Hemtentamen i survival analysis

Hampus Beijer

2025-12-05

Question 1 (2p)

Discuss in what sort of problems is survive analysis useful. How can it help decision makers in their work?

In order to discuss when survival analysis is useful, one must first understand what it is. Survival analysis is a collection of methods used to study how long it takes until an event occurs. We study something over time, which can be a person, a machine, and so on. Survival analysis is not restricted to one method, but rather includes multiple methods that use the concept of the survival function. The survival function is the core of survival analysis, which tells us the probability that something is alive (event has not happened yet) at a given time: $S(t) = P(T > t)$. Here T is the random survival time, a random variable, and t is a specific timepoint (Bartoszek, 2025, s.10). When the survival function has been defined, one can estimate the Kaplan-Meier curve, which is an estimated line (or rather step function) of how long the population survival time, which can be expanded to two lines (meaning 2 survival functions). Then one can compute confidence intervals and confidence bands, however, to do so we must use Greenwood's formula (s. 25). This formula makes it possible to estimate the variance of the survival function. Hence, the standard error for the confidence interval/bands can be computed (s.26). However, when doing survival analysis one is not restricted to only looking at Kaplan-Meier curves. Another method for doing survival analysis is Cox regression (s. 42), which allows one to see how multiple covariates affect the hazard (instantaneous risk of the event) over time. But Cox regression has assumptions that needs to be met, and in some cases this can be avoided by using another methods, such as decision trees (s. 67). Decision trees are non-parametric which means that the proportional hazards assumption do not have to be met.

Now when survival analysis has been introduced the discussion can begin on when it is useful. Problems that survival analysis is useful for is when studying how long something lasts until failiure. Example, Samsung announced their new phone, the Samsung Flip. In short, this phone can be open and closed. The manufacture of this phone might be interested in how many times the phone can be open and closed before the folding mechansim fails. This can be studied with survival analysis. So, survival analysis can be useful to get an understand of how long something last until it breaks. In this case, it can help the manufacture to see whether the folding mechansim can last over a long time.

Another important application is in the study of medicine. During the COVID-19 pandemic, pharmaceutical companies released vaccines to protect against the virus. During this period multiple studies (survival analysis) were conducted, overlooking the survival rate of individuals who took the vaccine. Survival analysis was useful in this field because it was possible to study how long a person could "survive" after taking the vaccine. However, surviving in this case must not mean death. So, survival analysis in this case could help decision makers to evalue the vaccines safety and effectiveness.

Question 2 (4p)

The dataset i presented in table 1, where the time variable is the age of each individual at the follow-up, and the status is the status of the individual. In this study the event is outcome, *diagnosis*. When doing the survival analysis without censoring the event is coded as 1. When censoring was taken into account the event is coded as 1, otherwise 0.

Table 1: Dataset of breast cancer in ascending order.

time	event
36	Death from heat attack
37	Death from heat attack
44	Diagnosis
49	Death from heat attack
63	Diagnosis
67	Diagnosis
73	Death in car accident
77	Emigrated
78	Diagnosis
80	Emigrated

Question 2a)

In the file breast cancer.csv (e–mailed to you), in the first column, age, you will find the the time (since birth) in years till diagnosis of breast cancer for 10 women carrying the BRCA1 or 1 BRCA2 (high risk of cancer genes). Question: Calculate a life–table. Present your calculations

```
# Loading required packages
require(kableExtra)
require(dplyr)

# Reading the data
data <-
data <- read.csv2("\\\\filur03.it.liu.se\\students\\hambe399\\Downloads\\breast_cancer.csv")


# Rearranging the data in ascending order
data <- data %>%
  arrange(age) %>%
  rename( "time"=age,  "event"=follow_up)

# Presenting the dataset in a table
kable(data, caption ="Dataset of breast cancer in ascending order.")
```

Since time is a continuous random variable, it is divided into intervals of 20 years. In table 2, the life table is presented. The first variable x is the intervals, *start* is n_x which is the number of individuals who enter the interval x, the variable *Censored* can be ignored for now, *Atrisk* are the number of people who are at risk which equals n_x since censoring is not important, *Diagnosis* are the amount of people who didn't go on to the next interval, $Pr.surv.x$ is the probability of diagnosis within each interval, $Pr.sur.x$is the probability of survivng the given interval, and lastly $Prsurvgtx$ is the probability of surviving after the given interval.

Interpretation of table 2

In table 2 the interval is divided into 20 year gap intervals. At the begining of interval $(0, 20]$, 10 people were at risk, with 0 diagnosis, 0 probability of diagnosis, and 100% probability of surviving this interval and beyond it. This makes sense since there are no individuals dying at this age range. In the second interval $(20, 40]$, 10 people were at risk, 2 people were diagnosed resluting in a 20% probability of diagnosis. Hence, the probability of surviving this interval is 80%. In the third interal $(40, 60]$ there were 8 people at the start/at risk. Two people were diagnosed, the probability of diagnosis was 25%, the probability of surviving this interval was 75%, and the probability of surviving beyond this interval was 60%. In the last interval $(60, 80]$ 6 people were at risk/start and all people were diagnosed, resluting in a diagnosis rate of 100%. The probability of surviving this interval and beyond is 0%.

