Survival Analysis

Patrik Dennis

January 22, 2025

1 Authors note

Due to the study and therefore data being about peoples lives, the author of this rapport would like to briefly acknowledge the lives that were lost during the study. The data and assignment itself has been handled with the upmost respect for the lives behind the study. Acknowledgement is also due to the fact that mathematics can save lives.

Contents

2 Introducing the Data

In this lab the Worcest Heart Attack Study (Hosmer and Lemeshow, 2008) will be used to conduct survival analysis. The data presented comprises from 500 subject with a variety of variables. The variables that are present in the data set. The variables that will be of importance for the assignment at hand will be,

- Lenfol: Length of follow-up stored in days. Lenfol measurement is terminated either by death or censoring. These two events are the outcomes of the study.
- Fstat: Censoring variable i.e the indicator for what type of outcome the participant had. The values for Fstat are: loss to follow-up **0**, death **1**
- Age: Age of participant when hospitalized.
- BMI: Body mass index.
- HR: Initial heart rate of participant.
- Gender: gender of participant; males **0** and females **1**.

Note There are some observations i.e. participants that do not have information on the day of death after a heart attack took place.

3 Preliminaries

3.1 Survival Analysis Packages in R

For the study at hand there are certain packages in R that are required. The required packages are,

- Tidyverse: Enables one to produce summary statistics tables.
- GGally: Extension to ggplot2 (data visualization package) which enables one to produce similar plots to ggplot2.
- Survival: Provides one with function for Survival Analysis.
- Survminer: Enables one to produce plots for Survival Analysis. This is also an add-on to ggplot2.

3.2 Distributions in Survival Analysis

3.2.1 Plotting the distribution of the survival times

To plot the distribution of the survival time one can provide,

```
ggplot(dt[dt$FSTAT==1,],aes(x=LENFOL))+geom_histogram(col="white")
```

The above filters the data such that the data observations of Fstat value 1 is present,i.e death as outcome. After providing the relevant data frame to be plotted, the x axis is provided, which in this case is the days of survival (Lenfol). One then provides the type of plots to be produced and what the y-axis should be. In this case the y-axis is the density with both histograms and the density being present. The plot provided by the previous code is presented below.

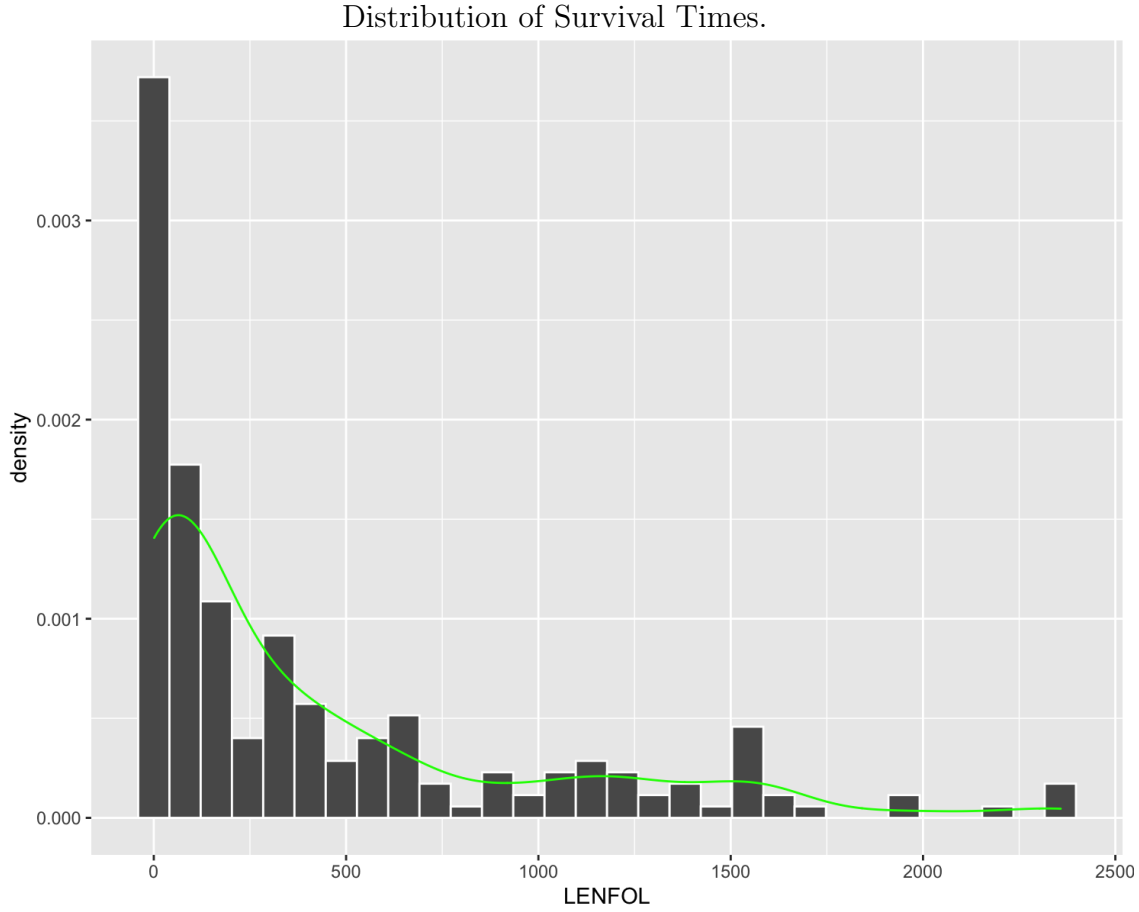


Figure 1: Plot showing the distribution of survival times. The density curve is also present in green.

From the above one can observe that the density curve is approximately exponential. This can be seen as the values are non-negative, and that the density decreases in an exponential fashion. This is further discussed in later chapters where one can observe the hazard function. In these situations it is reasonable for the survival function to be exponential due to the loss of memory property. The loss of memory property is present in this case, due to the probability of failure being the same for any time period no matter the age.

3.3 Finding the ECDF

It should be noted that the CDF produced in this assignment is naturally the emperical cumulative density function, denoted $\hat{F}_n(t)$, is provided by,

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}.$$

where $\mathbf{1}_{\{X_i \leq t\}}$ is the indicator, taking on the value 1 if $X_i \leq t$.

One can produce a visualization of the ecdf by,

```
ggplot(dt[dt$FSTAT==1,],aes(x=LENFOL))+stat_ecdf()
```

returning the plot,

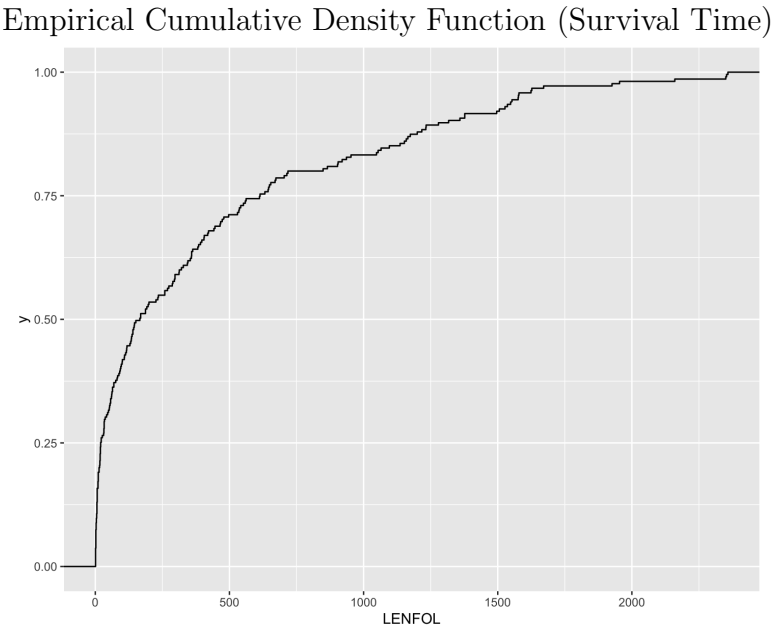


Figure 2: Plot showing the ecdf of the surival time.

