quality of life, and greater health-care expenditures as a result of this demographic change. Patients in this group are frequently referred to be frail [SPS$^+$15].

Frailty might feel like an embarrassing or degrading title. Frailty, in fact, puts people at risk for health issues that might limit your freedom. Frailty can lead to problems with memory, thinking properly, and emotional and physical well-being. Frailty might start with minor changes that go unnoticed, and eventually worsen. As you become older, frailty becomes increasingly frequent. It can still have an impact on younger individuals. In certain situations, we can avoid or at least reduce the risk of frailty. The trick is to spot the weakness early on, before it becomes a problem.

Frailty is a high-risk state that renders a person more vulnerable to bad health consequences like falls. Frailty can be avoided or improved with the appropriate efforts, but ageing cannot be prevented. The comfortable atmosphere of their rest home was replaced with an unexpected environment for fragile elderly people. When individuals get frail, for example, they go from being cheerful and engaging to becoming non-interactive and reclusive.

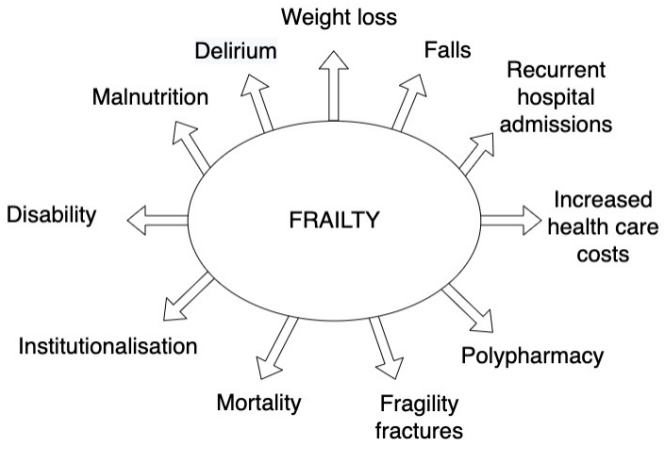

Figure 1.2: Effect of living with frailty

Someone who is frail may need to adjust their lifestyle and learn new techniques to manage day-to-day duties. This may also be said about their family. Frail people are more likely to encounter public and commercial services that aren't suited to their requirements. They are especially exposed to the impacts of low-quality healthcare and disconnected services.

People are living longer than ever before. At the moment, 16.1 percent of the European population is over the age of 65, and this percentage is predicted to grow to 22 percent by 2031 [SPS+15].

1.1.1 Models of frailty

Frailty has two broad models. The Phenotype model, for example, describes a set of patient features (unintentional weight loss, decreased muscular strength, decreased walking speed, self-reported tiredness, and low energy expenditure) that, when present, might indicate inferior outcomes. More than two of the characteristic's presence is defined as frailty (Although this model allows for the presence of fewer features and, therefore, pre-frailty, it is also possible). The Cumulative Deficit model is the second model of frailty. It is described by Rockwood in Canada as an accumulation of weaknesses (ranging from symptoms such as loss of hearing or poor mood, to indications such as tremor, to various illnesses such as dementia) that can develop with ageing and combine to increase the 'frailty index,' which in turn increases the likelihood of an unfavorable result. After a complete examination of an older person, Rockwood presented a clinical frailty scale; this suggests a rising amount of frailty, which is more in line with clinical experiences [Tur14].

Loss of skeletal muscle function (sarcopenia) is a fundamental characteristic of physical frailty as described by the phenotypic model, and there is a growing amount of research detailing the primary causes of this process. Age is the most significant risk factor, and prevalence increases with age. There is also a gender impact, with women having a greater incidence in community-dwelling elderly adults. For example, a UK research from 2010 reported a prevalence of 8.5 percent in women and 4.1 percent in men aged 65 to 74, using the phenotypic method to define frailty [Tur14].

1.2 ELSA

The English Longitudinal Study of Ageing (Elsa) is a panel study of men and women in England who are over 50 years old. It is a multidisciplinary research that includes the collecting of economic, social, psychological, cognitive, health, biological, and genetic data. It was created as a sister study to the Health and Retirement Study in the United States. The research began in 2002, and the sample has been followed up on every two years since then. Data is gathered using computer-assisted personal interviews and self-completion surveys, with four-year nurse visits for biomarker testing. There were 11,391 people in the initial sample, with ages ranging from 50 to 100. To promote international comparisons, ELSA has been harmonised with ageing studies from other countries and is connected to financial and health registry data. Researchers and analysts can access the data set immediately after it is collected [AS12].

1.2.1 ELSA Dataset

There are 18489 observations and 4349 columns in our Elsa dataset. These columns are variables that contain observational data such as health, insurance, financial and housing wealth, family structure, job history, pension, and other personal information. The first seven waves of ELSA data are stored in a single file called Harmonized ELSA. The data is saved in "fat format," which means that each observation corresponds to one responder. The person is the unit of observation. The unique identifier merge id is used to identify each individual [SB18].

The names of variables in the Harmonized ELSA Data follow a pattern. The first character denotes whether the variable pertains to the reference individual ("R"), spouse ("S"), entire household ("HH"), or financial unit household ("H"). The second character denotes which wave the variable belongs to: "1", "2", "3", "4", "5", "6", "7", or "A." The "A" stands for "all," implying that the variable isn't exclusive to any one wave. The respondent's birth date, for example, is RABDATE. The variable's concept is described by the remaining letters. Consider the following scenario: [SB18]

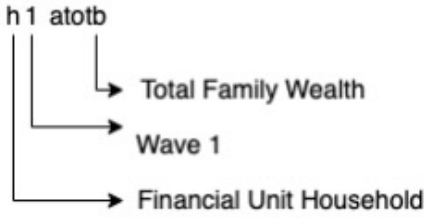

Figure 1.3: Scenario (Variable Name)

For our study, we'll just take columns like person id, biological sex, date of birth, and total family wealth and frailty score from the main dataset. As a result, these columns will be filtered from the primary dataset and stored in the gen_min dataset. This information will be utilised in our study.

Chapter 2

Methodology

2.1 Survival Analysis

Survival analysis is a field of statistics that probability that event occurs and realisation of the particular outcome after time T. Death, injury, sickness start, illness recovery (binary variables), or transition over or below the clinical threshold of a meaningful continuous variable are all examples of events. [TF16]

The outcome variable, Y, is what differentiates survival analysis. It is divided into sections. This section includes Time "T," and a random variable that indicate the event. For example, we followed someone for two and a half years until they become frail. This indicates that an event has occurred. Or someone else who we followed for a year and found out they were still living after their research finished. This indicates that the event doesn't take place. We know they lived beyond the study time, but how much longer we don't know. That result has two parts: the time(T) we tracked and the indicator of whether or not an event occurred.

$$Y = Surv(T, E)$$

Here,

Y- stands for the result variable in this case.

Surv – Function which Creates a survival object

T - The amount of time we spent following each individual

E - Event Indicator

Terms for a survival analysis:

- Time-to-event: The time it takes for a participant to reach a certain outcome after enrolling in research.
- Censoring: Subjects are said to be censored if they drop out of the study or lose to follow-up, or if the trial terminates before they die or have a significant result. For the time being, they are considered as living or disease-free.

The Kaplan-Meier curve and the Cox proportional hazards regression are the two primary methods in this survival analysis that we will utilize in this study.

2.1.1 Right Censoring

Censoring is a method of determining whether or not an event occurred. Let's take a closer look at what it means.

Right censoring, left censoring, and interval censoring are the three primary types of censoring. In this project I have used Right censoring. Because our project is based on people who are becoming frail from pre-frail stage. [Cha16] When a patient is censored, we don't know the patient's real survival period

When the research is finished, the individual has not experienced any of the events.

- During the study period, individuals are lost to follow-up.
- A participant drops out of the study.

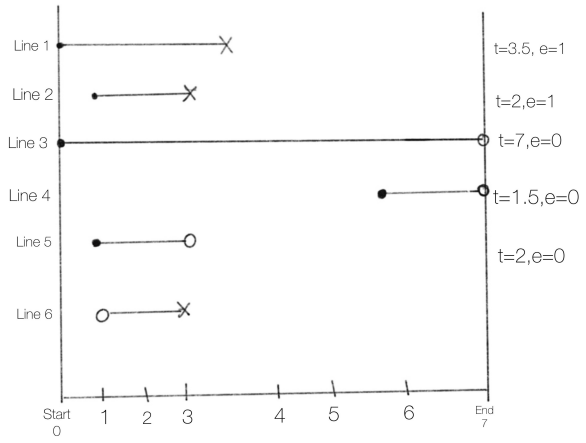

Figure 2.1: Censoring

Line 1:

Consider someone who enrolled in the research at the beginning and was tracked for three and a half years before becoming frail. So, this person's time is T=3.5 years and their event is E=1. This 1 denotes the occurrence of an event.

Line 2:

We may also have someone join our research in year one, and in year three, they become frail. This individual was tracked for two years, and the event occurred. That is T=2 years and E=1.

Line 3:

We also have someone who joined the study at the start, and they survived. Now our study ends, and they are still vulnerable. The open circle at the end of the line indicates censored. At five years, they were still vulnerable, they survived beyond seven years, but we don't know how long. This means our data is censored at that point. We are no longer following. This person here has a time T=7 years and Event E=0. Zero indicates event didn't occurs. They were censored.

Line 4: We may also have someone enter the research in year five and stay vulnerable until the end of the study. This person has time one and half year, and the event did not occur. That is T=1.5 year and E=0.

Line 5:

It might also include things such as these. Someone entered our research in the first year, but by the third year, they had dropped out. So, let's say they stopped responding calls in year three, or they stopped coming up for studies, or we have no idea what happened to them. We don't know how long they are in vulnerable stage, although they might vulnerable for a long time. So, this person is censored, has a time(T) of 2 years, and the event did not occur.

Line 6:

There are also some individuals. This of we may start following someone at year one and follow up and see how long it takes to become frail. But we don't know when they actually contracted the disease. So, suppose we are looking at the time till death after the contraction of the disease. We know they have the disease at year one, but we don't know at what point they contracted it.

The important thing here is, Line 2, 3, 4 are called Right censoring. This means the right endpoint is censored. We can see in these lines it is censored on the right side of the line.

Line 6 is referred to as "left censoring." Because the left side of the line is censored. In our report, we will not discuss left censoring.

All of the survival models we'll look at will assume that ELSA participants who drops out of the study because they are too frail and thus this will be an informative censoring. An example of informative censoring would be if a person in line 5 was in the research and was becoming significantly worse, and they knew they were going to become frail, they may stop showing up for the study. This censoring is informative.

2.1.2 Survival Function, Hazard and Hazard Ratio

Survival Function: It's a function that tells us whether or not a person will live beyond a certain period of time. A survival function is mathematically denoted as S, which is obviously a function of time. It is frequently written as S (t). This is the probability that T is greater than t. That is the probability that a survival time (T) is beyond time t [Kar16].

$$S(t) = P(T > t)$$

In other words, it is the probability of surviving T. The probability that survival time (T) is more than six years is the probability of surviving beyond six years if we are looking at survival in years.

Hazard: The abbreviation for hazard is *HAZ*. This is the probability that *T* is less than t plus delta. Given T is greater than t. It's also known as the rate of decrease of the curve or risk of dying [Kar16].

$$Hazard(HAZ) = P(T < t + \delta)|T > t$$

T $>$ t means individuals are still alive. In our case, it is a pre-frail stage.

T $<$ t means individuals are dead. In our case, individuals have become frail prior to time T.

In other words, provided that you are alive right now, this is your chance of becoming frail within next time interval. For example, if you are living at the six-year mark, what is the probability that you will become frail with next instance. When we look at the hazard ratio, it starts to make sense.

Hazard Ratio: The slope of the survival curve – a measure of how rapidly people die — is defined as hazard. When two treatments are compared, the hazard ratio is used. The Hazard Ratio will be abbreviated as HR. Let's say that the hazard of X is male divided by the hazard of X equal to female.

$$HR = \frac{HAZ, X = male}{HAZ, X = Female} = 2$$

This means, hazard for someone who is exposed relative to someone who is not exposed. For a more meaningful interpretation, suppose the Hazard ratio came out to be 2. What this tells us is that at the given instant of time, male who is exposed to the risk factor their risk of dying is double that of female. So now hazard become more meaningful to interpret.

The hazard ratio is calculated using all of the data in the survival curve, rather than just one time point. Because there is only one hazard ratio presented, it can only be evaluated if the population hazard ratio remains constant over time and any deviations are attributable to random sampling [pad16].

2.2 Kaplan-Meier Analysis

The Kaplan-Meier estimator is a non-parametric statistic that allows us to estimate the survival function. It was separately described by Edward Kaplan and Paul Meier and jointly published in the Journal of the American Statistical Association in 1958. [Sch19]

The Kaplan-Meier survival curve is the likelihood of surviving for a given amount of time when time is divided into numerous little intervals. This study is based on three assumptions. To begin, we assume that patients who are censored have the same chances of survival as those who are tracked indefinitely. Second, we assume that the survival rates for participants enrolled early and late in the research are the same. Finally, we suppose that the event occurs at the provided time. This might be problematic in some cases if the occurrence would be noticed during a routine inspection. [GM10]

The "product limit estimate" is another name for the Kaplan-Meier estimate. It entails calculating the chances of an event occurring at a specific point in time. To reach the final estimate, we multiply these sequential probabilities by any previously estimated probabilities[GM10]. The formula below is used to identify the probability of surviving at any given time:

$$S_t = \frac{Number\ of\ subjects\ at\ risk - Number\ of\ subjects\ become\ frail}{Number\ of\ subjects\ at\ risk}$$

The number of individuals surviving divided by the number of patients at risk is used to determine survival probability for each time interval. Subjects who have become frail, dropped out, or moved out are not classified as "at risk," i.e., those who have gone missing are labelled as "censored" and are not included in the denominator.

Let's have a look at the table that shows how Kaplan-Meier estimates are computed. We'll use a small dataset to help us understand things better. There are 12 people in this dataset. n=12, in other words.

TIME	#RISK	#DIED	HAZ	1-HAZ	SURV = S(t)
0	12	0	0/12	12/12	1 = 100%
2	12	1	1/12	11/12	(11/12) = 0.917
6	10	2	2/10	8/10	= 0.917(8/10) = 0.734
7	8	1	1/8	7/8	= 0.734(7/8) = 0.642
15	6	2	2/6	4/6	= 0.642(4/6) = 0.428
16	4	1	1/4	3/4	= 0.428(3/4) = 0.321
27	3	1	1/3	2/3	= 0.321(2/3) = 0.214
30	2	1	1/2	1/2	= 0.214(1/2) = 0.107
32	1	1	1/1	0/1	= 0.107(0/1) = 0

Table 2.1: Calculating Kaplan-Meier Estimates for a Dataset of 12 People

We can see from the table that there are 12 persons in the study at time zero, and there is no death at that moment. As a result, the risk of dying is 0/12. Remember, the hazard is the risk of dying in the given instant of time that is given as you are alive. As a result, the chance of dying at time zero is nil. 1 – Hazard represents the chance of not dying, and it is one for time zero. For this moment of time, the survival time S(t) is one. i.e. a hundred percent This means that everyone is alive at the beginning of time.

At t= 2, there are 12 people living and one person dead, and the hazard is 1/12. As a result, the chance of not dying is 11/12. So, survival S(t), or the chance of surviving for more than two years, is 11/12 = 0.917. This means that 91.7 percent of people will live for more than two years. At t=6, ten individuals were on the risk of dying. This is because in t=2, one person is dead, and in t = 3, someone is censored. As a result, two persons have dropped out of the research. As a result, we'll subtract two people from the overall number of people alive. At the six-year milestone, two people had died. So the risk of dying here is a 2/10, while the chance of not dying is an 8/10. To get the survival at t=6, multiply the probability of survival at t=6 by the probability of survival at last time. When you multiply 0.917 by 8/10, you get 0.734, which is the survival of time 6. The following steps will be followed in the

following steps.

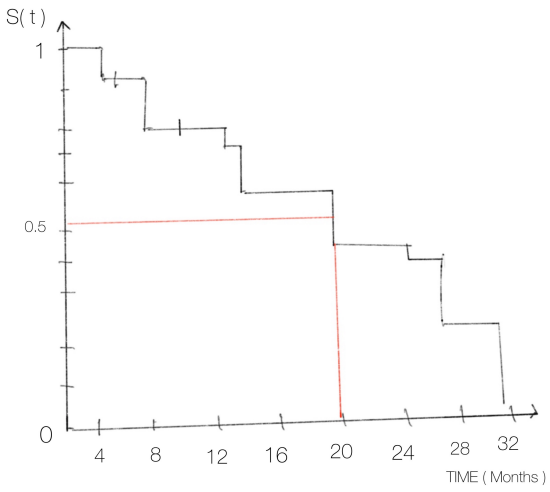

Figure 2.2: Kaplan-Meier Survival Curve for a Dataset of 12 People

From the above plot. Horizontal and vertical lines are displayed on the graph between estimated survival probabilities or estimated survival percentages on the Y-axis and time since enrollment into the research on the X-axis. The survival curve is represented as a step function, which means that even if there are some censored observations in between, the fraction of those who survive remains constant. Joining the computed points with sloping lines is wrong.

2.2.1 Advantage and Disadvantages of Kaplan-Meier

Advantage:

- Interoperability is straightforward.
- Another advantage is that you can calculate the survival S(t). As a result, you can estimate your chances of living for an extended period of time.

Disadvantage:

- We can't estimate the Hazard Ratio.
- Only a few categorical variables may be included in this K-M model. The numerical variables cannot be included in the K-M model. It can only deal with categorical X's.

2.2.2 Log-Rank Test

Survival functions for various groups of participants can be compared. We may, for example, examine the survival patterns of male and female patients in the male and female datasets. The distance between the curves may be examined in both horizontal and vertical dimensions.

The two survival curves may be statistically compared by testing the null hypothesis, which states that there is no change in survival between the two therapies. A statistical test known as the log-rank test is used to evaluate this null hypothesis. In the log-rank test, Let *E1 and E2* be the expected number of events in each group (Male and Female), whereas the total number of observed events in each group are O1 and O2 [GM10].

$$Log - rank\ test\ statistic = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

The log-rank test is used to see if the difference in survival durations between the two groups is statistically significant. However, it does not allow for the influence of additional independent factors to be tested[GM10].