Table 2: Life table for cancer breast dataset without caring for censoring.

x	Censored	Start	At.risk	Deaths	Pr.death	Pr.surv.x	Pr.surv.gt.x
(0,20]	-	10	10	0	0.00	1.00	1.0
(20,40]	-	10	10	2	0.20	0.80	0.8
(40,60]	-	8	8	2	0.25	0.75	0.6
(60,80]	-	6	6	6	1.00	0.00	0.0

```r
# Calculating the life table without caring for censored data.

life_table <- data.frame(
  # Bcs time is continuous I made it into intervals
  x = c("(0,20]", "(20,40]", "(40,60]", "(60,80]"),

  # Censored i detta fall är de som dött
  Censored = c("-","-","-","-"),

  Start = c(10,
              10, # two people died here
                8, # one person diagnosed one person dead here
                  8-2), # all people died (no censoring included)

  # Individuals at risk
  "At risk" = c(10,10,8,6),

  # Deaths i detta fallet är de som fått en diagnos inte dött
  Diagnosis = c(0,2,2,6),

  # probability of death
  "Pr diagnosis" = c(
    0/10, # no one dead yet
    2/10, # two people had the event, 10 remaining
    2/8, # two people had the event, 8 remaining
    6/6 # six people had the event, 6 left.
  ),

  #probability of surviving
  "Pr surv x" = c(
    1 - 0, # prob of death is 0%
    1 - (2/10), #prob of death is 2/10
```

```
    1 - (2/8), # prob of death is (2/8)
    1- (6/6) #prob of death is 100%
   ),

  "Pr surv gt x" = c(
    1*1, # This is S(20), but is missing event
    1*0.8, # This is S(40),
    0.8*0.75, # This is S(60)
    0.6*0 # This is S(85)
  )
)
```

Question 2b)

It the second column, follow up you will find information that some women were lost for follow up. Question: Recalculate the life–table by taking into account censoring. Present your calculations

In table 3, the life table of cancer diagnosis when taking into account for censoring, is presented. At the beginning of the interval $(0, 20]$ and $(20, 40]$ no events occurs, resulting in a 0% probability of diagnosis, survival rate, and a 100% cumulative survival rate. It should be noted that at the first interval that no individual was censored, however, at the second one person was censored. At the third interval $(40, 60]$, one person was diagnosed with cancer (event occured) resluting in 7.5 (≈ 8) people at risk. There was a 13% chance of being diagnosed, with a survival rate of 87%, the probability of surviving beyond this interval was 87%. During the interval $((40, 60])$ 1 person was censored. During the last interval $(60, 85]$, three people were censored, 6 individuals were at start and 4.5 (≈ 5) at risk. This meant that there was a 67% (Six seeeeven) chance of diagnosis, a 33% chance of not being diagnosed during that interval, and 29% chance of surving beyong the interval.

Table 3: Life table for cancer breast dataset when caring for censoring.

x	Censored	Start	At.risk	Diagnosis	Pr.diagnosis	Pr.surv.x	Pr.surv.gt.x
(0,20]	0	10	10.0	0	0.0000	1.0000	1.0000
(20,40]	2	10	9.0	0	0.0000	1.0000	1.0000
(40,60]	1	8	7.5	1	0.1333	0.8667	0.8667
(60,85]	3	6	4.5	3	0.6667	0.3333	0.2889

```
life_table_cen <- data.frame(
  # Bcs time is continuous I made it into intervals
  x = c("(0,20]", "(20,40]", "(40,60]", "(60,85]"),

  # Censored i detta fall är de som dött
  Censored = c("0","2","1","3"),

  Start = c(10,
            10, # two people died here
            8, # one person diagnosed one person dead here
            8-2), # all people died (no censoring included)

  # Individuals at risk
  "At risk" = c(
    (10-(0/2)), #10 at nx, 0 censored
    (10-(2/2)), #10 at nx, 2 censored
    (8-(1/2)), #8 at nx, 1 censored
    (6-(3/2))  #6 at nx, 3 censored
    ),

  # Deaths i detta fallet är de som fått en diagnos inte dött
  Diagnosis = c(0,0,1,3),

  # probability of diagnosis
  "Pr diagnosis" = c(
    round(0/10, 4), # no one diagnosed yet
    round(0/9,4), # two people had the event, 10 remaining
```

```
    round(1/7.5,4), # two people had the event, 8 remaining
    round(3/4.5,4) # six people had the event, 6 left.
  ),

  #probability of surviving
  "Pr surv x" = c(
    1 - 0, # prob of diagnosis is 0%
    1 - 0, #prob of diagnosis is 2/10
    1 - round(1/7.5,4), # prob of diagnosis is (2/8)
    1- round(3/4.5,4) #prob of diagnosis is 100%
  ),

  "Pr surv gt x" = c(
    round(1*1,4), # This is S(20) prob of surviving beyond this interval is 100%
    round(1*1,4), # This is S(40), prov of surv beyoung is 1*1
    round(1*0.8667,4), # This is S(60), previous was 100% and 87% surviving the intervall
    round(0.8667*0.3333,4) # This is S(80)
  )
)

kable(life_table_cen, caption = "Life table for cancer breast dataset when caring for censoring.")
```

Uppgift 2c)

Calculate the life–tables in a) and b) with the help of computer software

In this section the life tables were calculated automatically using SAS. The dataset was written in code because it was easier writing the status variable. Proc lifetest was used in construction the life tables where intervals specified.