In order to find the probability of surviving the first 200 days, one must note that the ecdf computes and provides the probability of dying up to a time unit. This implies that one must take the complement of the ecdf, i.e,

$$1 - P(T_i \leq t).$$

The above will provide one with the probability dying after day t. This results in the code below giving the probability of surviving the first 200 days.

```
P = ecdf(x=dt$LENFOL[dt$FSTAT==1])
1-P(200)
```

The above prints the result 0.4651163 which means that the probability of surviving the first 200 days are approximately 46.512%. The formula to find the probability of survival further than a certain amount of days can be defined as the Survival function $\hat{S}(t)$. This meaning that the defintion of the survival funciton is,

$$\hat{S}(t) := 1 - P(T_i \leq t) = 1 - \hat{F}(t).$$

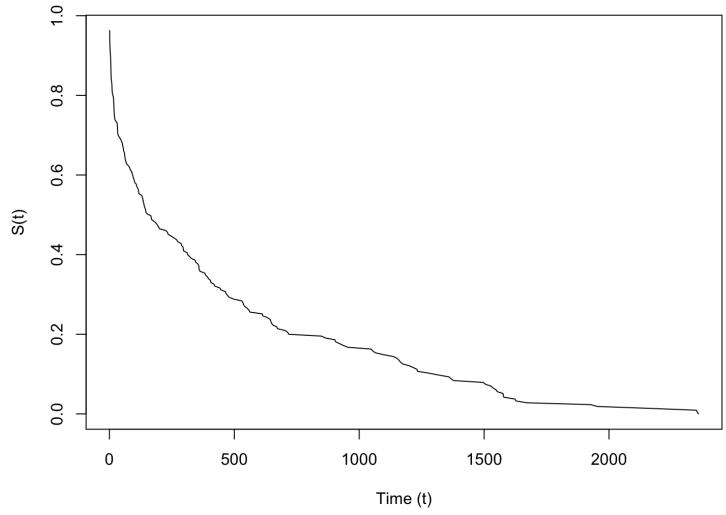
The above shows that the survival function was used to compute the probability of survival up to 200 days. The survival function can also be used to find the probability of surviving further than a specified time unit, in this case days. To find the probability of surviving more than 1000 days, one can provide,

```
S_{T}(t) = S(t>T) = (1-P(1000))
```

resulting in 0.1674419, i.e. approximately a 16.74419% chance of survival. The reason for again taking the complement of the survival function ($1 - \hat{S}(t)$). To further analyze and gain knowledge of the survival function $\hat{S}(t)$ one can plot the function. This is done by,

```
plot(x=sort(data1), s_emp, type="l")
```

returning the plot,



Plot of $\hat{S}(t)$

Figure 3: Plot of survival function $S(t)$.

From the plot above, one can see that the computed probabilities are correct with the plot.

4 Survival Analysis

4.1 Exploratory Data Analysis (EDA)

Note In this section there is code that is used multiple times to provide descriptive statistics on the chosen variable. In order to not be repetitive the skeleton code is shown below. There also exists an example of how the code is used under the Exploring Lenfol section.

```
dataframe %>% summarise(N=n(),Mean=mean(name.of.variable),SD=sd(name.of.
    variable),Min=min(name.of.variable),Max=max(name.of.variable))
```

In the skeleton code above, the variable dataframe is the data frame where the desired variable to be analyzed is stored. The name.of.variable is the variable one wishes to investigate and provide descriptive statistics about.

4.1.1 Exploring Lenfol

To further understand the data, one can provide the **summarise()** function. If one runs,

```
dt %>% summarise(N=n(),Mean=mean(LENFOL),SD=sd(LENFOL),Min=min(LENFOL),
    Max=max(LENFOL))
```

returning the values,

N:	Mean:	Standard Deviation:	Min:	Max
500	882.436	705.6651	1	2358

Table 1: Results from providing summarise() with the variable LENFOL.

The table above shows that the number of samples, in this case participants, is 500. The mean time of participation in days is 882.436. As stated in the introduction the time of participation is stopped either by not showing up or death. This surmounts to being right-censored. The mean seems reasonable if one observes the histogram **??**, where the distribution of survived days are heavily concentrated below the 800 day period. This would imply that the mean should be lower. This is however not the case when further looking at the distribution of values beyond 800. The small but nevertheless existent values beyond the 800 day time-period, skews the mean further to the right, i.e. to be larger. This implies and concludes why the mean is greater than what one would guess (for a student).

Due to the variation of values in Figure 1, one could presume that the standard deviation would be large. This presumption would be correct. The presumption of a large standard deviation could also be drawn by the shifting of the mean by the existent but few large values for time. The standard deviation of the number of days participated by the subjects is 705.6651, which is large for the interval of time measured, as presumed. Previously the interval of time measured was used as a relative measure for the so to speak extremeness of the standard deviation. The time interval can be seen from the minimum and maximum values the time variable has which is shown in the table under min and max. The interval is therefore [1,2358].

4.1.2 Exploring Age

To show the descriptive statistics of the variable, age, one uses the skeleton code 4.1.1, now with AGE as the name.of.variable. Providing the previous returns the values,

N:	Mean:	Standard Deviation:	Min:	Max
500	69.846	14.49146	30	104

Table 2: Results from providing summarise() with the variable AGE.

From the above table one can observe that the mean age is approximately 70 years. This can be considered reasonable due to the studying heart attacks, which usually occurs during the later stages of ones life. When however looking at the minimum and maximum value for the age, one sees that there is still a presence of a variety of ages. The observation of a large interval of ages can be used to prove that the criteria stated in the protocol on recruiting subjects of various ages is not missed.

The standard deviation of approximately 14.5 years can be considered a reasonable of the test, due to the mere fact that, as stated before, heart attacks most commonly occurs during the later stages of ones life. One could however argue that the values that are 2 standard deviations away could be considered outliers. In this situation a person in their thirties may be considered outliers. This however can cause later clinical trials to lose focus on younger individuals when testing treatment in heart disease. This can happen due to the extensive research that goes into designing a clinical trial. An example of research that is gathering prior data and results from similar studies.

The exclusion of age groups and specifications a person can take is however never a problem, due to the strict guidelines when providing criteria during the protocol design phase. The argument stated in the previous paragraph is to provide an extreme case of what could happen. Human error is unfortunately one of the very characteristics that define us. This should therefore be taken into consideration.

4.1.3 Exploring BMI

The skeleton code is yet again used for the BMI variable which provides the values,

N:	Mean:	Standard Deviation:	Min:	Max
500	26.61378	5.405655	13.04546	44.83886

Table 3: Results from providing summarise() with the variable BMI.