2.3 Cox Proportional-Hazard Model

The Cox proportional-hazards model (Cox, 1972) is a regression model that is often used in medical research to look into the relationship between patient survival time and one or more predictor factors. It is a semi parametric model because the result distribution is uncertain even though the regression parameters (betas) are known [FES19].

$$HAZ = e^{b_0 + b_1 x_1 + \ldots + b_k x_k}$$

$$Log(HAZ) = b_0 + b_1 x_1 + \ldots + b_k x_k$$

"b_0" is a time function. It can vary, which means it can rise and fall over time. The b_0 in this case is $log(h_0(t))$. This model came with the ability to estimate $b_1, b_2, ..., b_k$ without explicitly specifying the baseline function. We can't estimate this survival function using the proportional hazards model since the baseline function isn't specified. We do not obtain an estimate of the intercept if we do not give this. Let's say the intercept can fluctuate up and down over time, and we'll describe how it varies up and down. We can't estimate the hazard to estimate this survival function since we don't have the baseline function b_0. However, it does let us to estimate the other coefficients, and if all we want to know is the Hazard ratio, the Cox proportional model is a good fit. If our objective is to be more predictive, such as determining the probability of surviving beyond particular line points, the Cox proportional model will not be able to help us. The Cox proportional model is excellent for estimating hazard ratios in effect size models.

The hazard will change over time in the Cox proportional hazard model. However, the hazard between the groups is a proposal, or to put it another way, the hazard ratio remains constant. Consider the example of the hazard ratio for biological sex. So the hazard for the males relative to the hazard for the females and what the cox proportional hazard model assumes is that this hazard ratio is constant over time.

$$HR = HAZ(males)/HAZ(females)$$

Assume the Hazard Ratio is two. That indicates that a male is twice as likely as a female to die at any given time. At any point in time, this will be doubled.

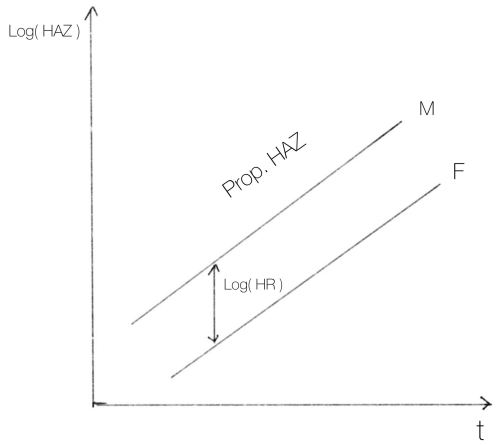

Figure 2.3: The graph for demonstrates proportional hazards.

Log(HAZ) is the function of X variables, as shown in the diagram above. In order to understand better, In the diagram is shown as two straight lines. Lines can be in different shapes. So, we fit male and female lines, and the distance between male and female lines is the log hazard ratio. This one demonstrates proportional hazards. Over time, the hazards for males and females are proportional. The hazard ratio is always the same. The term proportional hazards refer to a situation in which the risks are proportional to,

$$HR = HAZ(males)/HAZ(females) = 0.5/0.25 = 0.6/0.3 = 0.9/0.45 = 2$$

The male has a hazard of 0.5, whereas the female has a hazard of 0.25, as seen above. HR is therefore 2. As time passes, if the male HAZ is 0.6 and the female HAZ is 0.3, the HR is 2. It states that no matter what the Haz is, HR will remain constant over time.

From then, as we saw at the commencement of the Cox proportional hazard model,

$$Log(HAZ) = Log(h_0(t)) + b_1x_1 + b_2x_2 + ... + b_kx_k$$
$$HAZ = h_0(t) * e^{b_1x_1+b_2x_2+...+b_kx_k}$$

The log baseline function $log(h_0(t)$ is a time-dependent function. It permits the reference group's hazard to grow or shrink over time. For the reference group, this will operate as an intercept term.

This type of significant breakthrough with the Cox proportional hazard model is that he devised a method for estimating $b_1, b_2, ..., b_k$. Without needing to give the baseline hazard function, the coefficient may be calculated.

2.3.1 Cox Proportional-Hazard Assumptions

- Non-Informative censoring — This means that the censored observation has nothing to do with the likelihood of an event occurring. To put it another way, just because someone was censored doesn't indicate they were more or less likely to die. We assume it's a random event.
- Survival times(t) are independent – that is, person one's survival time is unaffected by person two's or three's survival time.
- Hazards are proportional. The primary assumption of the Cox proportional model is this. Another way to express it is that the hazard ratio remains constant throughout time.
- We assume log(HAZ) is a linear function of the X numerical variables as a linear function of the X numerical variables. We don't need to worry about linearity if we're looking at categorical variables like biological male or female. Residual plots can be used to verify this.
- This may be called assumption or fact about the Cox Proportional hazard model. Baseline hazard $h_0(t)$ is unspecified.

2.3.2 Solution If Assumptions Fails

- One option is to satisfy the variable if the hazards are not proportionate over time or the hazard ratio is not constant. We satisfy the biological sex and fit the model for men and the separate model for females if the Hazard ratio for males and females did not remain constant throughout time.
- If the connection between the Log hazard ratio and the numerical axis isn't linear, we can use a transformation such a log of X, root of X, or another. Polynomials can be included. X and X2, as well as others. We can attempt classifying X. As a result, numerical data is used to create categories.

2.3.3 Advantage and Disadvantages of Cox Proportional-Hazard Model

Advantage:

- The hazard ratio can be calculated. We can't estimate the hazard ratio in the K-M model.

Disadvantage:

- The survival function cannot be estimated since the baseline hazard $h_0(t)$ is unknown.

Chapter 3

Dataset

We have a dataset called gen min for now. This dataset contains 18489 participants. The length of the period in this dataset is from 2002 to 2015, which is from wave 1 to wave 7. The primary ELSA dataset has been filtered to create this dataset. Gender, age of the person at the time of the survey, wealth at each wave, and frailty score of individuals at each wave are all columns in gen min. The names and details of the columns are listed below.

Column Name	Column Details
idauniq	Unique ID of the individuals
ragender	Biological Sex
rabyear	Date of Birth
r1agey ,r2agey, r3agey, r4agey, r5agey, r6agey, r7agey	Age at Interview (In Years)
fraill1, fraill2, fraill3, fraill4, fraill5, fraill6 ,fraill7	Frailty score
h1atotb, h2atotb, h3atotb, h4atotb, h5atotb, h6atotb, h7atotb	Total Family Wealth

Table 3.1: Columns Details

3.1 Missing Data

The most common issue in data analysis is missing data. The is.na() function in r is used to determine the number of missing values in each column of the table. The number of null values in each column is shown below

idauniq	ragender	rabyear	r1agey	r2agey	r3agey	r4agey	r5agey
0	0	1	6390	9057	8719	7439	8215
r6agey	r7agey	fraill1	fraill2	fraill3	fraill4	fraill5	fraill6
7888	8823	6580	9059	8783	7448	8224	7893
fraill7	h1atotb	h2atotb	h3atotb	h4atotb	h5atotb	h6atotb	h7atotb
8825	6598	9179	8966	7734	8446	8127	9072

Table 3.2: Number of Null values in each columns

All of the columns, apart from iduniqe and ragender, have null values, as seen above. Our investigation will be limited by these Null values. We must fill the missing values in each column to fix the difficulties that are caused by missing data.

We will use the date of birth of each participant and subtract it from the year of the survey to fill in the missing values for the age of the individual at the time of the survey. This will provide the age at the time of the survey. We'll now utilise the mean of wealth across all waves to fill in the missing values in the total wealth of people in each wave, replacing the null with the mean. We will be able to replace the null values in the wealth columns as a result of this. We can't fill null values in the frailty score because the person is dropping out from the studies. This will be handled with throughout the censoring process. As a result, all of the necessary column's null values are filled.

3.2 Censoring

We require columns like status, time, the age of the individual at the time of becoming vulnerable and frail, and the total wealth of the individual at the time of becoming vulnerable and frail for survival analysis. These columns may be found in the gen min database. We're utilising the right censoring for it. In the last part, we learned what right censorship is. We're employing the same strategy.

It is said to be at a vulnerable stage if the frail score is between 0.14 and 0.24. It is considered to be frail if the frail score is more than 0.24 [ABM01]. There are two values in the status column: 1 and 2. 1 denotes a person who has reached the vulnerable stage, and 2 denotes an event that occurred, that is person reached the frail stage. To acquire the values for a status column, we'll start by looking for the vulnerable stage from the wave's beginning. If they entered this stage, the status column is set to 1, and they will only check for the frail stage in subsequent waves if they are in the vulnerable stage. If they reach the frail stage, 1 will be replaced by 2 in the status column. The column has been filled as a result of this status. The time column represents the amount of time it takes for an individual to become frail after entering the vulnerable stage. Time will be zero if an individual does not encounter frail after being vulnerable. In new columns, the age and wealth of the individual at the time of vulnerability and frailty will be noted. Only the needed columns will be filtered and stored in a new dataset named cox_fd.

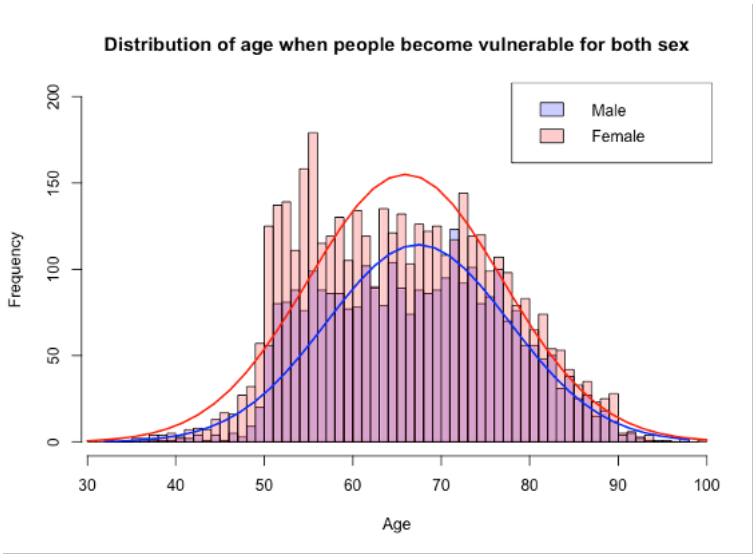

Figure 3.1: Distribution of age when people become vulnerable for both sex

The distribution of age of both men and females at the age when they become vulnerable is shown in the graph above. We may conclude from the plot that the male has a lower frequency than the female. This indicates that the female count is high in this dataset vulnerable stage. Also, we can observe that the majority of persons, both male and female, are between the ages of 55 and 80. This means that the chances of being vulnerable are high in this range.

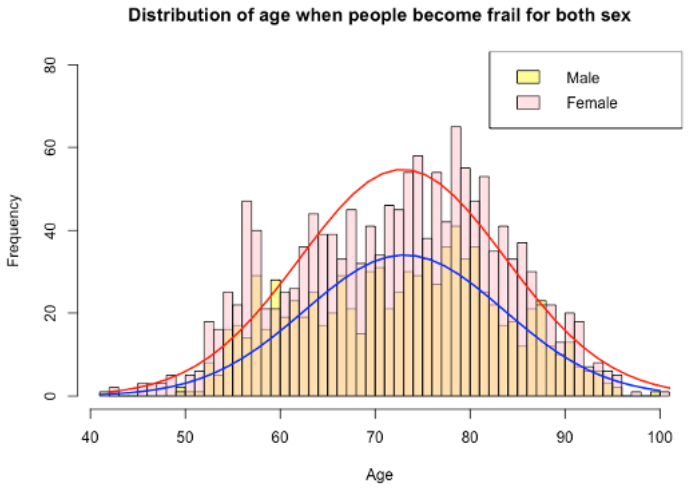

Figure 3.2: Distribution of age when people become frail for both sex

The above figure, like the one before it, shows the age distribution of both males and females at the time when they become frail. We can observe that the majority of persons, both male and female, are between the ages of 60 and 85. This suggests that if you're in this range, you're at a significant risk of getting frail.

3.3 Adding Required Variables

We're also compiling a new set of variables for age and wealth. These variables will be utilised in the future for K-M models. First, I'm going to make an age_vc variable. This will have three catalogue numbers: 1, 2, and 3. Where 1 denotes persons between the ages of 0 and 65, 2 denotes people between the ages of 65 and 75, and 3 denotes those above 75. In the same way, I'm going to make a new variable named wealth_vc. It also has three catalogue numbers: 1,2,3. Where 1 denotes persons with total wealth of less than 60,000 pounds, 2 denotes people with total wealth of between 60,000 and 300,000 pounds, and 3 denotes people with total wealth of more than 300,000 pounds. The cut function on age and wealth variables is used to construct these variables. The datatypes of age_vc and wealth_vc are transformed to factors once they are created. The cox_fd dataset contains all of these variables. This is the final dataset we've prepared.

Column Name	Column Details
status	Frailty Status
time	Time taken to become frail from vulnerable stage
age_v	Numerical age when individuals become vulnerable
wealth_v	Numerical wealth when individuals become vulnerable
age_f	Numerical age when individuals become frail
wealth_f	Numerical wealth when individuals become frail
age_vc	catalog age
wealth_vc	catalog wealth

Table 3.3: Columns Details

Chapter 4

Data Analysis

The cox_fd dataset that we produced in Data Preparation will be used. Our dataset contains all of the columns that we will need for this study. Let's have a look at cox_fd:

This will display the data:

idauniq	ragender	status	time	age_v	wealth_v	age_vc	wealth_vc
100007	1	1	8	58	229580	1	2
100016	2	1	2	54	135700	1	2
100018	1	1	2	50	15000	1	1
100024	1	1	2	79	183088	3	2
100025	1	1	8	75	534100	2	3

Table 4.1: Cox_fd Dataset

This table includes gender, time, status, numerical age, and wealth before and after frail, as well as catalogue age and wealth.

4.1 Kaplan-Meier Analysis

Let us say the "Kaplan-Meier" technique to calculate the survival curve and choose the best model that fits to see how age, gender, and wealth impact the development of frailty. To begin with, let's start with the simple model.

To start we need to load the survival library in order to use these survival commands.

4.1.1 K-M model: without x

To compute a kaplan-Meier survival estimate, use the function survfit(). Its key points are as follows a Surv() - created survival object, as well as the data set containing the variables. Let's compute a Kaplan-Meier survival estimate without any X variable. Without X variables mean without any covariance like age, wealth, sex. We'll use the surv fit command to fit the model, and then save it in the km.model object. Then we use Surv() to define the survival time or our Y variable. In the parenthesis, we must include both the time when the individual's end was tracked as well as the indicator of whether the event happened, i.e., is they became frail or were censored. If there are no x variables, Tilda one is used. This is how we tell R that we're only interested in predicting survival without utilising any specific X variables.

Now, let's use km.model to get the following results:

```
    n events median 0.95LCL 0.95UCL
 7219   2383     10      10      10
```

The total number of people involved in the research is 7219. In that case, 2,383 events happened, and 4,836 censored observations were made, indicating that individual events did not occur. We can figure this out by subtracting the total number of persons from the number of events that occur. After then, the median survival time will be ten years. That is, half of the persons lived for more than ten years while the other half did not. It also provides a 95% confidence interval for the median. As a result, we have a 95% confidence that the median survial is 10.

Let's ask for the summary of the km.model using summary() function

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   5798    1145    0.803 0.00523        0.792        0.813
    4   3783     529    0.690 0.00638        0.678        0.703
    6   2574     327    0.603 0.00718        0.589        0.617
    8   1609     223    0.519 0.00807        0.503        0.535
   10    928     103    0.461 0.00895        0.444        0.479
   12    434      56    0.402 0.01077        0.381        0.424
```

summary() provides a data frame with the following columns:

- time: the points in time(in years) where the curve takes a stride.
- n.risk: the number of people who are at risk at time t.
- n.event: the total number of events at time t.
- n.censor: the number of occurrences that have been suppressed.
- surv: a probability estimate of survival.
- standard error of survival (std.err)
- upper: the confidence interval's upper limit
- lower: the confidence interval's lower end

The Kaplan-Meier survival function will be defined by this collection of steps. There were 5798 persons at risk at time two years, and 1145 people became frail. The chance of surviving for more than two years is 80.3 percent. We are 95% confident that we have a 79.2 percent to 81.3 percent probability of living for more than two years. At time four, 3783 people are at risk, 529 people become frail, and their chances of living beyond four years are 69 percent. We are 95 percent confident that somewhere between 67.8% and 70.3 percent of people will live for more than four years. This continues until time 12, when the number of persons at risk is 434 and 56 people are censored, resulting in a survival rate of 40.2 percent after 12 years. This is a table comparable to the one we're making in section 4.2 using our hands.

Plot of this model display the fit of the Kaplan-Meier model.

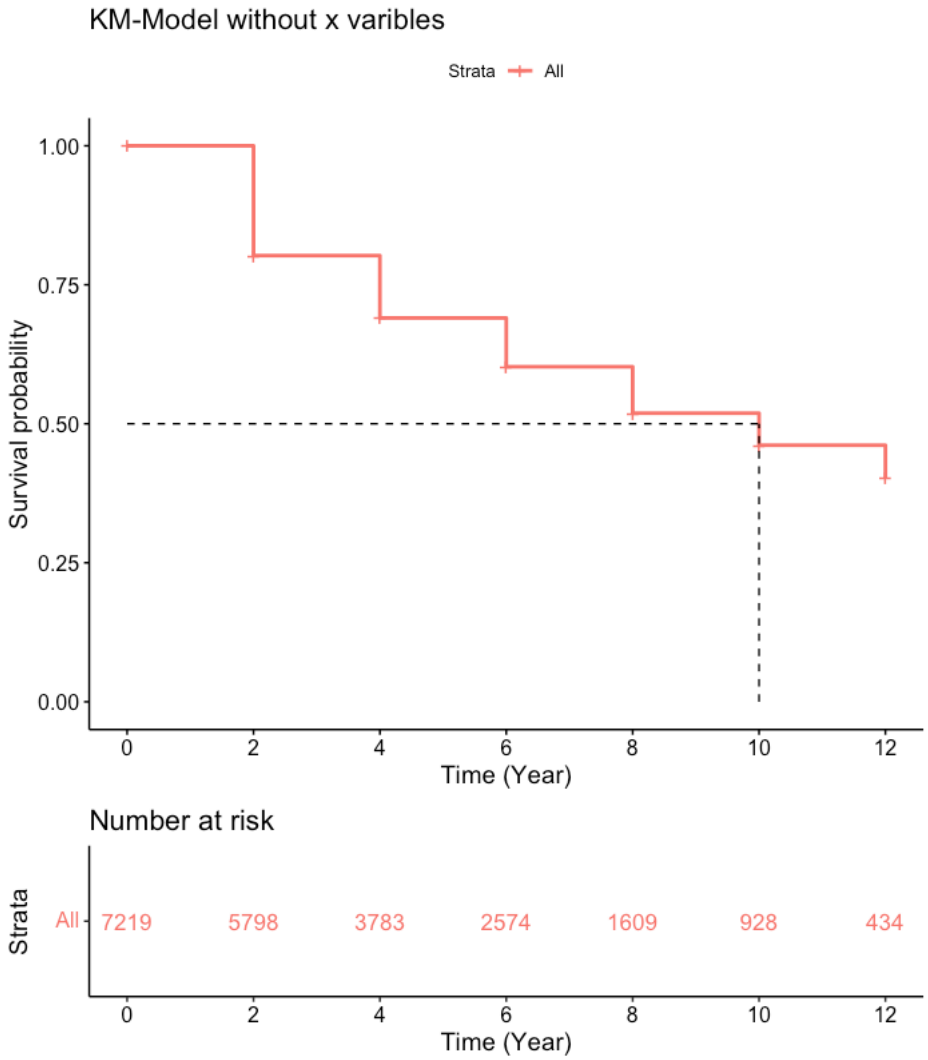

Figure 4.1: KM-Model without x variables

The plot for the km.model will be this. The median survival line, which crosses the survival function at 10 years, is shown as a dotted line in 50 percent at survival probability. That is, it reveals that this model has a ten-year median survival rate. The tick marks may be seen in the surviving line. There is a censor observation where that mark appears. From the figure, we can see that the observation is censored at periods 0, 2, 4, and so on, and we can summarize the survival probability of each year. Like at time 6, the probability of surviving is 60%.

4.1.2 K-M model with Gender

Let's do a Kaplan-Meier survival estimate with the x variable gender now. We'll fit the model with the surv_fit command, then save it in the km.model1 object. Surv () will be the Y variable, while ragender will be the X variable. This variable contains 1 as male and 2 as female, and data for the surv_fit will be cox_fd as seen before everything is same.

Lets return km.model1 this will give:

```
              n events median 0.95LCL 0.95UCL
ragender=1 2964    896     10      10      10
ragender=2 4255   1487     10       8      10
```

From the above result, we can see that the total number of males is 2964. In that number of events that occurred for males is 896, and the censored observation will be 2068. The median survival time for males is 10. Also, it returns the 95% confidence interval for the median. That is, we are 95% confident that the median survival is 10. The number of females who become frail is 1487 out of 4255. The median survival time is 10. So half of the females survived beyond ten years, half did not. Also, we are 95% confident that the median survival for the female is somewhere between 8 and 10.

Now let's get the brife summary of the model:

```
               ragender=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   2294     418    0.818 0.00806        0.802        0.834
    4   1497     200    0.709 0.01002        0.689        0.728
    6    996     129    0.617 0.01153        0.595        0.640
    8    605      80    0.535 0.01313        0.510        0.562
   10    316      41    0.466 0.01526        0.437        0.497
   12    154      28    0.381 0.01912        0.345        0.420

               ragender=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   3504     727    0.793 0.00685        0.779        0.806
    4   2286     329    0.678 0.00826        0.662        0.695
    6   1578     198    0.593 0.00918        0.576        0.612
    8   1004     143    0.509 0.01023        0.489        0.529
   10    612      62    0.457 0.01110        0.436        0.480
   12    280      28    0.412 0.01292        0.387        0.438
```

From summarising the km.model1, we get two data frames one is for males, and another one is for females. This data frame is the step that defines the KM survival function. Ragender 1 refers to male. At time two there 2294 male individuals are at risk, in that 418 become frail. So the probability of survival beyond two years is 81.8%, and we have 95% confidence that somewhere between 80.2% up to 83.4% chance of surviving beyond two years. Ragender 2 refers to female. A summary of females shows at time two, there are 3504 people who are at risk, and the number of events that occurs is 727, and the probability of survival beyond two years is 79.3%. It has a lower 95% confidence interval of 77.9% and an upper 95% confidence interval of 80.6%. So this is a general summary of the K-M model with biological sex.

Plot for this model is called using ggsurvplot.

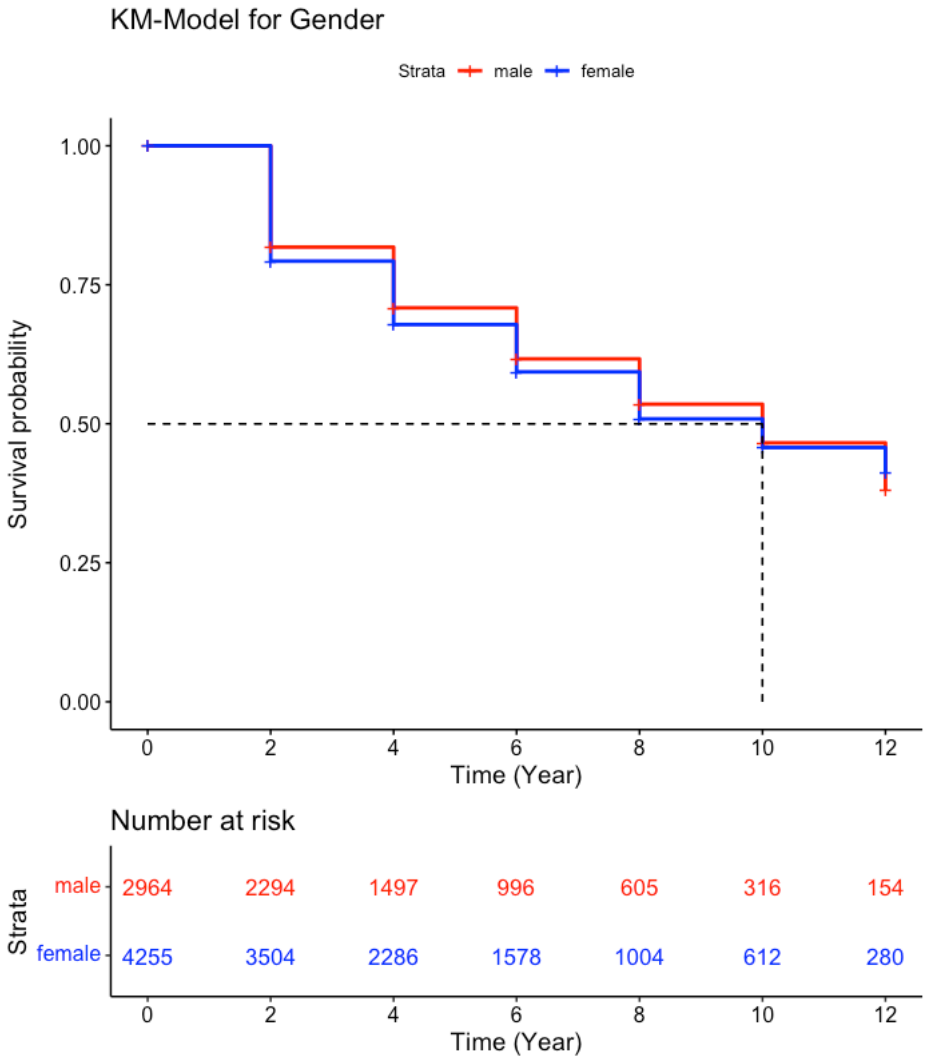

Figure 4.2: KM-Model for Gender

In this plot, the X-axis refers to the time in years, and the Y-axis refers to survival probability. In this plot, there are two survival function lines. The Red line refers to the male, and the blue line is for the female. From the plot, we can say the male has a little high chance of surviving than the female. And the median survival time for both males and females is ten years. So, half of the males and half of females in the study have survived beyond ten years, and Half did not.

Log-Rank test for km.model1: In order to check the assumption, that is, do we think two survival functions are statistically different? In other words, do we think that survival differs depending on if someone is in these two groups? We are going to do a formal test for this. The log-rank test is a way of testing that is there is a significant difference between the two survival curves. This test has a null hypothesis that the survival of two groups is the same and the alternative hypothesis that survival of two groups is not the same. To check this in this model survdiff function is used. In parentheses, we must specify the model that was fit.

when we will return the following result:

```
              N Observed Expected (O-E)^2/E (O-E)^2/V
ragender=1 2964      896      928      1.08      2.12
ragender=2 4255     1487     1455      0.69      2.12

 Chisq= 2.1  on 1 degrees of freedom, p= 0.1
```

In the end, we can see the test statistic is 2.1, and the p-value is 0.1. based on the large p-value, It fails to reject the null hypothesis. It is not proven that the male survival curve and female survival curve are different.

4.1.3 K-M model: with Age

K-M model with three age catalog: Let's now use the age catalogue variable to estimate the Kaplan-Meier survival. Because we can't utilise the numerical variable in the Kaplan-Meier model, I'm utilising the catalogue variable here. Surv_fit is used to fit the model, and age_vc comprises three age categories. 1 refers to those under the age of 65, 2 to people between the ages of 65 and 75, and 3 to people above the age of 75 in age_vc. Model is stored in km.model2.

Returning km.model2 will give us the following result:

```
            n events median 0.95LCL 0.95UCL
age_vc=1 3450    911     NA      NA      NA
age_vc=2 2131    787      8       8       8
age_vc=3 1638    685      6       6       6
```

We can see from the results that there are 3450 persons under the age of 65 who are at risk, and 911 people become frail in them. There are 2131 persons between the ages of 65 and 75 who are at risk, and 787 of them will become frail. That group's median survival time is eight years. There are 1638 adults over the age of 75 who are at risk. 685 of them grow frail. And median survival time is six years.

Summary of km.model2:

```
               age_vc=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   2842     443    0.844 0.00680        0.831        0.858
    4   1988     196    0.761 0.00834        0.745        0.777
    6   1456     112    0.702 0.00935        0.684        0.721
    8   1002      97    0.634 0.01069        0.614        0.656
   10    599      40    0.592 0.01189        0.569        0.616
   12    306      23    0.548 0.01416        0.520        0.576

               age_vc=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   1732     345    0.801  0.0096        0.782        0.820
    4   1143     178    0.676  0.0118        0.653        0.700
    6    762     127    0.563  0.0134        0.538        0.590
    8    430      78    0.461  0.0152        0.432        0.492
   10    252      35    0.397  0.0165        0.366        0.431
   12    107      24    0.308  0.0205        0.270        0.351

               age_vc=3
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	1224	357	0.708	0.0130	0.6833	0.734
4	652	155	0.540	0.0154	0.5106	0.571
6	356	88	0.406	0.0169	0.3746	0.441
8	177	48	0.296	0.0184	0.2624	0.334
10	77	28	0.189	0.0200	0.1531	0.232
12	21	9	0.108	0.0233	0.0704	0.165

As a consequence of the summary, we'll obtain three data frames for each group. At time two for age_vc 1, 2842 people are at risk, and 443 persons become frail. The chance of living for more than two years is 84.4 percent. We have a 95% confidence level that we will survive beyond two years, with a probability of 83.1 percent to 85.8%. People between the ages of 65 and 75, on the other hand, have a lower chance of living than those under the age of 65. That is, age_vc 2 has a lower chance of living at 80.1 percent than age_vc 1. By seeing age_vc 3, it is less than the other group. Whenever a result, as people get older, their chances of survival decline. This demonstrates that age has a significant impact on survival estimates.

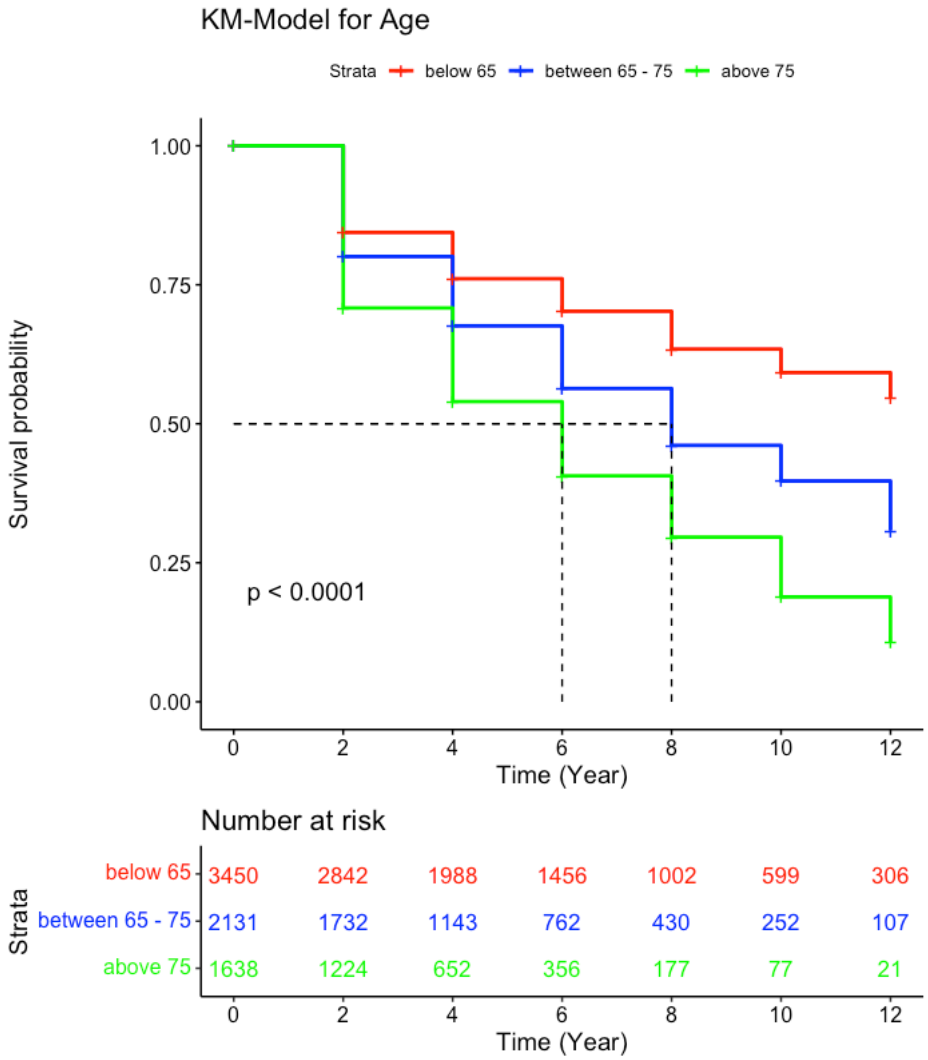

Figure 4.3: KM-Model for Age

We can achieve the plot displayed above by showing the survival curve for km.model2. The red line is at the top, followed by the blue and green lines, which are at the bottom. Younger individuals, have a better chance of living than older people. People above the age of 75 have a median survival time of eight years, while those between the ages of 65 and 75 have a median survival time of six years.

Log-Rank test: In order to check survival functions are statistically different. We are getting the result using survdiff function.

```
            N Observed Expected (O-E)^2/E (O-E)^2/V
age_vc=1 3450      911     1269     101.0     261.8
age_vc=2 2131      787      700      10.8      18.3
age_vc=3 1638      685      414     177.6     262.4

 Chisq= 354  on 2 degrees of freedom, p= <2e-16
```

In the end, we can see the test statistic is 354, and the p-value is less then 0.05. based on the small p-value, It reject the null hypothesis. It is proven that the survival curves are different.

K-M model with four age catalog: As we've seen before, as people become older, their chances of surviving drop. To double-check, I divided the age groups into four categories and fitted the model. The model's plot demonstrates this. We can observe that the elderly have a lower survival rate than the young.