From the table above one can observe that the mean of approximately 26.61 is reasonable due to a person that can be considered moderately healthy having a bmi of 18.5-24.9. [1]

Observing the standard deviation one can conclude that the standard deviation seems reasonable as well, due to extremes i.e. obesity and underweight lying above 30 and below 18.5. [1]

4.1.4 Exploring HR

Using the skeleton code now for the variable HR, one get the results below.

N:	Mean:	Standard Deviation:	Min:	Max
500	87.018	23.58623	35	186

Table 4: Results from providing summarise() with the variable BMI.

From the table, one can observe that the mean is approximately 87.02 which is considered an average resting heart rate. [2]

The standard deviation of approximately 23.59 is understandable, due to the fact that the majority of subjects are not normal in the sense that they have experienced heart issues, implying that the subjects heart functionality may be weaker than an average healthy non-participating individual.

4.2 EDA using Pairplots

4.2.1 Pair plot analysis non-deceased

To provide a pairplot for variables: lenfol, gender, age, hr and bmi, for respective outcome, i.e. non-deceased, one can provide the code below.

[1]Center for Disease Control and Prevention. About adult BMI. *U.S. Department of Health Human Services.* 2021. https://www.cdc.gov/healthyweight/assessing/bmi/adult$_b$*mi/index.html*(*Sourced*2022 − 02 − 08).

[1]Ibid

[2]British Heart Foundation. What is a normal heart rate? *British Heart Foundation.* 2021. https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/ask-the-experts/pulse-rateHeading1 (Sourced 2022-02-08).

```
1
2 dt_survived <- dt[dt$FSTAT==0,]
3
4 ggpairs(dt_survived[,c("LENFOL","GENDER","AGE","BMI","HR")])
```

resulting in the figure below.

Pair plot for non-deceased subjects

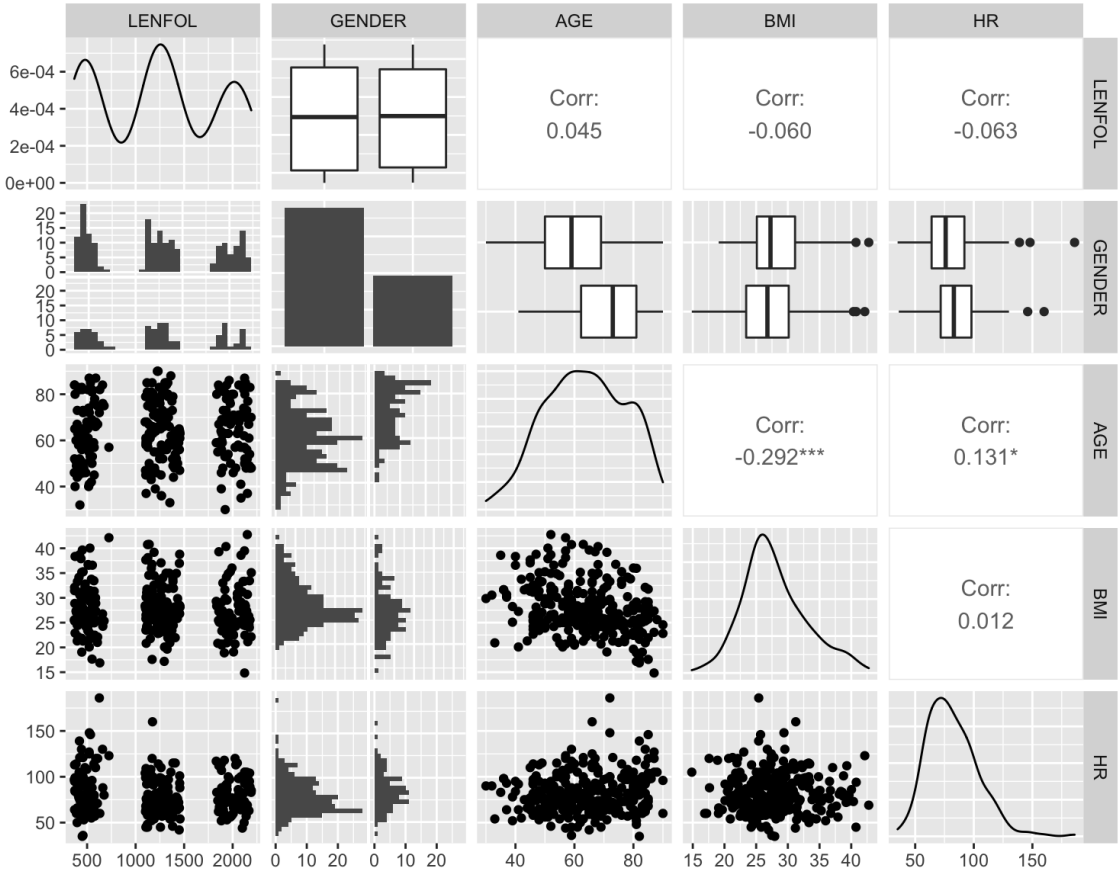


Figure 4: Pair plot of the variables lenfor, gender, hr, age and bmi for the non-deceased subjects.

From the above plot one can observe that the distribution of the subjects participation periods are in 3 distinctive groups or compartments. One would assume that the distribution of a non-deceased subject would build a vertical line at the time the study was terminated. This is however not the case, due to the study experiencing subjects that fail to follow up. From the group of 3, there is are three distinct

periods, where the subjects failed to show up. One could reason that it might be due to a certain testing frequency the study had, e.g. testing once per year or every day etc. If the test was to occur once per year the participants are surely more likely to forget about the study than a study that tests every day. On the other hand, one could also argue that the distribution of events, could be due to a more frequent testing schedule. This being because the subjects may have gotten tired of testing etc.

Another enlightening piece of the plot is the bar plot for the lenfol variable, which is stratified into gender. One can see from the above, that the median is approximated the same with roughly the same quantiles. One can however see that the correlation between age and lenfol is small, meaning that the time until loss to follow up is not informed or dependent on the age. If one was to go further into analysing gender, one can observe that the bmi and hr for the respective genders have roughly the same median. A difference can be seen in the quantiles. If one however observes the age box plots, one sees that the median for one of the gender is distinctively higher than the other, with the other median lying around the 25% quartile for the other gender. Further looking at gender, one also sees that there exists outliers regarding BMI and HR. These outliers could be considered as initially more ill individuals. One can also observe that the bmi with respects to age has a high variance, where the distribution of the data point are spread widely over the plot.

Lastly when looking at hr and age it is clear how the standard heart rate of approximately 60-100 is considered normal.[3] One can see from the scatter plot that the majority of the subjects have a heart rate approximately lower than 100.

4.2.2 Pair plot analysis deceased

The procedure to provide a pair plot for the deceased subjects is analogous to the previous,

```
dt_died <- dt[dt$FSTAT==1,]

ggpairs(dt_died[,c("LENFOL","GENDER","AGE","BMI","HR")])
```

resulting in the plot below.