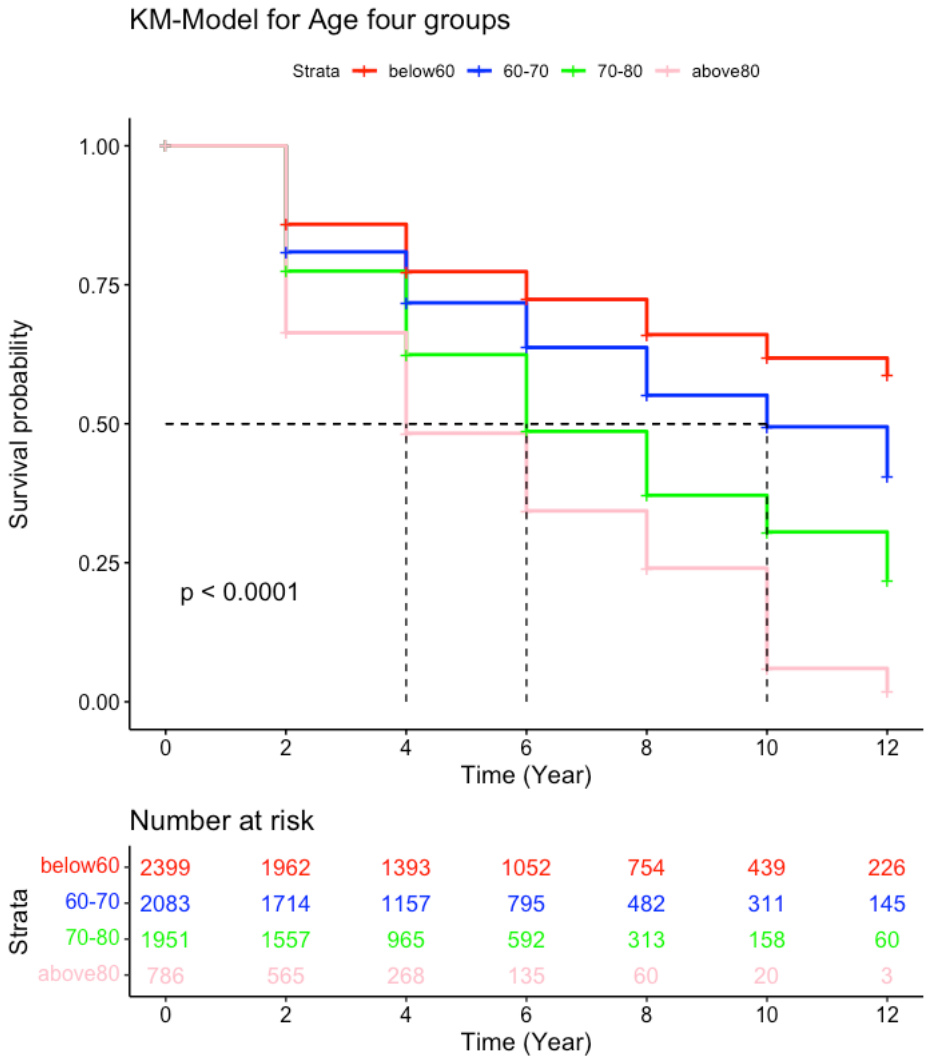

Figure 4.4: KM-Model for Age four groups

K-M model: with Wealth

K-M model with three wealth catalog: We're utilising wealth to fit the K-M model. This variable, Wealth, is a catalogue variable with three groupings. Wealth_vc 1 is a group of persons with a net worth of less than 60,000, wealth_vc 2 is a group of people with a net worth of 60,000 to 300,000 , and wealth_vc 3 is a group of people with a net worth of more than 300,000. We're fitting the model and saving it in the km.model3 object.

```
                n events median 0.95LCL 0.95UCL
wealth_vc=1 1893    761      6       6       8
wealth_vc=2 3649   1218     10      10      10
wealth_vc=3 1677    404     NA      NA      NA
```

Returning km.model3 gives us the conclusion that 1893 persons are at risk for wealth vc 1. 761 individuals grow frail as a result of it. And the median survival time is six years. Then there's wealth vc 2, which puts 3649 individuals in risk. It causes 1218 people to become frail, with a ten-year median survival rate. Finally, 1677 persons are at danger in wealth vc 3, with 404 becoming frail.

Summary:

```
               wealth_vc=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   1469     363    0.753  0.0113        0.731        0.775
    4    885     179    0.601  0.0136        0.575        0.628
    6    572     105    0.490  0.0147        0.462        0.520
    8    362      67    0.400  0.0156        0.370        0.431
   10    202      28    0.344  0.0166        0.313        0.378
   12    106      19    0.283  0.0187        0.248        0.322

               wealth_vc=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   2987     581    0.805 0.00724        0.791        0.820
    4   1981     268    0.697 0.00881        0.679        0.714
    6   1365     163    0.613 0.00987        0.594        0.633
    8    863     113    0.533 0.01110        0.512        0.555
   10    509      60    0.470 0.01241        0.447        0.495
   12    252      33    0.409 0.01470        0.381        0.438

               wealth_vc=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   1342     201    0.850 0.00974        0.831        0.870
    4    917      82    0.774 0.01195        0.751        0.798
    6    637      59    0.702 0.01403        0.676        0.731
    8    384      43    0.624 0.01682        0.592        0.658
   10    217      15    0.581 0.01899        0.545        0.619
   12     76       4    0.550 0.02334        0.506        0.598
```

By summarising km.model3, we will get three data frames. At time 2, wealth_vc 1 has 1469 people at risk. In it, 363 people become frail. Surviving probability of it is 75.3%. Where the surviving probability of wealth_vc 2 and wealth_vc 3 is 80.5% and 85%. As a result of this, we learned that as money rises, so does the likelihood of survival. As a result, money has a noticeable influence on survival projections.

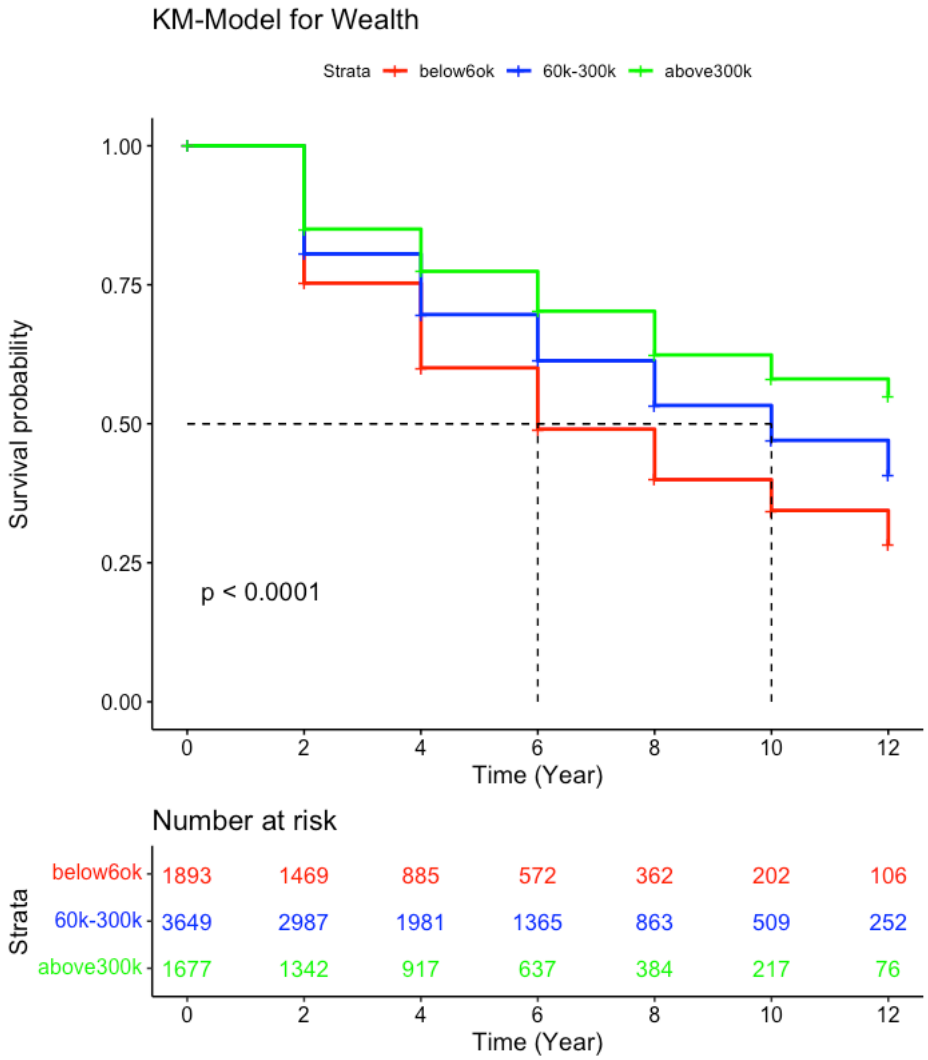

Figure 4.5: KM-Model for Wealth

By displaying the survival curve for km.model3 , we may create the figure shown above. The top line is the green line, which is followed by the blue and red lines at the bottom. This means that those who have more money have a higher chance of surviving than those who have less. The median survival time for people with less than 60,000 pounds and those with 60,000 to 300,000 pounds is six years and 10 years, respectively.

Log-Rank test: In order to see if survival functions are different statistically. The result will be obtained using the survdiff function.

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
wealth_vc=1 1893      761      573    61.842     97.60
wealth_vc=2 3649     1218     1249     0.766      1.93
wealth_vc=3 1677      404      561    44.076     68.98

 Chisq= 128  on 2 degrees of freedom, p= <2e-16
```

We can see that the test statistic is 128, and the p-value is less than 0.05 in the conclusion. It rejects the null hypothesis due to the tiny p-value. It is proven that the survival curves are different.

K-M model with four wealth As we've seen in the past, as people's wealth rises, so do their chances of survival. I separated the wealth groups into four categories and fitted the model to double-check. The plot of the model illustrates this. People with less money have a lower survival rate than those with more money, as can be shown.

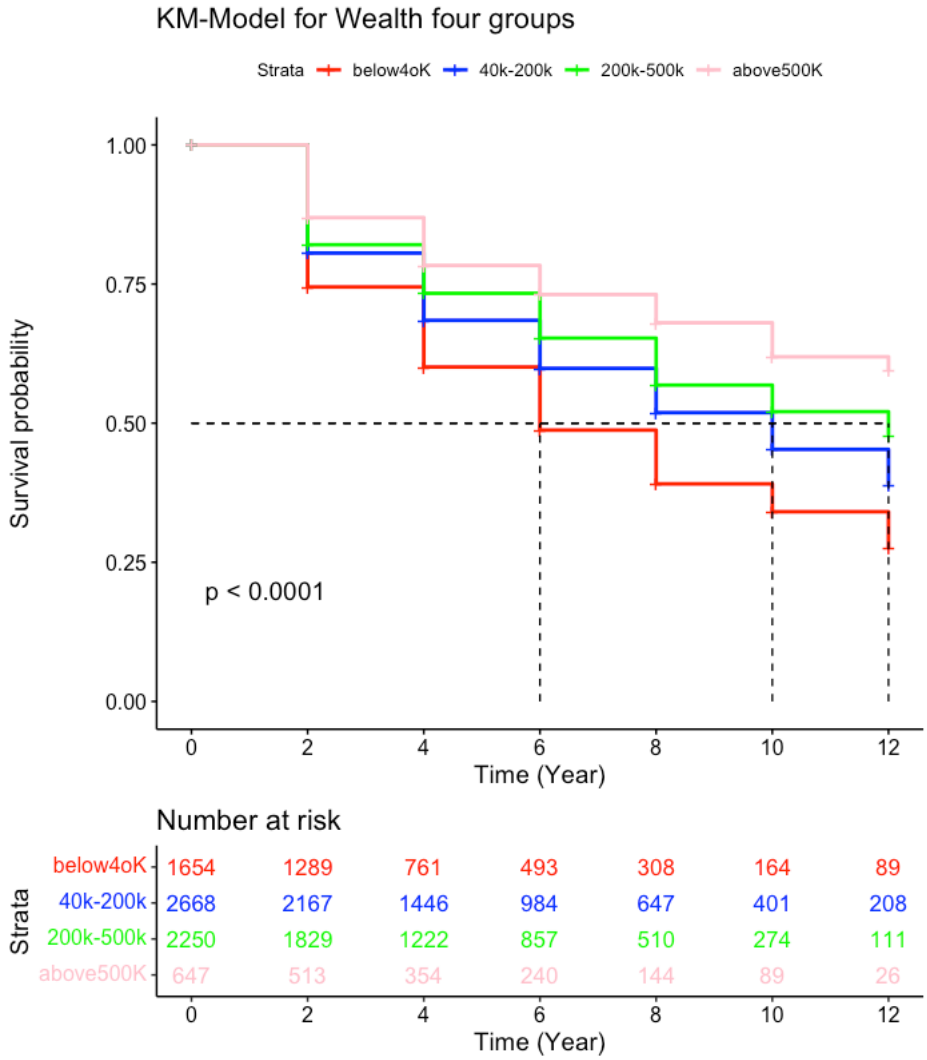

Figure 4.6: KM-Model for Wealth four groups

4.1.4 K-M model: with Age and Wealth

Let's use the various x variables to get a Kaplan-Meier survival estimate. We'll use the age and wealth variables to fit the model with the surv_fit command, then store it in the km.model4 object. The variables of age and wealth are both catalogue variables.

By returning km.model4 we will get following result:

```
                          n events median 0.95LCL 0.95UCL
age_vc=1, wealth_vc=1   903    324      8       8      10
age_vc=1, wealth_vc=2 1778    461     NA      NA      NA
age_vc=1, wealth_vc=3   769    126     NA      NA      NA
age_vc=2, wealth_vc=1   515    217      6       6       8
age_vc=2, wealth_vc=2 1072    415      8       8      10
age_vc=2, wealth_vc=3   544    155     10       8      NA
age_vc=3, wealth_vc=1   475    220      4       4       6
age_vc=3, wealth_vc=2   799    342      6       4       6
age_vc=3, wealth_vc=3   364    123      8       6       8
```

For any combination of age and wealth, the above result will tell us the number of people at risk, the number of events that occur, and the median survival time. We can do so for age_vc 1 and wealth_vc 1, and 903 persons are at danger, with 324 being frail. That is, people with age below 65 and wealth below 60,000 pounds there are 903 are at risk, and in it 324 became frail. The outcome of the other combination is shown in the same way.

Now let's get the brife summary of the km.model4:

```
               age_vc=1, wealth_vc=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    714     147    0.794  0.0151        0.765        0.824
    4    464      82    0.654  0.0188        0.618        0.692
    6    310      33    0.584  0.0203        0.546        0.625
    8    218      40    0.477  0.0226        0.435        0.523
   10    118      13    0.424  0.0243        0.379        0.475
   12     70       9    0.370  0.0272        0.320        0.427

               age_vc=1, wealth_vc=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2   1487     227    0.847 0.00933        0.829        0.866
    4   1063      95    0.772 0.01127        0.750        0.794
    6    789      60    0.713 0.01271        0.688        0.738
    8    547      45    0.654 0.01436        0.627        0.683
   10    340      21    0.614 0.01595        0.583        0.646
   12    179      13    0.569 0.01899        0.533        0.608

               age_vc=1, wealth_vc=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    641      69    0.892  0.0122        0.869        0.917
    4    461      19    0.856  0.0144        0.828        0.884
    6    357      19    0.810  0.0170        0.777        0.844
    8    237      12    0.769  0.0198        0.731        0.809
   10    141       6    0.736  0.0230        0.693        0.783
   12     57       1    0.723  0.0260        0.674        0.776
```

We will obtain nine data frames after summarising the km.model4, which is a combination of age_vc and wealth_vc. According to the findings, those under the age of 65 who have less than 60,000 pounds in wealth had a survival probability of 79.4 percent at time 2, with 95 percent confidence ranges of 76.5 percent to 82.4 percent. People under the age of 65 who have a net worth of 60,000 to 300,000 pounds have an 85.7 percent chance of

surviving. According to the findings, those under the age of 60 and with a net worth of more over 300,000 pounds had the highest survival chance of 89.2 percent at time 2.

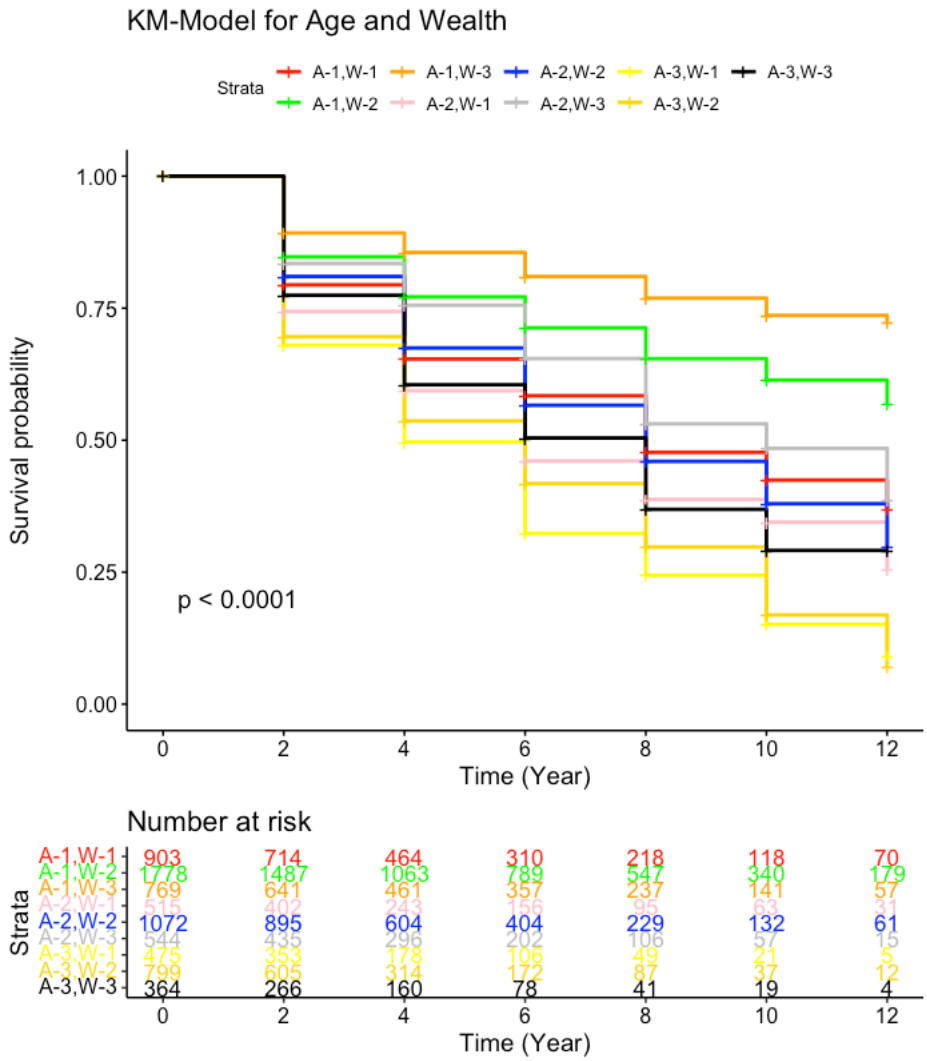

Figure 4.7: KM-Model for Age and Wealth

In this plot, there are nine survival function lines. Each line refers to the combo of age and wealth. Legend explains which colour belongs to which combo.

The given graphic does not provide a good indication of the outcome. So that we can view the output more clearly, I've separated the survival curves. This will provide a clear picture of the model.

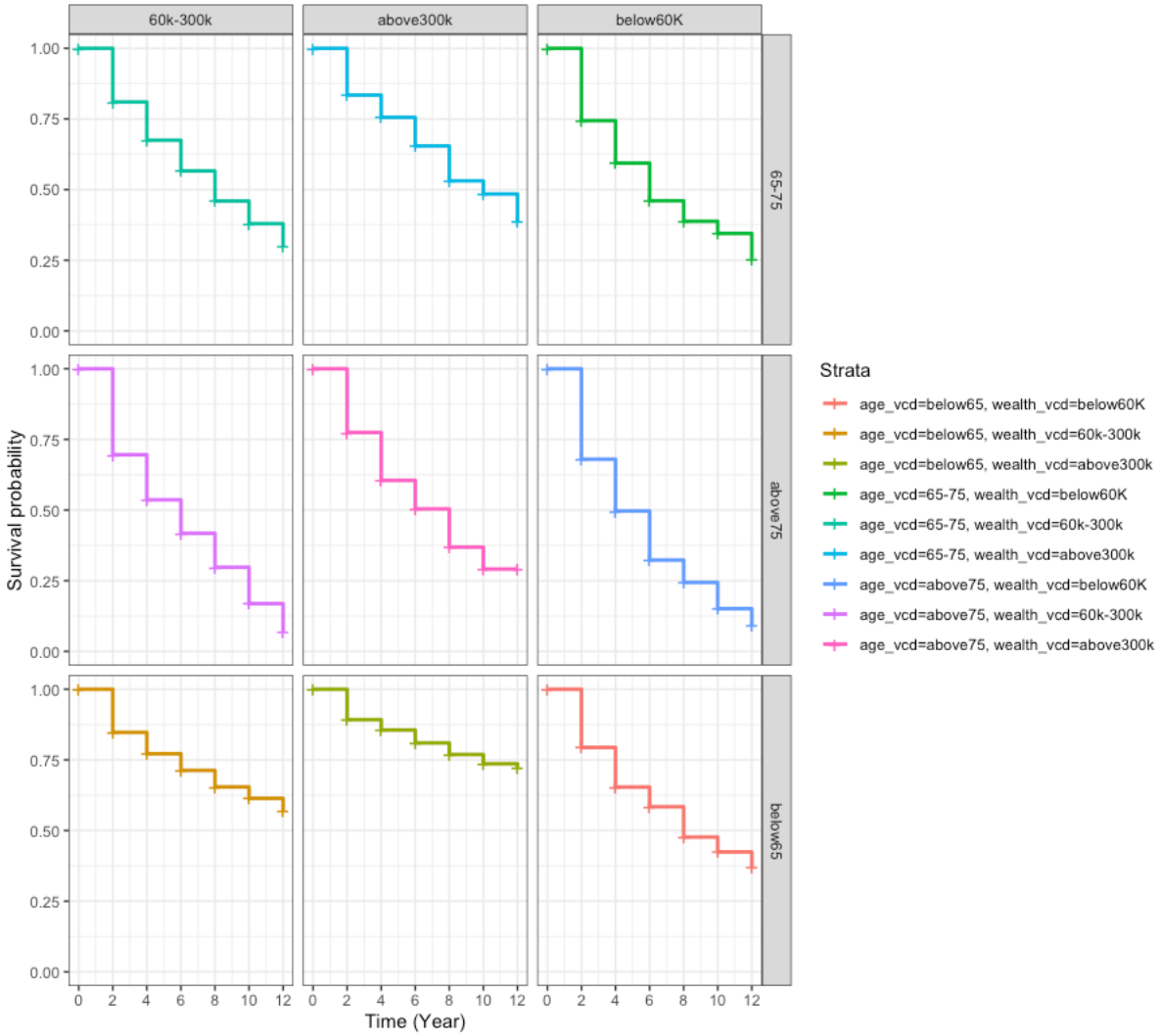

Figure 4.8: KM-Model for Age and Wealth (Survival Curves with separate view)

According to the above plot, persons with a wealth of over 300,000 pounds have a greater survival probability than the rest of the population, and people under the age of 65 have a higher survival probability. The green survival curve in the lower middle has the highest survival probability compared to the other survival curves, according to this graph. This curve represents those under the age of 65 who have a net worth of more than 300,000 pounds. The survival curves in the middle left and middle right have an extremely low likelihood of surviving. Both curves belong to adults over 75 years old.

Log-Rank test: In order to see if the survival functions in km.model4 are statistically different, The result will be obtained using the survdiff function.

```
                          N Observed Expected (O-E)^2/E (O-E)^2/V
age_vc=1, wealth_vc=1   903      324    297.6      2.34      3.20
age_vc=1, wealth_vc=2 1778      461    679.2     70.09    117.78
age_vc=1, wealth_vc=3   769      126    292.3     94.58    128.73
age_vc=2, wealth_vc=1   515      217    157.3     22.62     29.02
age_vc=2, wealth_vc=2 1072      415    366.8      6.34      8.96
age_vc=2, wealth_vc=3   544      155    175.9      2.49      3.22
age_vc=3, wealth_vc=1   475      220    117.8     88.58    112.85
age_vc=3, wealth_vc=2   799      342    202.9     95.27    126.38
age_vc=3, wealth_vc=3   364      123     93.1      9.59     12.05

 Chisq= 477  on 8 degrees of freedom, p= <2e-16
```

The test statistic is 477, and the p-value is less than 0.05, as shown at the end. It rejects the null hypothesis due to the tiny p-value. The survival curves have been demonstrated to be different.

4.1.5 Effect of Age and Wealth on Gender

Now we'll use age and wealth as x factors to estimate Kaplan-Meier survival for both genders. With these x variables, we'll fit the model and save it in km.model5. All the X variables are catalogue variables. This model will produce a result that combines gender, age, and wealth. We can extract the number of persons at risk, the number of occurrences, and the median survival time for all gender, age, and wealth combinations by returning km.model5.

	n	events	median	0.95LCL	0.95UCL
ragender=1, age_vcd=below65, wealth_vcd=below60K	368	126	8	8	10
ragender=1, age_vcd=below65, wealth_vcd=60k-300k	682	175	NA	12	NA
ragender=1, age_vcd=below65, wealth_vcd=above300k	278	46	NA	NA	NA
ragender=1, age_vcd=65-75 , wealth_vcd=below60K	225	82	6	6	8
ragender=1, age_vcd=65-75 , wealth_vcd=60k-300k	436	146	10	8	12
ragender=1, age_vcd=65-75 , wealth_vcd=above300k	254	62	12	8	NA
ragender=1, age_vcd=above75, wealth_vcd=below60K	182	70	6	4	8
ragender=1, age_vcd=above75, wealth_vcd=60k-300k	342	129	6	6	8
ragender=1, age_vcd=above75, wealth_vcd=above300k	197	60	8	6	NA
ragender=2, age_vcd=below65, wealth_vcd=below60K	535	198	8	8	10
ragender=2, age_vcd=below65, wealth_vcd=60k-300k	1096	286	NA	NA	NA
ragender=2, age_vcd=below65, wealth_vcd=above300k	491	80	NA	NA	NA
ragender=2, age_vcd=65-75 , wealth_vcd=below60K	290	135	6	6	8
ragender=2, age_vcd=65-75 , wealth_vcd=60k-300k	636	269	8	8	8
ragender=2, age_vcd=65-75 , wealth_vcd=above300k	290	93	8	8	NA
ragender=2, age_vcd=above75, wealth_vcd=below60K	293	150	4	4	6
ragender=2, age_vcd=above75, wealth_vcd=60k-300k	457	213	6	4	6
ragender=2, age_vcd=above75, wealth_vcd=above300k	167	63	6	4	8

Summary:

```
                ragender=1, age_vcd=below65, wealth_vcd=below60K
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    285      56    0.804  0.0235        0.759        0.851
    4    184      34    0.655  0.0299        0.599        0.716
    6    115      11    0.592  0.0325        0.532        0.660
    8     78      14    0.486  0.0371        0.419        0.564
   10     37       7    0.394  0.0434        0.318        0.489
   12     21       4    0.319  0.0487        0.237        0.430

                ragender=1, age_vcd=below65, wealth_vcd=60k-300k
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    561      78    0.861  0.0146        0.833        0.890
    4    410      37    0.783  0.0180        0.749        0.819
    6    300      26    0.715  0.0208        0.676        0.757
    8    209      19    0.650  0.0237        0.606        0.698
   10    119       7    0.612  0.0263        0.563        0.666
   12     71       8    0.543  0.0328        0.483        0.611
```

The survival probability of each combination of all times is obtained from km.model5 summary. According to the findings, females under the age of 65 who have a net worth of 300,000 pounds had a 90.3 percent chance of surviving at time 2, which is the highest of all the combinations.

I divided the survival curves and compared gender in each plot to get a clear result. By comparing the gender in a single image, this will offer a clear picture of the model.

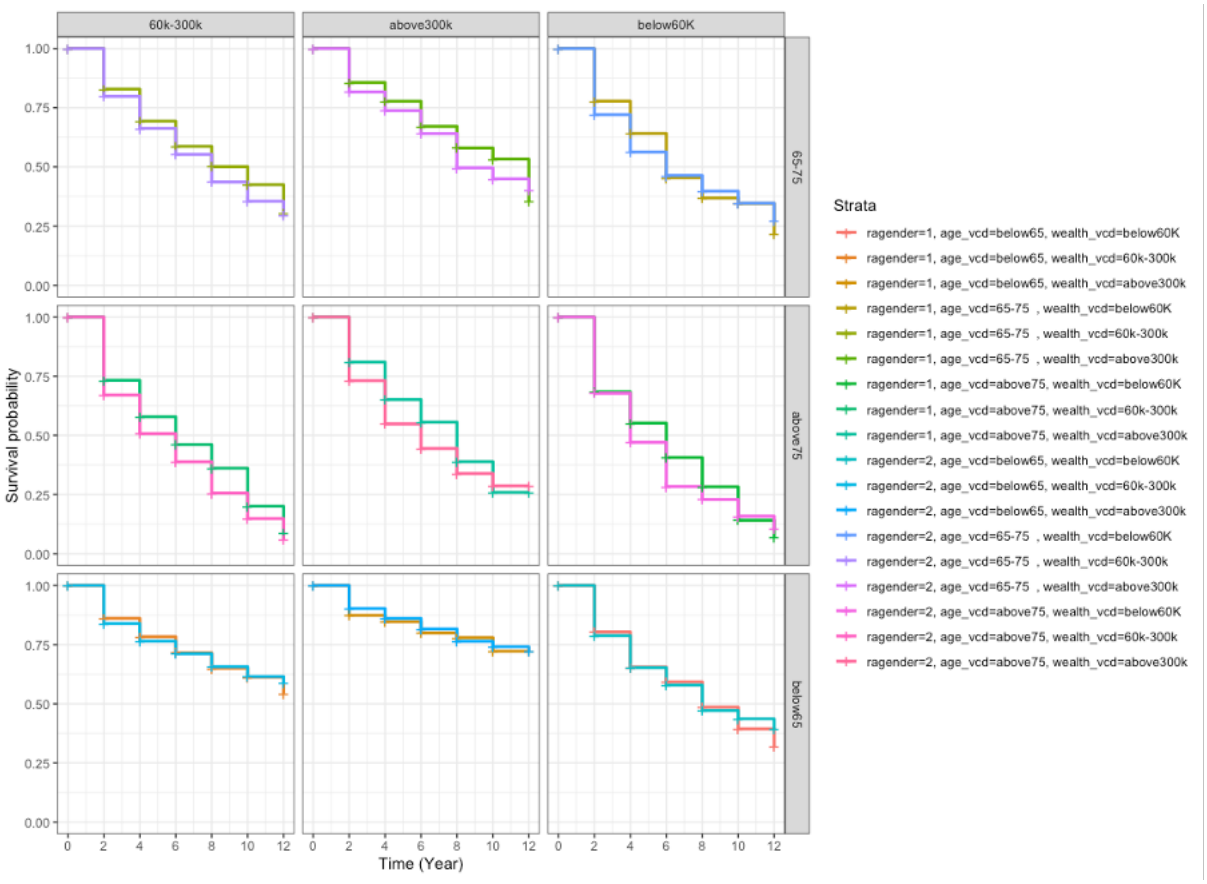

Figure 4.9: KM-Model for Age and Wealth comparing Gender

There are two survival lines in each plot from the above plot. These lines are for both men and women. Males have a higher chance of surviving than females in most plots. We can observe that the bottom middle plot has the highest survival probability curve, which belongs to those under 65 and with a net worth of more than 300,000 pounds, compared to the other curves. For below age of 65, both genders have the same chance of surviving. However, we can clearly observe that males have a better survival rate than females once they reach the age of 75.

Log-Rank test: The survdiff function will be used to determine whether the survival functions in km.model5 are statistically different.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
gender=1, age_vcd=below65, wealth_vcd=below60K	368	126	114.2	1.209	1.520
gender=1, age_vcd=below65, wealth_vcd=60k-300k	682	175	257.6	26.468	35.440
gender=1, age_vcd=below65, wealth_vcd=above300k	278	46	101.6	30.464	37.973
gender=1, age_vcd=65-75 , wealth_vcd=below60K	225	82	61.1	7.184	8.853
gender=1, age_vcd=65-75 , wealth_vcd=60k-300k	436	146	139.9	0.263	0.334
gender=1, age_vcd=65-75 , wealth_vcd=above300k	254	62	77.5	3.105	3.846
gender=1, age_vcd=above75, wealth_vcd=below60K	182	70	40.2	22.015	27.010
gender=1, age_vcd=above75, wealth_vcd=60k-300k	342	129	84.8	23.030	28.834
gender=1, age_vcd=above75, wealth_vcd=above300k	197	60	50.7	1.718	2.121
gender=2, age_vcd=below65, wealth_vcd=below60K	535	198	183.4	1.168	1.513
gender=2, age_vcd=below65, wealth_vcd=60k-300k	1096	286	421.6	43.621	63.428
gender=2, age_vcd=below65, wealth_vcd=above300k	491	80	190.6	64.186	83.221
gender=2, age_vcd=65-75 , wealth_vcd=below60K	290	135	96.3	15.564	19.415
gender=2, age_vcd=65-75 , wealth_vcd=60k-300k	636	269	226.9	7.830	10.346
gender=2, age_vcd=65-75 , wealth_vcd=above300k	290	93	98.4	0.296	0.370
gender=2, age_vcd=above75, wealth_vcd=below60K	293	150	77.6	67.554	84.576
gender=2, age_vcd=above75, wealth_vcd=60k-300k	457	213	118.1	76.162	97.150
gender=2, age_vcd=above75, wealth_vcd=above300k	167	63	42.5	9.948	12.193

Chisq= 489 on 17 degrees of freedom, p= <2e-16

As indicated at the conclusion, the test statistic is 489, and the p-value is less than 0.05. The null hypothesis is rejected due to the small p-value. It has been established that the survival curves differ.

4.2 Cox Proportional-Hazard Model

In this section, We will fit the model for sex and age individually in this Cox Proportional-Hazard Model. Then see if the model's fit is improved by combining wealth with sex and age, and we'll get the perfect fit Cox Proportional-Hazard Model.

4.2.1 Cox model for Individual X variables

Gender: Let's start with a single covariate Cox proportional model. We can't get survival probabilities directly from the Cox model since, as previously stated, it doesn't estimate the baseline hazard. To do so, we'll need to combine it with a non-parametric baseline hazard function estimator. These probabilities are generated by function survfit() for a fitted Cox model from package survival. To fit the cox model, the coxph function is employed. Let gender be the covariate in this case. Model is stored in cox.mod1 object.

we will return the summary of cox.mod1 using summary function.

```
  n= 7219, number of events= 2383

             coef exp(coef) se(coef)     z Pr(>|z|)
ragender2 0.06326   1.06530  0.04231 1.495    0.135

          exp(coef) exp(-coef) lower .95 upper .95
ragender2     1.065     0.9387    0.9805     1.157

Concordance= 0.514  (se = 0.006 )
Likelihood ratio test= 2.25  on 1 df,   p=0.1
Wald test            = 2.24  on 1 df,   p=0.1
Score (logrank) test = 2.24  on 1 df,   p=0.1
```

The first thing we notice in this summary is that it returns n, which is the number of observations we have, which is 7219, as well as the number of occurrences or persons that become frail, which is 2383. We may conclude that 4836 observations have been censored as a result of this. We can see that we are retuning the model coefficient in the first column, and one thing to note is that we are not given the B_0 or an interceptor. We can't predict the survival function in this model. The Hazard ratios, on the other hand, may be estimated. It also offers us the hazard ratio by returning the exponentiated coefficient. We can observe that the hazard ratio for ragender 2, which is females, is 1.065 by looking at the exponentiated coefficient. The interpretation is that females are 1.065 times more likely than males to become frail at any given time. We may calculate the percentage by removing one from the hazard ratio. This equals 1.065 – 1. We're going to receive 0.065. Females are 6.5 percent more likely than males to become frail at any particular time. It returns the exponentiated negative coefficient and 95 percent of the hazard ratio's confidence interval if we notice additional outputs down there. Females had a hazard ratio of 1.065, with a 95 percent confidence range of 0.98 to 1.157. As a result, we have a 95% confidence that the true hazard ratio is within this range. The reciprocal of the hazard ratio is the exponentiated negative coefficient. To put it another way, the reference group is shifting. The exponentiated negative coefficient hazard ratio is 0.938. This is the male-to-female hazard ratio. This means that males are 0.934 times more likely than females to become frail at any particular time. Finally, it returns the likelihood ratio test, Wald test, and log-rank test at the bottom. These are the tests for the null hypothesis that b_1 equals b_2 up to the point where b_k equal to zero. At least one that isn't zero is an alternative. This is a test of the model's overall significance.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Now let's look at the Cox proportional model's proportional assumption. We'll utilise the cox.zph() method to do so. Hazards are proportional is the null hypothesis, whereas hazards are not proportional is the alternative hypothesis. The p values of the covariates will be returned.

We will receive the following result.