[3] Ibid

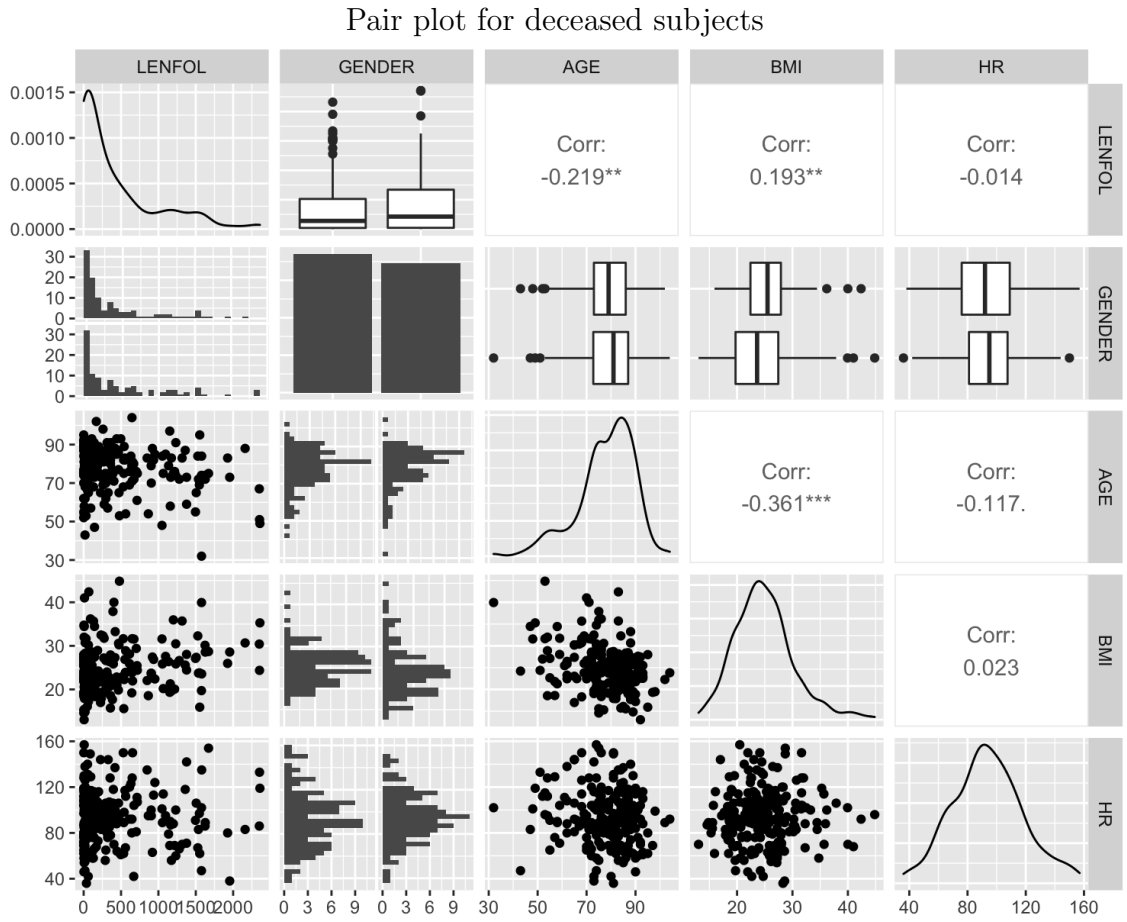


Figure 5: Pair plot of the variables lenfor, gender, hr, age and bmi for the deceased subjects.

From the plot above one can see that the distribution of days to event and age show that the amount of days passed until event is lower for the older the subjects. One can note that the scatter plots of day until event shows a concentrated cluster to the left of all three plots. This shows that if one were to have an event (death) during the study, it was more likely to occur in the early stages (approximately before 500 days). After approximately 500 days the data points progress its sparsity and decreases in quantity. One could reason from the hr and lenfol scatter plot that the lower a subjects heart rate, the longer one was to survive. This is due to the larger amount of data points below 80 that surpass a 1500 survival threshold, whereas there are is scarce amount of subject that survive past 1500 days with a heart rate of approximately above 90. This seems reasonable, due to a higher heart rate making one more liable to have a heart attack.

4.2.3 Comparing the pair plots

Now that an undergone analysis of the respective pair plots, the time is nigh for a comparison. From the lenfol variable and the respective scatter plots for the other variables, one can see that the distribution of the non-survivors are not grouped. This simply being that the non-survivors cannot decide when not to die. When comparing the distribution of gender for the non-survivors, one can observe that the distribution is roughly even, with approximately 30 subjects for the respective gender. This however is not the case for the survivors, where there exists a distinct difference in distribution of the respective gender.

One can also observe that there appears to be more variance when looking at heart rate with respect to age for the non-survivors, meaning that their was a larger spread of heart rates with respect to age. This can be due to non-deceased subject generally having a lower heart rate, implying that the risk of death is lower.

4.3 Non-parametric Analysis

To fully appreciate the results from the plots and value below, one must first be provided with the foundations of the Kaplan-Meier estimate of the survival function.

4.3.1 Survival Function with Kaplan-Meier

Since evaluating the true distribution of the survival time is cumbersome, one can use a non-parametric method estimate the survival function. The non-parametric method chosen is the Kaplan-Meier.

Definition: Kaplan-Meier Estimation

$$S(\hat{y}) = \prod_{y(k)<y} \left(1 - \frac{d_k}{n_k}\right).$$

The definition above is the Kaplan-Meier estimation of the survival function. To first compute the estimation one must order the distinct failure times $y_{(1)} < ... < y_{(n)}$. Now one defines d_k as the number of events at time y_k, i.e amount of deaths or withdrawals at time $y_{(k)}$. Further m_k is defined as the number of censored observations in the interval $(y_{(k)}, y_{(k+1)}$, with $k = 1, ..K$. In order to provide the risk of an event occurring, denoted n_k, one takes the amount of subjects that have survived infinitesimally prior to y_k and not having a survival time censored before this time. Due to independence regarding censoring, one can now define n_k,

$$n_k = (d_k + m_k) + \cdot + (d_k + m_k), k = 1, ..., K.$$

This meaning that the number of subject in risk just prior to the time y_k is denoted by n_k. Hence the Kaplan-Meier estimate interprets the probability that a subject is alive at y_k is equal to the doncitional probability that the subject is alive at y_k, given that the subject survived through all the proceeding time points $y_{(1)}, ..., y_{(k-1)}$.

4.3.2 Variance of Kaplan-Meier Estimate

Now that the estimate is defined, one can proceed in deriving the variance. This is needed to fully appreciate the results of the confidence intervals given later on in this section. Greenwood's formula,

$$Var(\hat{K}) \approx K^2 \sum \frac{1 - \hat{p}_t}{N_t \hat{p}_t} \approx \\ \approx \hat{K}^2,$$

is used to find the variance. Note that the assumption that of the random variable X_i as the amount of survivors for y_{k-} to y_{k+}, is binomial with X_i $Bin(n_t, p_t$. If one is to further presume that the $n_t p_t$ are large, thus resulting in the greenwood formula above.

With Greenwoods formula the variance of the Kaplan-Meier estimate arises,

Defintion: Variance of Kaplan-Meier estimate.

$$v[\hat{S}(y) = [\hat{S}(y)]^2 \prod_{y(k)<y} \left(\frac{d_k}{n_k(n_k - d_k} \right).$$

The above enables one to find and therefore compute the standard error of the Kaplan-Meier estimate.

The above then leads to Confidence Interval for the Kaplan-Meier estimate to be,

Kaplan-Meier Confidence Interval

$$\hat{S}(y) \pm Z(\alpha/2) \sqrt{v[\hat{S}(y)]},$$

resulting further in the estimated median survival time to be,

$$q_{0.5} = min\{y : 1 - \hat{S}(y) \geq 0.5\}.$$

4.4 Plots of the Kaplan-Meier Estimate

When providing the first Kaplan-Meier estimate,

```
surv.data <- with(dt,Surv(LENFOL, FSTAT))
fit1 <- survfit(surv.data~1,data=dt) #Median CI based on var(log(S(t)))
fit1
```

one is provided with the results below.