```
          chisq df     p
ragender  6.83  1 0.009
GLOBAL    6.83  1 0.009
```

The p-value of ragender is 0.009, which is less than 0.05, according to the results. The null hypothesis is rejected due to low p values. This indicates that it is not proportional. As a result, the proportional hazards assumptions are violated. As a result, satisfying the ragender will be the solution to this problem. Using the strata() function, we include the stratifying component in the model's calculation to fit the stratified Cox model. We don't have to worry about the proportional risks assumption if we satisfy.

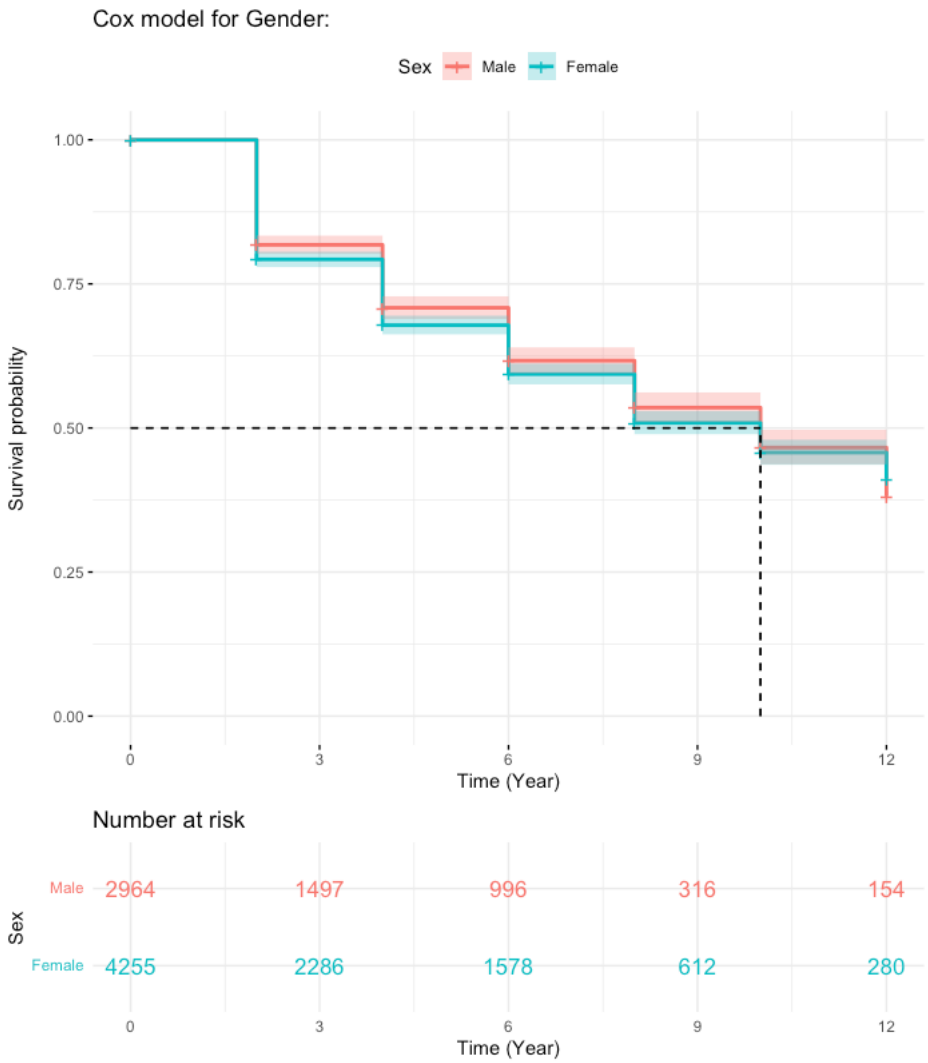

Figure 4.10: Cox Model for Gender

The ggsurvplot function is used to plot the Cox model for gender (cox.mod1.2). The x-axis in this graph represents time in years, while the y-axis represents survival chance. The survival lines in this plot are red for men and blue for females. In summary, females are 6.5 percent more likely than males to become frail, as we've seen before. This is seen in the graph. Males have a better chance of surviving than females.

AGE: The advantage of the Cox model is that numerical covariate can be used. We can't utilise numerical X in the K-M model. Let the numerical age be the x variable now. The Coxph function is used to fit and store the model in the cox.mod8 object.

summary of cox.mod8 is shown below:

```
  n= 7219, number of events= 2383

          coef exp(coef) se(coef)     z Pr(>|z|)
age_v 0.040194  1.041012 0.002027 19.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
age_v     1.041     0.9606     1.037     1.045

Concordance= 0.612  (se = 0.007 )
Likelihood ratio test= 395  on 1 df,   p=<2e-16
Wald test            = 393.2  on 1 df,   p=<2e-16
Score (logrank) test = 400.6  on 1 df,   p=<2e-16
```

By looking at the exponentiated coefficient for the numerical age, here comes out the hazard ratio is 1.041 with a 95% confidence interval of 1.037 up to 1.045. Now by interpreting this, we would say that at the given instant in time, the probability of becoming frail for someone who is one year older is 4.1% higher than the someone who is one year younger.

Checking LINEARITY assumption using Martingale: Only the numerical x variables can be used to test the linearity assumption. If we recall, we assumed that the relationship between any of the numeric x variables and the log hazard is linear. We may confirm the linearity by looking at the residual plot on the x-axis, then plotting the predicted values on the y axis, and finally plotting the residuals. In survival analysis, there are a few different types of residuals. We'll use martingale residuals in this case. The plot below is the result of plotting.

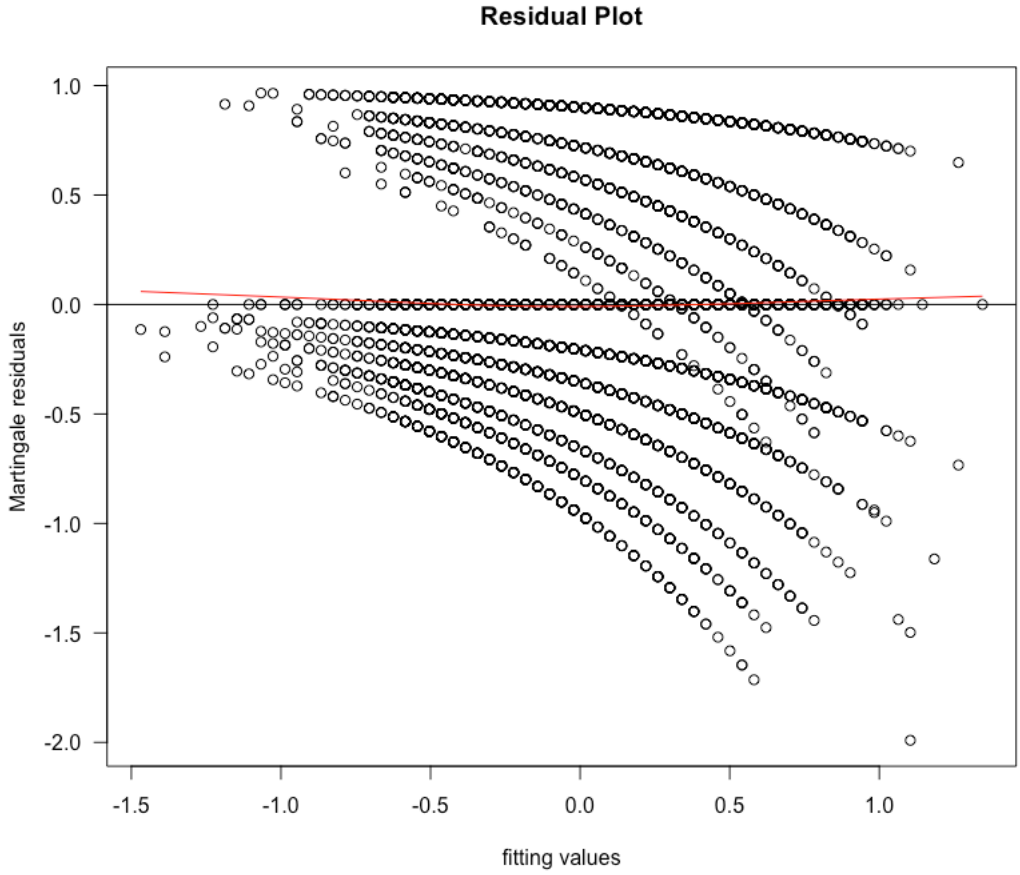

Figure 4.11: Residual Plot to Check Linearity

The red line in the above figure demonstrates little non-linearity. As a result, the LINEARITY assumption is violated. So what we can do is let's transform the X variable. i.e., a person's age.

The Cox proportional model uses the Square of age as an X variable, which is stored in the cox.mod8.1 object. Let's look at the linearity assumption for cox.mod8.1 now.

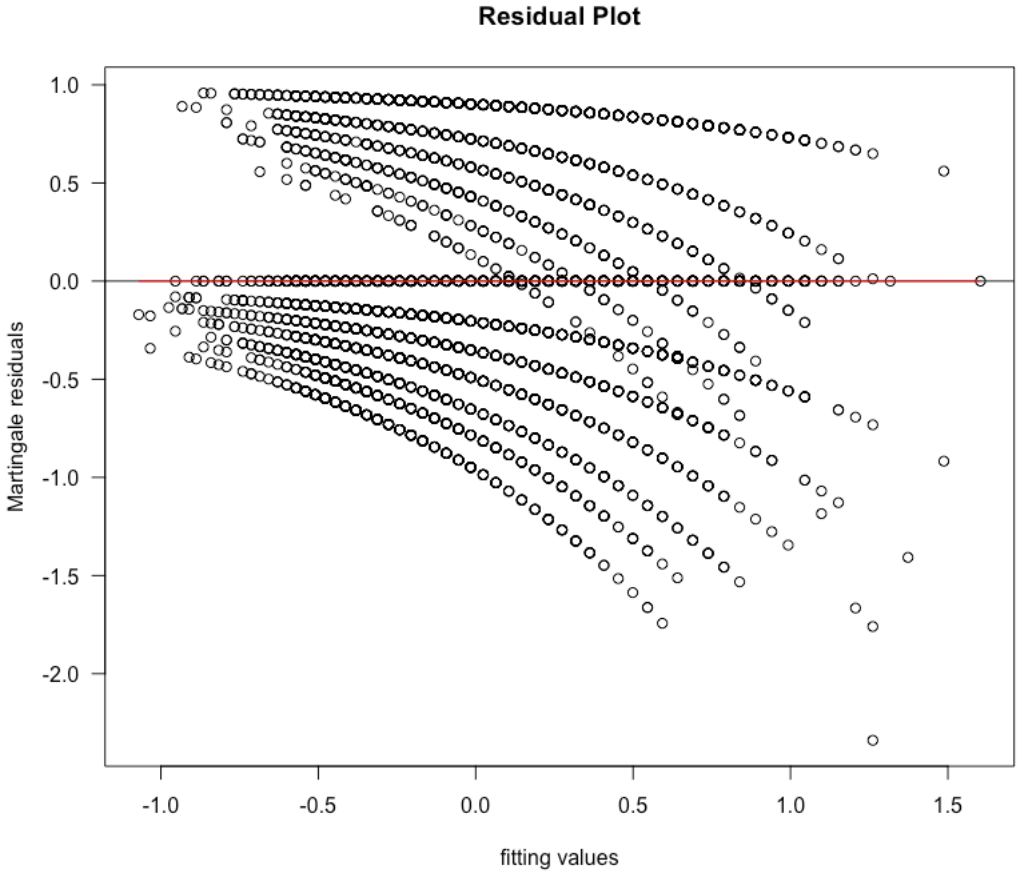

Figure 4.12: Residual Plot to Check Linearity

By transforming age, now we can see the red line is straight. So this model meets the linearity.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Now consider the proportional assumption of the Cox proportional model. The null hypothesis is that hazards are proportional, whereas the alternative hypothesis is that hazards are not proportional. If you employ the cox.zph function, you'll get the p values of the covariates.

```
            chisq df     p
I(age_v^2)   44.2  1 3e-11
GLOBAL       44.2  1 3e-11
```

Here we can see the p-value of the square of age is returned where the p-value is smaller than 0.05. As a result of the low p values, the null hypothesis is rejected. This shows that it isn't proportionate. The proportional hazards assumptions are therefore broken. As a result, we've decided to enter the catalogue age.

We don't need to check linearity for the catalogue variable.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: We may reach the following conclusion by testing the proportional hazards assumption for cox.mod8.2. That is, the p-value is lower. As a result, the null hypothesis is rejected, and the result is not proportionate. As a result, it is in violation of the assumption. The answer to this problem will be to satisfy the age_vc. The strata() function is used.

```
        chisq df     p
age_vc   37.9  2 6e-09
GLOBAL   37.9  2 6e-09
```

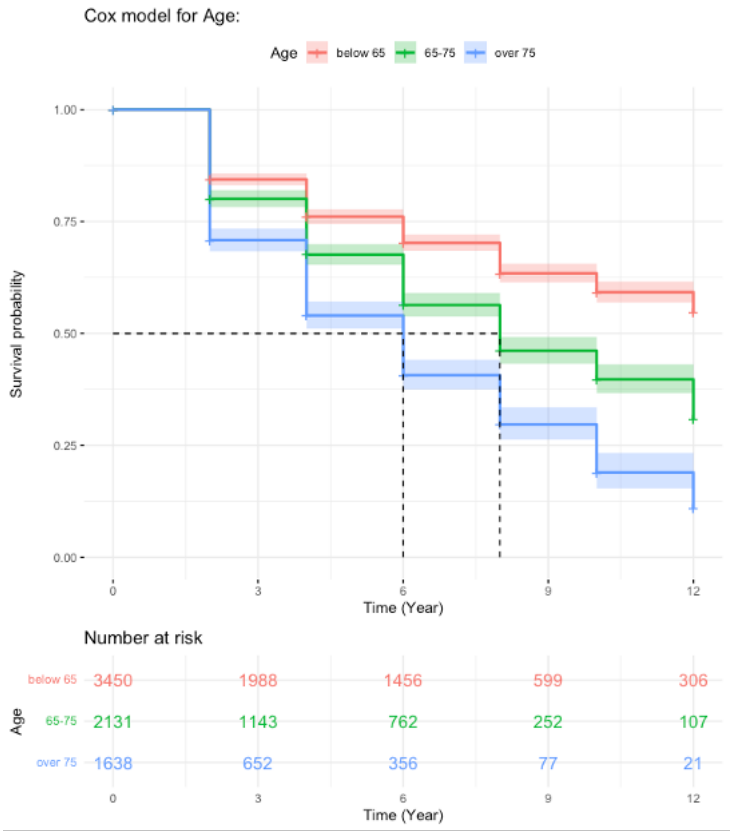

Figure 4.13: Cox Model for Age

The Cox model for age group is plotted using the ggsurvplot function (cox.mod8.3). The x-axis in this graph indicates years, while the y-axis shows the probability of survival. The survival curves in this graph are red for those under the age of 65, green for those 65 to 75, and blue for those above 75. The plot shows that those under the age of 65 have a better chance of surviving than people between the ages of 65 and 75, and then people above 75. This shows that younger individuals have a lower risk of becoming frail.

To double-check that younger individuals have a lower risk of becoming frail. I divided the age group into four categories. It was previously separated into three categories. Let's plot the survival probability for each age group now.

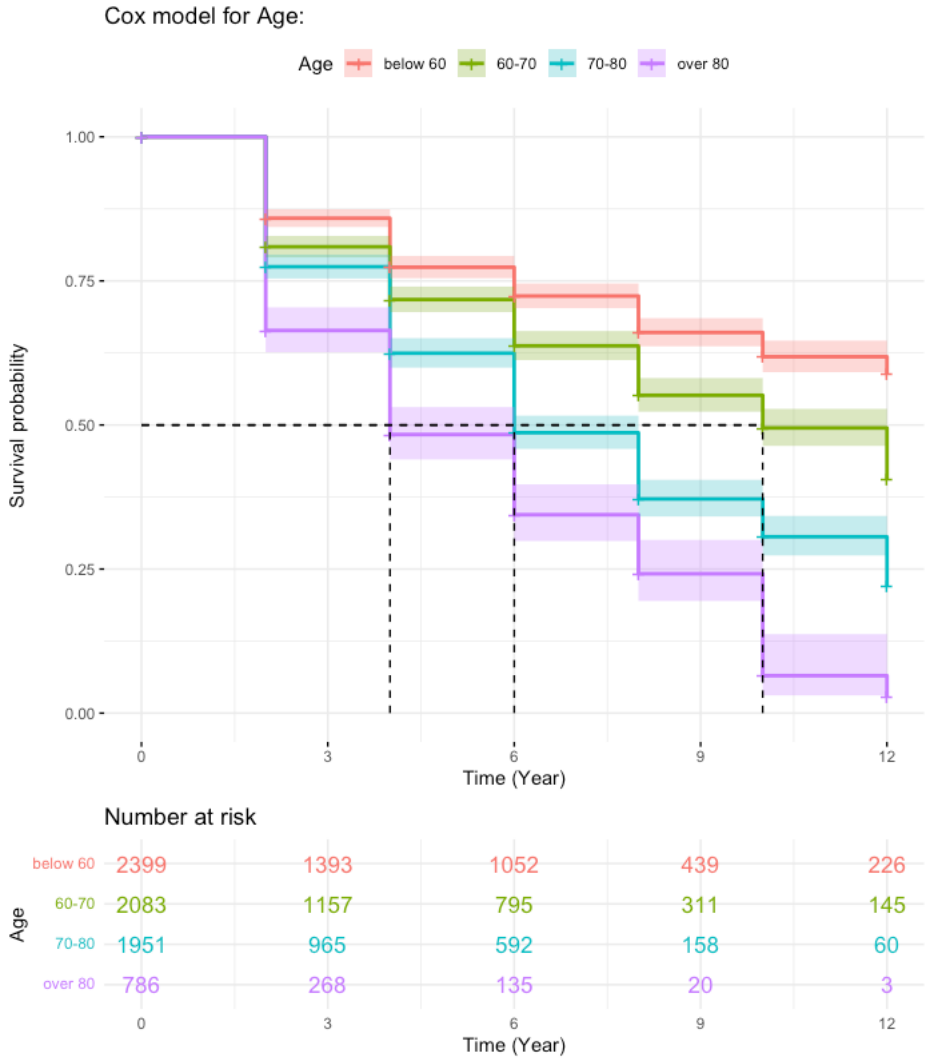

Figure 4.14: Cox Model for Age (4 Groups)

Here, we can see that the survival curves of all four groups are still in the same order, with younger individuals having a higher chance of surviving than older ones. This shows that older adults are more likely to become frail.

WEALTH: Let's use the wealth catalogue variable to fit the cox proportional model. This will show how people's frailty is affected by their wealth. As a result, the model is fit and saved as cox.mod9.1.

Summary of the cox.mod9.1 is given below.

```
  n= 7219, number of events= 2383

               coef exp(coef) se(coef)       z Pr(>|z|)
wealth_vc2 -0.34154   0.71068  0.04623  -7.387  1.5e-13 ***
wealth_vc3 -0.66816   0.51265  0.06159 -10.849  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7107      1.407    0.6491    0.7781
wealth_vc3    0.5127      1.951    0.4544    0.5784

Concordance= 0.564  (se = 0.006 )
Likelihood ratio test= 125.8  on 2 df,   p=<2e-16
Wald test            = 124.9  on 2 df,   p=<2e-16
Score (logrank) test = 127.7  on 2 df,   p=<2e-16
```

The hazard ratio for wealth 2 is 0.7107, with a 95 percent confidence range of 0.649 to 0.7781, according to the exponentiated coefficient. Now, analysing this, we can conclude that at any given time, someone with a net worth of 60,000 to 300,000 pounds has a 28.9 percent lower chance of getting frail than someone with a net worth of less than 60,000. People with a net worth of over 300,000 pounds have a lower risk of becoming frail than other categories of people.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Let's check the proportional hazards assumption using cox.zph() function on cox.mod9.1.

```
          chisq df    p
wealth_vc  1.74  2 0.42
GLOBAL     1.74  2 0.42
```

Where the p-value of the wealth is larger than 0.05. The null hypothesis fails to reject due to the huge p values. This demonstrates that it is in proportional. As a result, the proportional hazards assumptions are not violated.

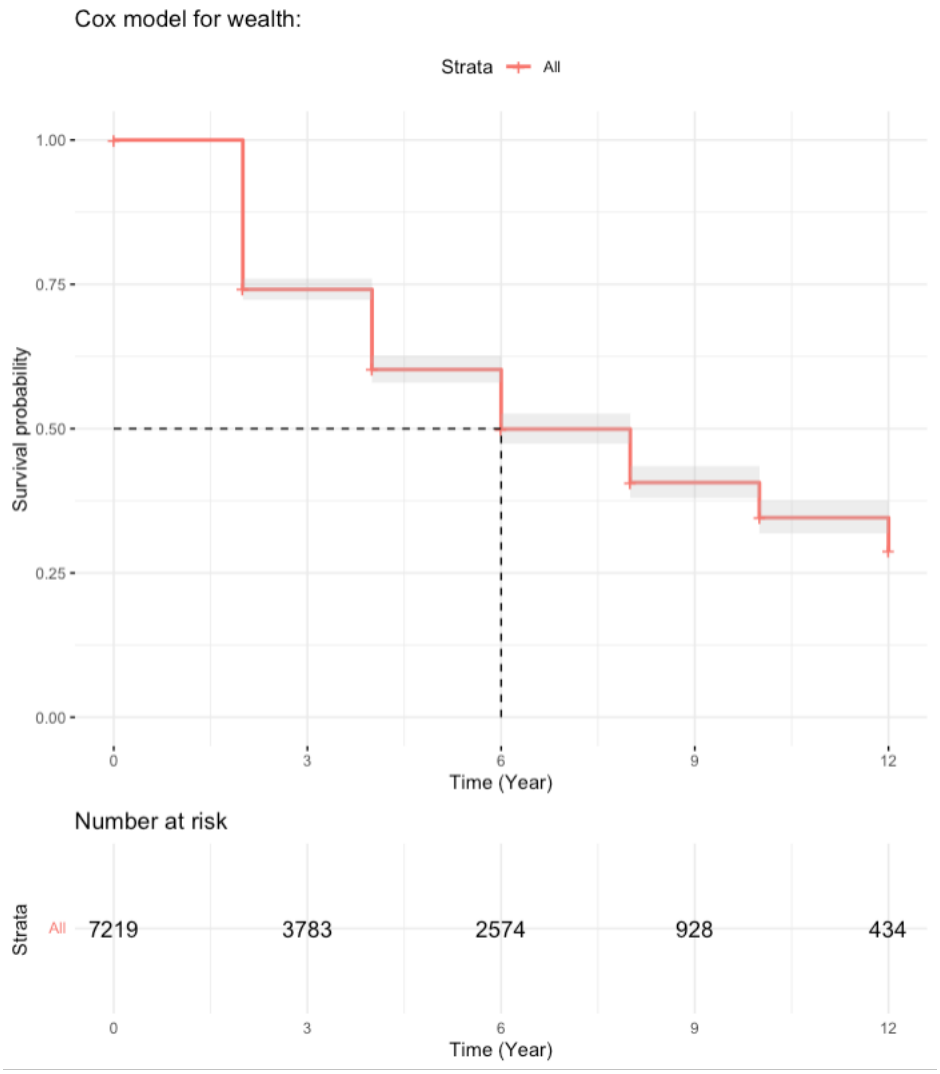

Figure 4.15: Cox Model for Wealth

When Cox.mod9.1 is plotted, we obtain a single survival curve that indicates how people's overall wealth influences their frailty. It does not illustrate how different levels of wealth have an impact on people's frailty. As a result, I was able to fulfil the wealth variable and obtain the plot below.

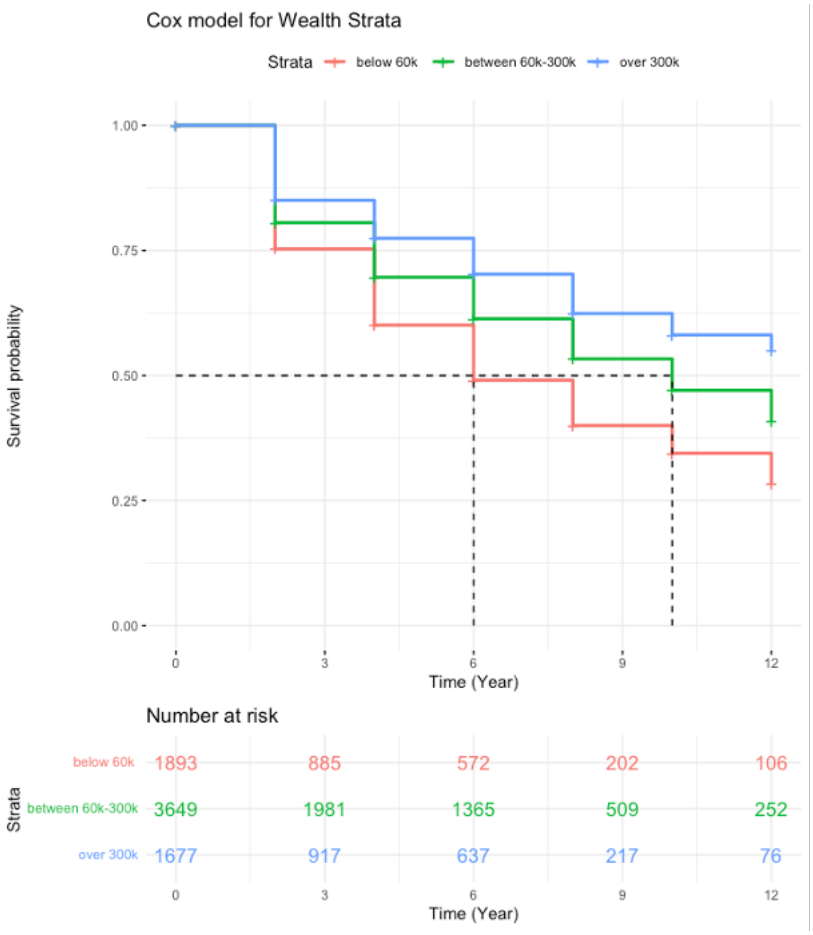

Figure 4.16: Cox Model for Wealth Strata

Each group has a different survival curve in this graph. The red curve represents individuals with wealth of less than 60,000 pounds, the green line represents those with wealth of 60,000 to 300,000 pounds and the blue line represents those with wealth exceeding 300,000 pounds. We can see from the plot that those with more money have a better chance of surviving than those with less wealth. This graph also demonstrates that money has a significant impact in the model.

4.2.2 Let's check adding wealth in X variable improves the model

Cox model for strata(ragender) + wealth_vc: Let's see if wealth helps the model when it's combined with other factors. To do so, I must first fit the cox model to the satisfied gender and wealth, and then store it in the cox.mod14 object.

By returning the summary for cox.mod14 we will get following report.

```
  n= 7219, number of events= 2383

               coef exp(coef) se(coef)       z Pr(>|z|)
wealth_vc2 -0.34358   0.70923  0.04625  -7.429 1.09e-13 ***
wealth_vc3 -0.66619   0.51366  0.06160 -10.814  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7092      1.410    0.6478    0.7765
wealth_vc3    0.5137      1.947    0.4552    0.5796

Concordance= 0.564  (se = 0.006 )
Likelihood ratio test= 125.2  on 2 df,   p=<2e-16
Wald test            = 124.5  on 2 df,   p=<2e-16
Score (logrank) test = 127.2  on 2 df,   p=<2e-16
```

The hazard ratio for wealth 2 is 0.709, with a 95 percent confidence interval of 0.647 to 0.7765, according to the summary. We may deduce from this that someone with a net worth of 60,000 to 300,000 pounds has a 29.1% lower risk of being frail than someone with a net worth of less than 60,000 pounds at any given time. Followed by wealth 3.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Let's check the proportional hazards assumption using cox.zph() function on cox.mod14.

```
          chisq df    p
wealth_vc  2.02  2 0.36
GLOBAL     2.02  2 0.36
```

The null hypothesis fails to reject when the wealth p-value is more than 0.05 due to the big p values. This proves that it is proportional. The proportional hazards assumptions are not violated as a result.

Likelihood Ratio test: The likelihood ratio test may be used to evaluate nested models and see if adding a variable improves the model or if the interaction or effect modification term is statistically significant. We'll perform the likelihood ratio test to see whether we can remove wealth_vc without significantly worsening our model. Here cox.mod14 is a model with ragender and wealth. Where cox.mod1.2 is with ragender alone.

cox.mod1.2 —- strata(ragender)

cox.mod14 —- strata(ragender) + wealth_vc

We will use the likelihood ratio test on both models using anova() function.

The following function will return the following result.

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ strata(ragender)
 Model 2: ~ strata(ragender) + wealth_vc
  loglik  Chisq Df P(>|Chi|)
1 -17749
2 -17686 125.19  2 < 2.2e-16 ***
```

Here we find the p-value is small that is less than 0.05. That is a statistical difference between the two models, and we can keep wealth_vc. Because of the small P-value, model 2 is the best. So, a model with wealth performs well than a model without wealth.

Cox model for strata(age_vc) + wealth_vc : Let's fit the cox proportional model using satisfied age and wealth as x variables and save it in the cox.mod13 object.

Summary of the cox.mod13 model is returned.

```
  n= 7219, number of events= 2383

               coef exp(coef) se(coef)       z Pr(>|z|)
wealth_vc2 -0.30229   0.73912  0.04632  -6.526 6.75e-11 ***
wealth_vc3 -0.64337   0.52552  0.06167 -10.433  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7391      1.353    0.6750    0.8094
wealth_vc3    0.5255      1.903    0.4657    0.5930

Concordance= 0.571  (se = 0.007 )
Likelihood ratio test= 114.9  on 2 df,   p=<2e-16
Wald test            = 112.5  on 2 df,   p=<2e-16
Score (logrank) test = 114.8  on 2 df,   p=<2e-16
```

The hazard ratio of wealth 2 is 0.7391, with a confidence range of 0.675 to 0.8094, according to the model's summary. In other words, someone with a net worth of 60,000 to 300,000 pounds has a 26% reduced chance of becoming frail at any given moment than someone with a net worth of less than 60,000 pounds.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Let's use the cox.zph() function on cox.mod13 to test the proportional hazards assumption.

```
          chisq df    p
wealth_vc  1.13  2 0.57
GLOBAL     1.13  2 0.57
```

Null hypothesis fails to reject when the wealth p-value is more than 0.05. This proves that it is proportional. As a result, the proportional hazards assumptions are not violated.

Likehood test: We'll use the likelihood ratio test to check whether we can get rid of wealth_vc without making our model much worse. Cox.mod13 is a model that includes age_vc and wealth. Where cox.mod8.3 is used just by ragender.

cox.mod8.3 —- strata(age_vc)

cox.mod13 —- strata(age_vc) + wealth_vc

We'll use the anova() function to perform a likelihood ratio test on both models.

$$anova(cox.mod8.3, cox.mod13, test = "LRT")$$

The code above will provide the following result.

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ strata(age_vc)
 Model 2: ~ strata(age_vc) + wealth_vc
  loglik Chisq Df P(>|Chi|)
1 -16552
2 -16495 114.9  2 < 2.2e-16 ***
```

The p-value is tiny, less than 0.05, in this case. We can maintain wealth_vc with age_vc since there is a statistical difference between the two models. Model 2 is the best due to the low P-value. As a result, a model with wealth outperforms a model without wealth.

Both likelihood ratio tests for wealth with gender and age yielded the same results. We can see how bringing wealth into the equation improves the model.

4.2.3 Find the perfect fit model

Now let's find the perfect model by fitting Cox proportional model using multiple covariates. Now we are going to use ragender, age_vc and wealth_vc as an x variable in the model and store it in cox.mod16 object.

Summary of the cox.mod16 model is returned.

```
  n= 7219, number of events= 2383

               coef exp(coef) se(coef)       z Pr(>|z|)
ragender2   0.07623   1.07921  0.04237   1.799    0.072 .
age_vc2     0.50002   1.64875  0.04878  10.249  < 2e-16 ***
age_vc3     0.92535   2.52276  0.05133  18.026  < 2e-16 ***
wealth_vc2 -0.30490   0.73720  0.04632  -6.582 4.65e-11 ***
wealth_vc3 -0.64130   0.52661  0.06171 -10.392  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
ragender2     1.0792     0.9266    0.9932    1.1727
age_vc2       1.6487     0.6065    1.4984    1.8142
age_vc3       2.5228     0.3964    2.2813    2.7898
wealth_vc2    0.7372     1.3565    0.6732    0.8073
wealth_vc3    0.5266     1.8990    0.4666    0.5943

Concordance= 0.625  (se = 0.007 )
```

```
Likelihood ratio test= 450.7  on 5 df,   p=<2e-16
Wald test            = 460.7  on 5 df,   p=<2e-16
Score (logrank) test = 481.8  on 5 df,   p=<2e-16
```

The hazard ratios for gender, age groups, and wealth groups may be seen in the summary. The 95 percent confidence range for the hazard ratios was also returned. Let's check for assumptions before interpreting the summary.

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: Let's check the proportional hazards assumption using cox.zph() function on cox.mod16.

```
           chisq df       p
ragender    6.19  1   0.013
age_vc     36.24  2 1.3e-08
wealth_vc   1.44  2   0.486
GLOBAL     42.73  5 4.2e-08
```

We can observe from the result that ragender and age_vc have tiny p values, indicating that the hazards are not proportional. We can also observe that the model's global p-value is less than 0.05. As a result, the model isn't proportional. As a result, both ragender and age_vc are satisfied.

The summary of cox.mod16.1 is shown below.

```
  n= 7219, number of events= 2383

                coef exp(coef) se(coef)       z Pr(>|z|)
wealth_vc2 -0.30029   0.74060  0.04639  -6.473 9.61e-11 ***
wealth_vc3 -0.62902   0.53312  0.06189 -10.163  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2     0.7406      1.350    0.6762    0.8111
wealth_vc3     0.5331      1.876    0.4722    0.6019

Concordance= 0.57  (se = 0.007 )
Likelihood ratio test= 109.1  on 2 df,   p=<2e-16
Wald test            = 107.2  on 2 df,   p=<2e-16
Score (logrank) test = 109.3  on 2 df,   p=<2e-16
```

CHECKING PROPORTIONAL HAZARDS ASSUMPTION: We obtain the following result when we examine the proportional hazard assumption again. Only wealth_vc is visible in it. This is due to the fact that the satisfied variable will not appear here. Now that wealth_vc and global have p-values larger than 0.05, the null hypothesis is no longer rejected. As a result, it is proportional. So that the proportional hazard assumption is not broken.