Summary Statistics for Fit 1

n	n events	median	0.95 LCL	0.95 UCL
500	215	1627	1527	NA

Table 5: Median Confidence Interval based on $var(\log(\hat{S}(t))$

.

From the above one can observe that the upper bound for the confidence interval has the value NA. One reason for this is due to the data being skewed. This is also due to the upper limit never dropping to 50%. This is seen in the figure below.

Figure 6: Confidence interval for the Kaplan-Meier estimate. Grey bands show the upper and lower limits of the confidence interval.

As seen from the figure above and discussed above, that the upper limit bands never drop to 50%, thus not being able to compute an upper limit for the median confidence interval.

To remedy the missing upper limit predicament, one can compute the confidence interval based on $var(log(-log(S(t))))$. This will shift the the interval given above downwards. This can be seen from the figure below.

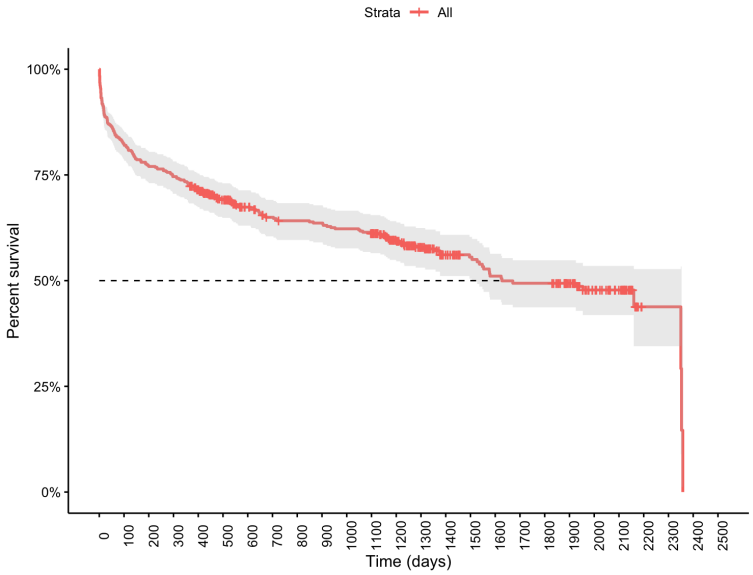


Figure 7: Caption

When comparing the two plots, one can observe how the upper limit for the confidence intervals for the survival percentage at a time unit t has been shifted downwards, now enabling the upper bound for the confidence interval at the median time, resulting in the table below. The log-log approach also makes the quantity unrestricted, thus making the confidence interval to be within the proper range when transforming back. The meaning of proper range, is based on the first approach, using Greenwood's formula which can result in upper limits above 1 and lower bounds below 0. The log-log approach therefore remedies this.

Summary Statistics for Fit 2

n	n events	median	0.95 LCL	0.95 UCL
500	215	1627	1506	2353

Table 6: Median Confidence Interval based on $var(log(-log(S(t))))$

.

4.4.1 Hazard Function

The hazard function, denoted $h(y)$, provides the instantaneous death rate. The probabilistic definition of the hazard function is the conditional probability that a subject fails over the next instant given that the subject has survived up to the beginning of the interval.

Definition Hazard Function.

$$\frac{f(y)}{1-F(y)} = \frac{f(t)}{S(y)}.$$

4.4.2 Kaplan-Meier Hazard- & Survival Plots

With the definition of the Kaplan-Meier estimate of the survival function and the needed data to compute the function, one can provide the needed data by providing the code below.

```
d <- data.frame(time = fit1$time,
                n.risk = fit1$n.risk,
                n.event = fit1$n.event,
                n.censor = fit1$n.censor,
                surv = fit1$surv,
                failure = 1-fit1$surv,
                se = fit1$std.err,
                upper = fit1$upper,
                lower = fit1$lower)

d$n.failed <- cumsum(fit1$n.event)
d$n.left <- fit1$n - d$n.failed

#Plot survival curve
ggsurvplot(fit1,
           risk.table = TRUE,
           surv.median.line = "hv",
           ggtheme = theme_bw())

#Plot cumulative hazard curve
ggsurvplot(fit1,
           fun = "cumhaz",
           risk.table = TRUE,
           ggtheme = theme_bw())
```

From the first segment of the above code, one can see how the number of events, censors and number of subjects at risk at some time unit is defined. The survival data is then computed and can be retrieved as a data frame. To get a grasp of the computed survival values and therefore survival function one can retrieve the figure below.

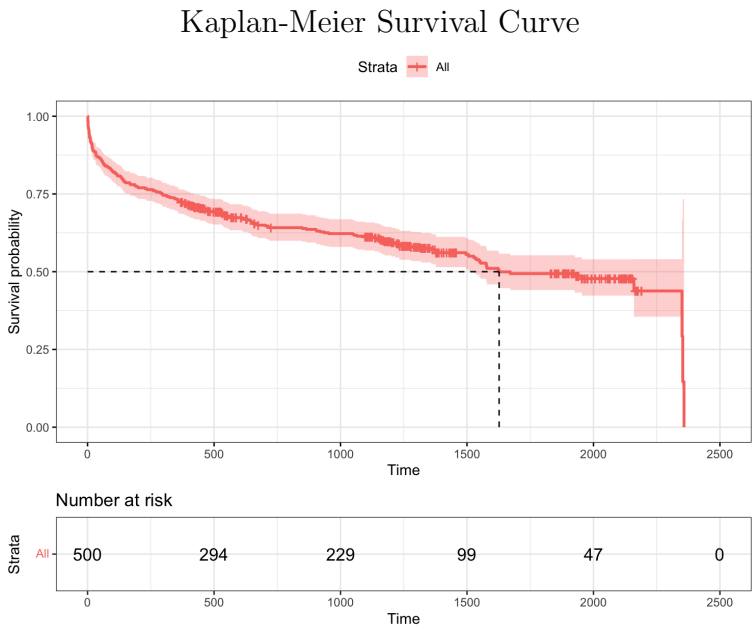


Figure 8: Survival Curve computed with Kaplan-Meier.

As seen from the figure, the number of subjects at risk just prior to a time point y_k is given below the plot of the survival function. To further investigate the results provided by the Kaplan-Meier estimate, one can retrieve the hazard functions plot as well.

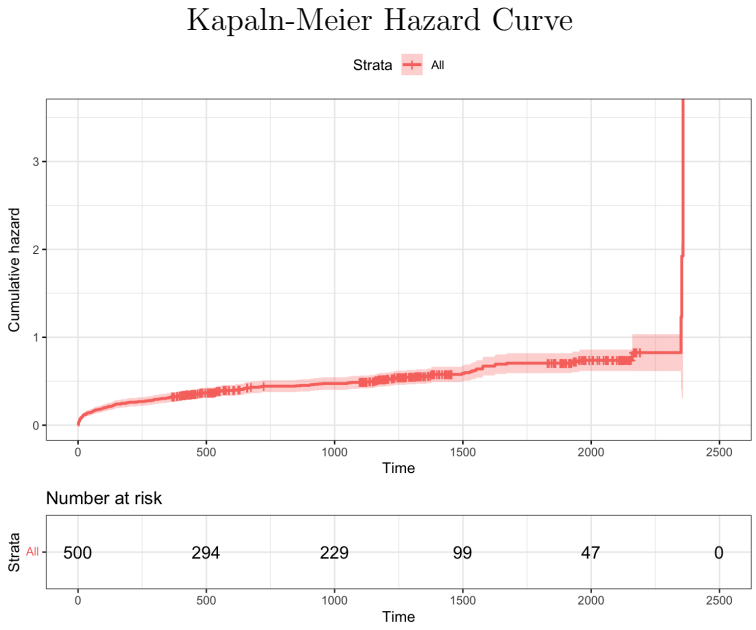


Figure 9: Hazard curve computed with Kaplan-Meier.

When observing the hazard step-function one can see that function is decreasing in a monotone fashion, thus resulting in the instantaneous death rate increasing as time increases. This is also seen from the figure showing the survival estimate. From the figure it can be seen that the survival probability decreases with time, thus implicitly saying that the death rate will increase as time increases.

One can also argue that the hazard function is somewhat constant, with small increasing increments throughout the time period, thus resulting in a small increase over the full time period. Due to the constant nature of the hazard function, which provides information on the slope of the survival function, one can conclude that the survival function is memory-less. Memory-less, being that the life length does not regard the evolution of time. These properties further provide reasonable arguments for the survival function being exponential, with the parameter λ being the hazard value.

4.5 Mantel-Heaenszel test for Gender

In a clinical trial one might be interested in the differences of the efficacy levels or death-rates between for example genders. This results in one wanting to compare for example,

$$H_0 : S_{male}(y) = S_{female}(y) \tag{1}$$

vs

$$H_1 : S_{male}(y) \neq S_{female}(y). \tag{2}$$

One of many methods used to undergo such a test as the above is the logrank test. The logrank test is considered powerful, due to testing proportionality for the hazard functions,

$$H_0 : S_{male}(y) = S_{female}(y)$$

vs

$$H_1 : S_{male}(y) \neq [S_{female}(y)]^{\lambda}.$$

Under the assumption of a hypergeometric distribution, the conditional expected number of subjects with number of events for d_{1k} at the event time $y_{(k)}$ is given by,

$$e_{1k} = \frac{n_{1k} d_k}{n_k}$$

with variance of d_{1k} being,

$$v_{1k} = \frac{n_{1k} n_{2k} d_k (n_k - d_k}{n_k^2 (n_k - 1)}, k = 1, ..., K.$$

Due to the logrank test statistic being the same as the Mantel-Haenszel statistic, one can use the logrank test denoted,

$$X_{LR} = \frac{\left[\sum_{k=1}^{K} (d_{1k} - e_{1k}\right]^2}{\sum_{k=1}^{K} v_{1k}} = \frac{(d_1 - e_1)^2}{v_1}$$

with,

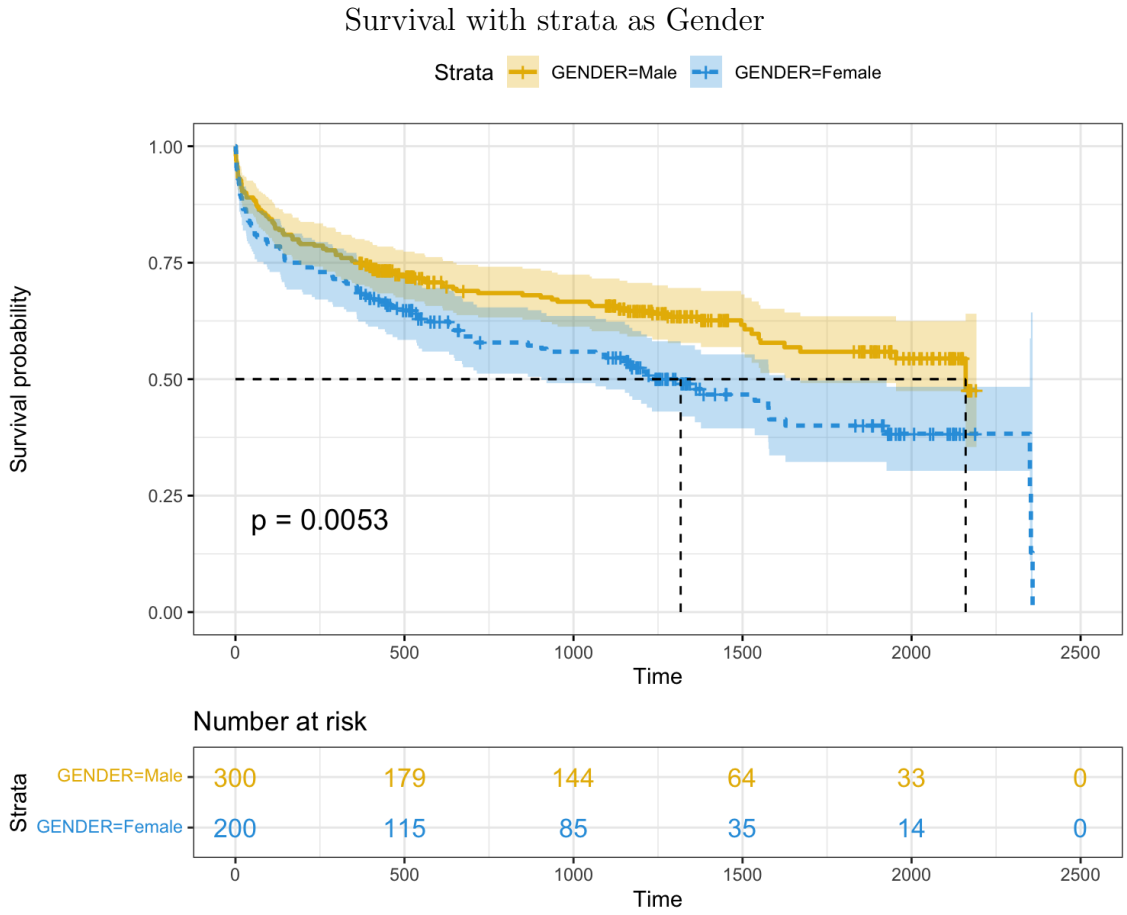
$$d_1 = \sum_{k=1}^{K} d_{1k},$$
$$e_1 = \sum_{k=1}^{K} e_{1k},$$
$$v_1 = \sum_{k=1}^{K} v_{1k}.$$

Further, as the logrank test statistic approximately following a central chi-square distribution with one degree of freedom, the null hypothesis is therefore rejected at the αth significance level if,

$$X_{LR} = \chi^2(\alpha, 1). \tag{3}$$

4.5.1 Results

Provided with the resulting survival functions for each gender respectively,



From the plot above, the upper limit of the 95% confidence interval for the median will be non-existent regarding males (yellow). This is, as discussed before, due to the upper limit not intersecting the 50% mark. From the above one can deduce that during the first stages the hazard function will increase more for females, this being due to the death rate being larger. As seen in the figure above. It is however notable that the confidence intervals overlap during the first stages, thus leaving a probability that the survival could be the same during those stages.