```
           chisq df    p
wealth_vc   1.22  2 0.54
GLOBAL      1.22  2 0.54
```

cox.mod16.1 is plotted.

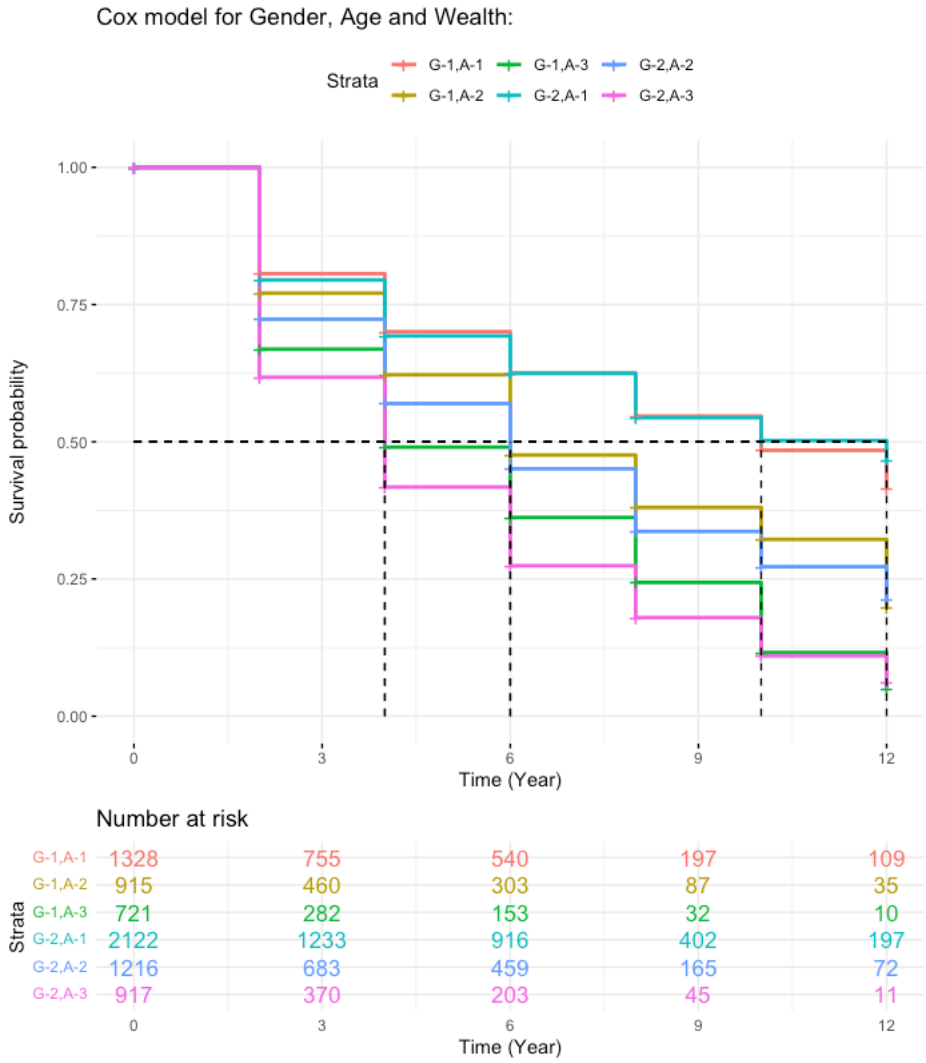

Figure 4.17: Cox Model for Gender, Age and Wealth

Females above the age of 75, which is shown by a pink survival curve in the graph above, have the lowest survival probability of all the categories of individuals. Males and females under the age of 65 have a better probability of surviving than other age groups.

Let's try increasing the number of age categories to see if the model improves or not. As a result, we'll use age_vc4 instead of age_vc.

AIC test is used to compare: Let's examine the cox.mod16.1 and cox.mod16.2 models to see if dividing ages into more groups helps the model improve. The quality of a group of statistical models is compared using Akaike's information criterion (AIC). When comparing models fitted to the same data using maximum likelihood, the smaller the AIC, the better the fit. We must first load the MASS library before we can begin AIC testing. After that, we'll use the AIC tool to compare the two models. The numeric values will be returned as a result of this. We may use this to identify the best model.

By using the function, we will get the following result.

```
> AIC(cox.mod16.1)
[1] 29831.96
> AIC(cox.mod16.2)
[1] 28575.61
```

The output of cox.mod16.2 is smaller than the model cox.mod16.1, as can be seen from the output. This demonstrates that the model has improved. The model is improved by dividing the age groups. So let's strata the entire age. Fit the model in the object cox.mod16.3.

Summary of the cox.mod16.3 is returned.

```
  n= 7219, number of events= 2383

               coef exp(coef) se(coef)       z Pr(>|z|)
wealth_vc2 -0.31372   0.73072  0.04772  -6.574 4.91e-11 ***
wealth_vc3 -0.64459   0.52488  0.06321 -10.198  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7307      1.369    0.6655    0.8024
wealth_vc3    0.5249      1.905    0.4637    0.5941

Concordance= 0.569  (se = 0.007 )
Likelihood ratio test= 109.6  on 2 df,   p=<2e-16
Wald test            = 107.9  on 2 df,   p=<2e-16
Score (logrank) test = 109.9  on 2 df,   p=<2e-16
```

We estimated the hazard ratio for wealth_vc 2 as 0.7307, with a 95 percent confidence range of 0.6655 to 0.8024, by fitting the final model. By translating, we can state that someone with a net worth of 60,000 to 300,000 has a 27% lower risk of being frail than someone with a net worth of less than 60,000 at any one time. Those with a net worth of over 300,000, however, have a better chance of surviving than those with less.

AIC test is used to compare: Let's use the AIC test to assess all three models and identify the best fit for the Cox proportional model. The following result is obtained using the AIC function.

```
> AIC(cox.mod16.1)
[1] 29831.96
> AIC(cox.mod16.2)
[1] 28575.61
> AIC(cox.mod16.3)
[1] 17328.12
```

As we can be seen, cox.mod16.3 has the lowest score of the three models, indicating that dividing the age groups improves the model significantly. As a result, we can confidently state that cox.mod16.3 is an excellent match for the model.

4.2.4 Comparing Gender

We've just fitted the cox proportional model for both genders together thus far. So, let's fit both the male and female Cox proportional models independently and evaluate the results individually.

Male: The male data from the cox_fd has now been filtered and saved in the male_dataset. We'll use male_dataset to fit the cox proportional model using satisfied age and wealth and save it to the cox.mod19.m object.

Summary of the cox.mod19.m is returned.

```
  n= 2964, number of events= 896

               coef exp(coef) se(coef)      z Pr(>|z|)
wealth_vc2 -0.35102   0.70397  0.07642 -4.594 4.36e-06 ***
wealth_vc3 -0.65350   0.52022  0.09801 -6.668 2.60e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7040      1.421    0.6061    0.8177
wealth_vc3    0.5202      1.922    0.4293    0.6304

Concordance= 0.569  (se = 0.011 )
Likelihood ratio test= 46.76  on 2 df,   p=7e-11
Wald test            = 46.95  on 2 df,   p=6e-11
Score (logrank) test = 47.97  on 2 df,   p=4e-11
```

According to the model's summary, males with wealth between 60,000 and 300,000 pounds had a 29.6% lower likelihood of becoming frail than males with wealth less than 60,000 at any given time, controlling for wealth_vc3.

Female: The female data from the cox_fd has now been filtered and saved in the female_dataset. We'll use female_dataset to fit the cox proportional model using satisfied age and wealth and save it to the cox.mod19.f object.

Summary of the cox.mod19.f is returned.

```
  n= 4255, number of events= 1487

               coef exp(coef) se(coef)      z Pr(>|z|)
wealth_vc2 -0.27094   0.76267  0.05836 -4.643 3.44e-06 ***
wealth_vc3 -0.61577   0.54023  0.07985 -7.711 1.25e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
wealth_vc2    0.7627      1.311    0.6802    0.8551
wealth_vc3    0.5402      1.851    0.4620    0.6318

Concordance= 0.571  (se = 0.009 )
Likelihood ratio test= 63.05  on 2 df,   p=2e-14
Wald test            = 61.1  on 2 df,   p=5e-14
Score (logrank) test = 62.24  on 2 df,   p=3e-14
```

According to the model's summary, females with wealth between 60,000 and 300,000 pounds had a 23.7 percent lower likelihood of becoming frail than females with wealth less than 60,000 pounds at any given moment, controlling for wealth_vc3. Similarly, wealthier females have a better probability of surviving than non-wealthy females.

Chapter 5

Conclusions

In this study, I included people who are pre-frail and are looking for signs of frailty. To be more specific, I may have included all individuals who are pre-frail when they join the cohort, and so I must account for the left censoring. Interval censoring is also present, which I have only observed every two years. There are some observations where they enter the study with frailty. We don't include these observations in the study since we don't know when they are pre-frail. As a result, we won't get the time it takes to get from pre-frail to frail.

Using survival analysis, we investigated the frailty of persons in England over the age of 50 in our study. This research aids us in determining the impact of gender, age, money, on frailty. To sum up the findings of this study, we can say that men have a larger chance of surviving than females. This suggests that in England, ladies have a larger risk of becoming frail than males. Frailty affects women 6.5 percent more than it does men. Furthermore, as people age, their chances of getting frail grow. People under the age of 65 have a greater probability of surviving than those between 65 to 75, and then those aged 75 and over. People who are younger have a lower risk of becoming frail than those who are older. In addition, overall family wealth has an impact on the onset of frailty. This is supported by the findings of a survival study. According to the report, as a family's overall wealth grows, the probability of getting frail reduces. People who have a total wealth of less than 60,000 pounds are more likely to become frail than those who have a total wealth of more than 60,000 pounds. People with a total wealth of more than 300,000 pounds are also at a lower risk of getting frail than those with less than 300,000 pounds. This clearly demonstrates that money has an impact on the development of frailty.

The Cox model is non-linear when stratifying for gender with age and wealth as continuous variables. So, in order to discover the optimal model, I employed categorised variables to fit the model. Also, I fitted the Cox model separately for age with three categories and age with four categories and obtained survival curves for both models. These plots will show that even if the age and wealth variables are divided into additional groups, the survival curves in the plots will still be in the same order. That demonstrates that as one's age grows, the probability of survival decreases, yet as one's wealth increases, the likelihood of surviving improves. The Akaike information criterion (AIC) method helps in determining how well these models fit the data. The cox model for stratified gender with stratified continuous age and categorised wealth gets the lowest AIC score of all the models when compared. This demonstrates that this model fits better than others.

Bibliography

[ABM01] Kenneth Rockwood Arnold B. Mitnitski, Alexander J. Mogilner. Accumulation of deficits as a proxy measure of aging. *TheScientificWorld*, 2001.

[AS12] James Banks James Nazroo Andrew Steptoe, Elizabeth Breeze. Cohort profile: The english longitudinal study of ageing. *International Journal of Epidemiology*, 2012.

[CAct] Iliffe S Rikkert MO Rockwood K Clegg A, Young J. Frailty in elderly people. *published correction appears in Lancet*, 2013 Oct.

[Cam19] Liberato Camilleri. History of survival analysis. *The Sunday Times of Malta*, 24/3/2019.

[Cha16] Fong Chun Chan. The basics of survival analysis. *Statistics, Survival Analysis*, May 12, 2016.

[FES19] Matthias Dehmer Frank Emmert-Streib. Introduction to survival analysis in practice. 2019.

[GM10] Kishore J Goel MK, Khanna P. Understanding survival analysis: Kaplan-meier estimate. *Int J Ayurveda Res*, 2010.

[Kar16] Christiana Kartsonaki. Survival analysis. *Diagnostic Histopathology*, July 2016.

[pad16] Graph pad. Hazard ratio from survival analysis. *KNOWLEDGEBASE - ARTICLE 1226*, March 16, 2016.

[SB18] Alexandra Crosswell Ashley Lin Drystan Phillips Marieta Valev Jenny Wilkens Victoria Yonter Jinkook Lee Sidney Beaumaster, Sandy Chien. Harmonized elsa documentation. November 2018.

[Sch19] Daniel Schütte. Survival analysis in r for beginners. *Datacamp*, December 17th, 2019.

[SPS+15] John Soong, Alan Poots, S Scott, K Donald, Thomas Woodcock, Derryn Lovett, and Derek Bell. Quantifying the prevalence of frailty in english hospitals. *BMJ Open*, 5, 10 2015.

[TF16] Yee Whye Teh Tamara Fernandez, Nicolas Rivera. Gaussian processes for survival analysis. *Centre Convencions Internacional Barcelona*, Dec, 2016.

[Tur14] Gill Turner. Cga in community settings, frailty. 11 June 2014.

[Wai01] Howard Wainer. Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology. Vol. 52: 305-335*, 2001.

Appendix A

Appendix

3) Dataset (Coding)

Python Code:

```
#creating columns ( status, time, age, wealth)
for index, row in gen_min_py_d.iterrows():
    if(gen_min_py_d.loc[index,'frailll1'] > 0.15 and gen_min_py_d.loc[index,
      'frailll1'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r1agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h1atotb']
        #cox_d_py.loc[index,'fraill_s'] = gen_min_py_d.loc[index,'frailll1']
        if gen_min_py_d.loc[index,'frailll2'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r2agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h2atotb']
            continue
        elif gen_min_py_d.loc[index,'frailll3'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 4
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r3agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h3atotb']
            continue
        elif gen_min_py_d.loc[index,'frailll4'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 6
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r4agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h4atotb']
            continue
        elif gen_min_py_d.loc[index,'frailll5'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 8
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r5agey']
```

```
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h5atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill6'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 10
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r6agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h6atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 12
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 12
                continue
            elif gen_min_py_d.loc[index,'fraill6'] != 0:
                cox_d_py.loc[index,'time'] = 10
                continue
            elif gen_min_py_d.loc[index,'fraill5'] != 0:
                cox_d_py.loc[index,'time'] = 8
                continue
            elif gen_min_py_d.loc[index,'fraill4'] != 0:
                cox_d_py.loc[index,'time'] = 6
                continue
            elif gen_min_py_d.loc[index,'fraill3'] != 0:
                cox_d_py.loc[index,'time'] = 4
                continue
            elif gen_min_py_d.loc[index,'fraill2'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill2'] > 0.15 and gen_min_py_d.loc[index,
'fraill2'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r2agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h2atotb']
        if gen_min_py_d.loc[index,'fraill3'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r3agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h3atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill4'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 4
```

```
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r4agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h4atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill5'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 6
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r5agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h5atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill6'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 8
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r6agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h6atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 10
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 10
                continue
            elif gen_min_py_d.loc[index,'fraill6'] != 0:
                cox_d_py.loc[index,'time'] = 8
                continue
            elif gen_min_py_d.loc[index,'fraill5'] != 0:
                cox_d_py.loc[index,'time'] = 6
                continue
            elif gen_min_py_d.loc[index,'fraill4'] != 0:
                cox_d_py.loc[index,'time'] = 4
                continue
            elif gen_min_py_d.loc[index,'fraill3'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill3'] > 0.15 and gen_min_py_d.loc[index,
'fraill3'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r3agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h3atotb']
        if gen_min_py_d.loc[index,'fraill4'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r4agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h4atotb']
```

```
            continue
        elif gen_min_py_d.loc[index,'fraill5'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 4
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r5agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h5atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill6'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 6
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r6agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h6atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 8
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 8
                continue
            elif gen_min_py_d.loc[index,'fraill6'] != 0:
                cox_d_py.loc[index,'time'] = 6
                continue
            elif gen_min_py_d.loc[index,'fraill5'] != 0:
                cox_d_py.loc[index,'time'] = 4
                continue
            elif gen_min_py_d.loc[index,'fraill4'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill4'] > 0.15 and gen_min_py_d.loc[index,
'fraill4'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r4agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h4atotb']
        if gen_min_py_d.loc[index,'fraill5'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r5agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h5atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill6'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 4
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r6agey']
```

```
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h6atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 6
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 6
                continue
            elif gen_min_py_d.loc[index,'fraill6'] != 0:
                cox_d_py.loc[index,'time'] = 4
                continue
            elif gen_min_py_d.loc[index,'fraill5'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill5'] > 0.15 and gen_min_py_d.loc[index,
'fraill5'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r5agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h5atotb']
        if gen_min_py_d.loc[index,'fraill6'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r6agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h6atotb']
            continue
        elif gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 4
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 4
                continue
            elif gen_min_py_d.loc[index,'fraill6'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill6'] > 0.15 and gen_min_py_d.loc[index,
'fraill6'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
```

```
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r6agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h6atotb']
        if gen_min_py_d.loc[index,'fraill7'] >= 0.25:
            cox_d_py.loc[index,'status'] = 2
            cox_d_py.loc[index,'time'] = 2
            cox_d_py.loc[index,'age_f'] = cox_d_py.loc[index,'r7agey']
            cox_d_py.loc[index,'wealth_f'] = gen_min_py_d1.loc[index,'h7atotb']
            continue
        else:
            if gen_min_py_d.loc[index,'fraill7'] != 0:
                cox_d_py.loc[index,'time'] = 2
                continue
        continue
    elif (gen_min_py_d.loc[index,'fraill7'] > 0.15 and gen_min_py_d.loc[index,
'fraill7'] < 0.25):
        cox_d_py.loc[index,'status'] = 1
        cox_d_py.loc[index,'time'] = 0
        cox_d_py.loc[index,'age_v'] = cox_d_py.loc[index,'r7agey']
        cox_d_py.loc[index,'wealth_v'] = gen_min_py_d1.loc[index,'h7atotb']
        continue
```

4.1 Kaplan-Meier (Coding)

```
#K-M model: R code
km.model <- surv_fit(Surv(time,status) ~ 1, data = cox_fd)
#ask for summaries of the model
km.model
summary(km.model)

#To pllot
ggsurvplot(km.model,
           pval = F, conf.int = F,
           risk.table = TRUE, # Add risk table
           risk.table.col = "strata", # Change risk table color by groups
           surv.median.line = "hv",title = "KM-Model without x varibles",
           xlab = "Time (Year)",break.time.by=2) # Specify median survival)

#test: LOG-RANK-TEST
#   H0 : survival in two groups is same
#   Ha : surv not
survdiff(Surv(time,status) ~ ragender,data = cox_fd)
```

4.2 Cox Propportional-Hazard Model (Coding)

```
#Cox model - R code

cox.mod1 <- coxph(Surv(time,status) ~ ragender, data = cox_fd)
#summary
summary(cox.mod1)

#   CHECKING PROPORTIONAL HAZARDS ASSUMPTION
# H0 : HAZARDS are PROPORTIONAL   Ha : HAZARDS are not PROPORTIONAL

cox.zph(cox.mod1)

# Plot

ggsurvplot(surv_fit(cox.mod1.2, data = cox_fd), data = cox_fd,
           ggtheme = theme_minimal(), conf.int = TRUE,risk.table = TRUE,
           surv.median.line = "hv"
           ,risk.table.col = "strata", title = "Cox model for Gender:",
           xlab = "Time (Year)", legend.title = "Sex",
           legend.labs = c("Male", "Female"))

# CHECKING LINEARITY
# Check for LINEARITY using Martingale
plot(predict(cox.mod8), residuals(cox.mod8,type = "martingale"),
     xlab = "fitting values",
     ylab = "Martingale residuals", main = "Residual Plot", las = 1)
abline(h=0)
lines(smooth.spline(predict(cox.mod8), residuals(cox.mod8,type = "martingale")),
      col="red")


# AIC test: to find the best model.
#model which ever has the low AIC score is the best model
AIC(cox.mod9.1)
AIC(cox.mod9.2)


###### Likehood test

# cox.mod1.2 ---- strata(ragender)
# cox.mod14 ---- strata(ragender) + wealth_vc
anova(cox.mod1.2,cox.mod14, test = "LRT")
```