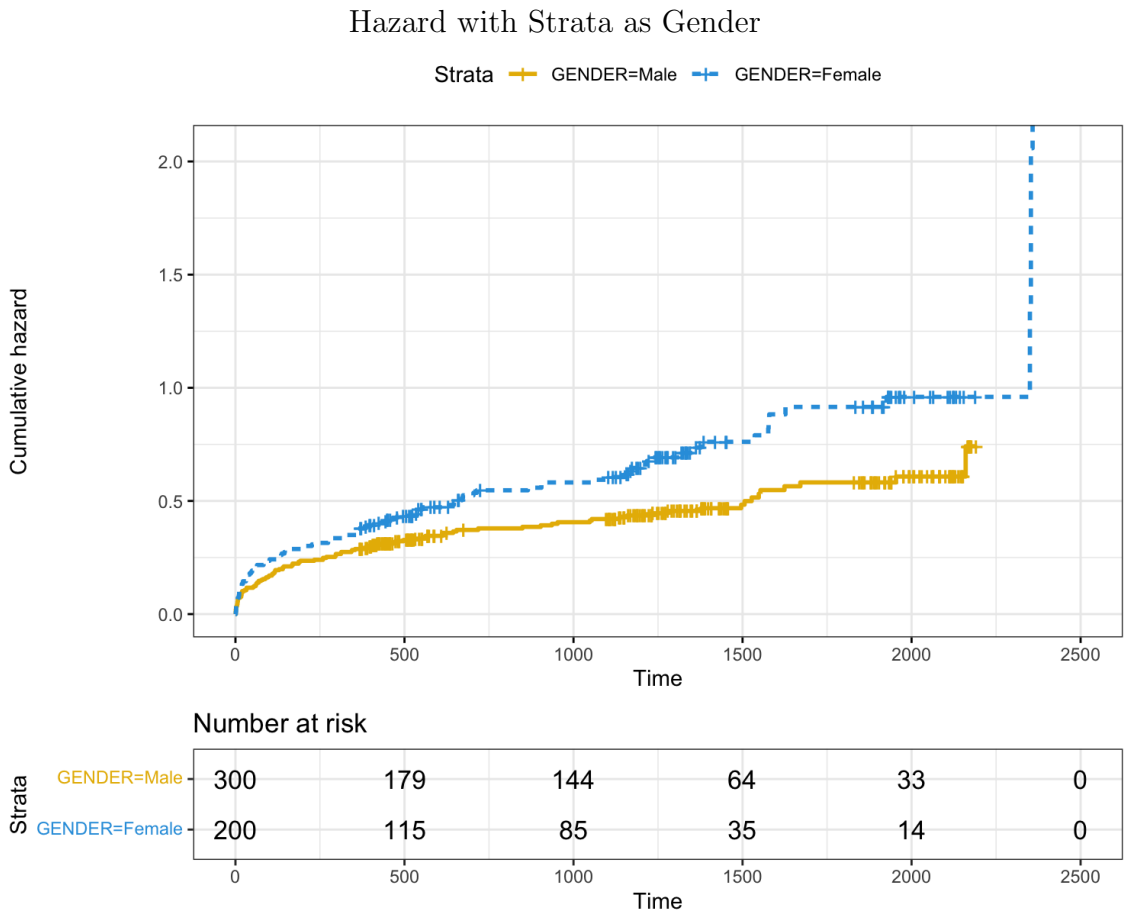
The figure also shows how the male function decreases abruptly at 2160. This value was first seen from the results of the maximum time evaluation for the males. That the maximum time period for the females is larger can be a result of many factors.

A possible factor is that, although the death rate is higher for the females, the rate levels out just after 1500. This meaning that the probability to survive after that time reaches an equilibrium, steady-state. This however may not be the case, where a simple explanation can be that the males withdrew earlier.

Another notable property in the figure that explains the maximum time period for the genders is the sudden drop for the respective genders. The figure however shows the females abruptly declining to 0, compared to the males that stop at a non-zero probability. This could show the right-censoring phenomenon, where the sudden decrease explains that the subjects for the respective subjects stopped attending the study. This could especially argue the case for the abrupt stop for the males, where the probability stops at a non-zero value.

One should also note that the number of females in the study are less than the males. This is further seen from the numbers at risk for the primary stage. As there are less females at the start, the same death number will decrease the probability to survive more for the females than the males. This meaning that, although the female death rate may be higher, the probability of survival can be misread as well.

To further investigate the death rate, the hazard functions for the respective genders is provided.



The figure above shows the hazard functions for the each gender, proving the hypothesis of higher death rates for females, as discussed above. Also seen from the figure is the female hazard function increasing for higher values. The figure also shows how the study stops, or the subjects fail to show up at an earlier stage compared to the females.

To further investigate the comparison between the survival function, the log-rank test is computed for the respective genders.

Log-Rank Test

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Gender: Male	300	111	130.7	2.98	7.79
Gender: Female	200	104	84.3	4.62	7.79

Table 7: Table showing log-rank test. **Chi-sqaure: 6.7**.

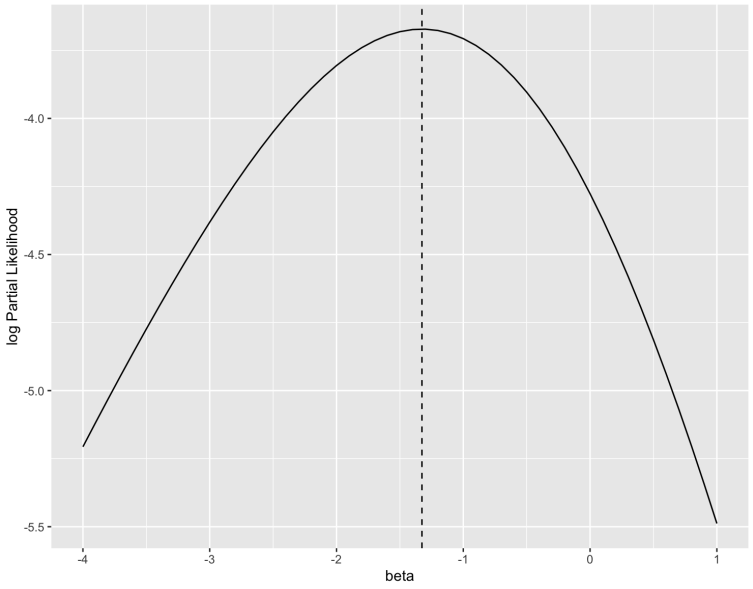
The table shows the log-rank test to be $X_{LR} = 7.9$ and the chi-square distribution being, $\chi^2 = 6.7$. From 3, the null hypothesis 1 can be rejected. This seems is reasonable, when observing the figure showing the survival functions for the genders.

4.6 Parametric Analysis with Cox Regression

Using the log-likelihood function,

$$l(\beta) = \beta \log(3 \exp \beta + 3) \log(3 \exp \beta + 1) \log(2 \exp \beta + 1),$$

the resulting plot for $l(\beta)$ is shown in the figure below



Log-likelihood Function

Optimizing the above function results in the values,

β	-1.326129
$l(\beta)$	-3.671981

Using the **coxph** to fit the proportional hazard regression model, the resulting β is equal to -1.3261. This is approximately equivalent to the β value produced by finding the maximum of the log-likelihood function. The difference is however that there now exists a standard error for the value. In this case the standard error is 1.2509. However the p-value for is 0.2715, thus resulting non-significance.

4.6.1 Cox Regression on Whas500

First Cox regression is performed with covariates of interest being age and gender. The interpretation of the model with two covariates of interest can be written as,

$$h(y; x_1, x_2) = h_0(y) \exp\{\beta_1 x_1 + \beta_2 x_2\}, \tag{4}$$

where X_1 represents the random variable for age and X_2 for gender. The β_1, in this case, measures the age after adjustment of difference in gender. The previous thus implies that β_2 reflects the effect of gender, adjusted for age. The above further provides the relative risk,

$$\lambda(x_1, x_2) = \exp\{x_1\beta_1 + x_2\beta_2\}. \tag{5}$$

After fitting the Cox regression model the following table is returned,

	Coef	$E(coef)$	se(coef)	z	$P(>\|z\|)$
Gender Female	-0.066285	0.935864	0.140585	-0.471	0.637
Age	0.066928	1.069218	0.006196	10.802	$< 2e-16$

	$E(-coef)$	$E(coef)$	lower 95%	upper 95%
Gender Female	0.9359	1.0685	0.7105	1.233
Age	1.0692	0.9353	1.0563	1.082

From the above, it can be seen at each time-point the hazard for dying is 6.4136% lower for females irrespective of age, and 6.9218% higher for the different age groups with correction for age.

The Wald,score and likelihood-ratio tests, test the following hypothesis,

- H_0: The reduced model being the true model,
- H_1: The reduced model not being the true model.

The Wald-test is computed on the weighted distance between the unrestricted estimate and the hypothesized value, with the weight being the precision of the estimate. The construction of the Wald-tests are,

$$W_i = \left[\frac{b_i}{se(b_i)}\right]^2 > \chi^2(\alpha, 1), \tag{6}$$

with b_i being the estimated β's for the different covariates and possibly the interaction terms i.e $\beta_{1_2} x_1 x_2$. Thus the confidence interval,

$$b_i \pm Z\alpha/2)se(b_i). \tag{7}$$

The Wald test does not test for interaction terms, meaning all parameters that are not common are set to 0. The Wald test then finds the difference between the computed test statistic against the assumed.

The likelihood-ratio test is computed by,

$$LR = -2ln(L_r - L_C), \tag{8}$$

with L_r being the reduced model, i.e. no interaction terms and L_C being the complex model with more variables in consideration. The likelihood-ratio test is therefore used to find the difference between the test statistic and the assumed statistic, stemming from the hypothesis.

The likelihood ratio tests, when one is considering multiple covariants and therefore interactions, do not include standard errors in the computation. When taking standard errors into account, such as the Wald score, the score can explode. This can occurs when there is complete separation in the model. This can be seen from 4.6.1, where Wald is computed with division of the standard error.Therefore if the standard error is small the score will explode.

The primary difference between the likelihood-ratio test is that the interaction terms are not included in the Wald test, but can be included in the likelihood-ratio test.

The Lagrange multiplier test (Score test), is computed by the slope of the likelihood function at the observed values of the included stratum only, i.e. not including other interaction terms. The formula for the Lagrange multiplier test is,

The score-test is used to determine whether an extra term, variable, should be included to improve the model fit.

The similarity for all tests, is that the conclusion made from the tests are the same. This being if the null hypothesis previously stated should be rejected or not. The main difference can occur for small sample sizes. When the sample size increases the test will converge to the same value.

The score for the three tests are provided below.

		df	p
Scoring	Likelihood ratio test: **142.4**	2	$\leq 2e-16$
	Wald test: **119.7**	2	$\leq 2e-16$
	Score (logrank) test: **126.9**	2	$\leq 2e-16$

From the table above, the p-values for all test are below 0.05. This means that the null-hypothesis is to be rejected, meaning that a larger model should be used with the chosen stratum. This could be achieved by including the interaction terms thus providing more β_i's.

4.7 Survival For Gender after Age correction

The plot below shows the survival function stratified by gender, correction for age.

Figure 10: Plot showing the survival function by gender after correction for age.

The plot above shows survival function by gender, after correcting for age. From the figure above, it can be seen that compared to **??** the differences between the step-functions are smaller. This can be explained when looking at table of coefficients, how the expected value for the coefficient both taking in the age and gender lowered the probability. This meaning that heavier changes in **??**, are due to the impact, weight, the age covariate has on the model. This was further seen from the scores provided in **??**, where the conclusion was to reject the simple model, with no interaction terms.

The figure above also shows how the gender did not play as big of role presumed when discussing **??**. This further enlightens the importance of investigating models after age correction.

4.8 Searching for a Better Model

4.8.1 Optimizing the Gender Age Model

First the Akaike information criterion can be measured for the model with the parameters, gender and age. The resulting Akaike measure is 2316.276. A reasonable

start could be to consider the rejection made from the score results. This meaning that a more complex model could be a better fit. This is achieved by including the interaction terms,

```
fit4 <- coxph(surv.data~GENDER+AGE+GENDER:AGE, data=dt)
```

resulting in the Akaike score 2312.443. From the results, it can be seen how taking the interaction terms in consideration optimized the model, meaning a better fitted model.

4.8.2 HR & Age Model

Covariates of interest could be the subjects heart rate and age. This is particularly interesting looking at the survival function for heart rate, when correcting for age. This being due to older subjects having a larger variance in heart rates, seen from 4 and 5. From the discussion on 5 (pair plot for deceased subjects), where the observation, higher heart rates may lead to event occurring quicker, was made. Such a model with parameters age and heart rate resulted in an Akaike score of 2298.431. This score is lower than the previous, thus rendering this model as a better fit.

The interaction terms can now also be introduced via the same procedure for the age and gender terms. The resulting Akaike score for was 2295.219, meaning that the model with interaction terms is deemed a better fit.

4.9 BMI & HR

Variables that could be considered interesting when observing survival of a subject, could further be other physical properties the subjects. Heart rate, as discussed above could very well be an indicator of ones probability of survival. The data set has however another variable which provides information on the individuals health, and therefore, maybe the survival of the subject. Fitting a model for BMI and HR, with no interaction terms results in the Akaike score of 2383.719. This result is therefore not the most optimal score found. If one was to take interaction terms into consideration, the Akaike score results in 2311.042. This results again in the more complex model performing better when looking at the goodness of fit. The best score for the BMI and HR model is however not the overall most optimal, being the HR and Age model.

4.10 Going for Gold

If one was to disregard computational power and time, one could consider the optimal model to be of a more complex nature. This means that one could fit and test a model for more than two covariates, with interaction terms. Such a model could be,

```
fit <- coxph(surv.data ~ BMI+GENDER+AGE+HR+(AGE:HR)+(GENDER:AGE)+(GENDER
    :BMI)+(HR:BMI), data=dt)
AIC(fit)
```

As seen above the BMI, GENDER, AGE, HR variables are all taken into consideration. Note that the interaction terms are also included, thus rendering this model to be complex. The resulting Akaike score for the above is 2291.986. This means that the best model, from a good of fitness perspective is the model above. In this model it can be seen that most variables regarding physical traits for the subjects is taken into consideration.

If one was to take reduce the model by retracting the gender component, the resulting Akaike score is 2293.66. If one was to now retract age from the model, the score is 2387.426. This implies that age is a better parameter to consider if was to choose between age and gender. This also further explains the figure **??**, where the survival functions for the gender after correction for age are arguably similar. With this in mind and then looking at the figure **??**, the age could be a variable that provides a considerable amount of information for a survival. This seems reasonable due to the fact that the older you are, the more difficult it is to recover and avoid future serious heart attacks.

4.11 Summary

To summarize the above, the optimal model found was with all of the physical traits a subject could have being present in the model. This seems reasonable, due to the correlations provided in the pair plots and due to the outcome studied i.e hear attack.

A suggestion to get the optimal number of parameters is to use LASSO (least absolute shrinkage and selection operator). Another approach could be to use cross validation. This meaning that one tests for different hyper-parameters for the model, one being the predictors.