Life table from a)

In figure 1, the life table from SAS is presented. The resluts is the same as in the previous exercise, however, additional outputs were produced such as survival standard error, median residual lifetime, PDF, PDF standard error, hazard, and hazard standard error. However since the resluts are exactly the same as previously, there is no point in re-interpreting them. The SAS code can be seen below.

Life Table Survival Estimates

Interval												Evaluated at the Midpoint of the Interval			
[Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	20	0	0	10.0	0	0	1.0000	0	0	64.1667	6.5881	0	.	0	.
20	40	2	0	10.0	0.2000	0.1265	1.0000	0	0	44.1667	6.5881	0.0100	0.00632	0.011111	0.007808
40	60	2	0	8.0	0.2500	0.1531	0.8000	0.2000	0.1265	28.3333	5.8926	0.0100	0.00632	0.014286	0.009998
60	85	6	0	6.0	1.0000	0	0.6000	0.4000	0.1549	12.5000	5.1031	0.0240	0.00620	0.08	0
85	.	0	0	0.0	0	0	0	1.0000	0	.	.	.	.	.	.

Figure 1: Life table on the breast cancer dataset, showing interval survival where censored observationes are NOT taken into account.

```
* Loading the dataset;
data cancer_nocens;
input time event;
datalines;
73 1
78 1
80 1
77 1
63 1
67 1
36 1
44 1
37 1
49 1
;

*Using proc lifetest to automatically calculate a lifetable;
proc lifetest data=cancer_nocens method=lt intervals=20 40 60 85;
   time time*event(0);
run;
```

Life table from b)

In figure 2, the SAS output for the life table when taking censoring into account is shown. It can once again be shown that the outputs in this life table is exactly the same as the previous. Hence, there is no need in interpreting the outputs.

Life Table Survival Estimates															
Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	Evaluated at the Midpoint of the Interval			
[Lower,	Upper)											PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	20	0	0	10.0	0	0	1.0000	0	0	75.8654	6.8415	0	.	0	.
20	40	0	2	9.0	0	0	1.0000	0	0	55.8654	7.2115	0	.	0	.
40	60	1	1	7.5	0.1333	0.1241	1.0000	0	0	35.8654	7.8998	0.00667	0.00621	0.007143	0.007125
60	85	3	3	4.5	0.6667	0.2222	0.8667	0.1333	0.1241	18.7500	8.8388	0.0231	0.00838	0.04	0.02
85	.	0	0	0.0	0	0	0.2889	0.7111	0.1970	.	.	.	.	.	.

Figure 2: Life table on the breast cancer dataset, showing interval survival where censored observationes are taken into account.

```
* Loading the dataset;
data cancer_nocens;
input time event;
datalines;
73 0
78 1
80 0
77 0
63 1
67 1
36 0
44 1
37 0
49 0
;

*Using proc lifetest to automatically calculate a lifetable;
proc lifetest data=cancer_nocens method=lt intervals=20 40 60 85;
   time time*event(0);
run;
```

Question 3 (7p)

A petrochemical company recorded survival times beyond age of 65 for its employees, recording whether they retired at age 65 or 55 (data in this exercise is made–up for simplicity, but such a study was conducted on Shell employees retiring between 1973 and 2003, see Tsai et. al. (2005) "Age at retirement and long term survival of an industrial population: prospective cohort study", BMJ, 331(7523): 995. if you are interested). The survival times beyond age 65 in years is for the two groups can be found in the file pertrochemical survival times.csv (e–mailed to you). The first column is the id of the employee, the second is the retirement age, the third survival time, the fourth status (2: uncensored, 1: censored).

For this exercise, the dataset is a set of a petrochemical company's survival times over the age of 65, comparing employees who retired at the age 55 or 65. The dataset is presented in table 4, where *invidID* is the employee's ID, age is the retirement age (65 or 55), time is the employee's survival time (1 or 2), and status is whether the event occured (1 = event occured, 0 = censored). It should be noted that the data was altered, where status was re-coded to 0 or 1.

Table 4: Survival dataset of petrochemical company

indivID	age	time	status
1	55	2	0
2	55	1	0
3	55	5	0
4	55	2	0
5	55	2	0
6	55	3	0
7	55	1	1
8	55	3	0
9	55	1	1
10	55	2	0
11	55	1	0
12	55	1	1
13	55	1	1
14	55	4	0
15	55	1	0

```
# Reading the dataset
data <- read.csv2("\\\\filur03.it.liu.se\\students\\hambe399\\Downloads\\pertrochemical_survival_times.
# Coding of the data for the status variable
# is causing some problems so I re-code
# the variable here with ifelse
data$status <- ifelse(data$status==2,1,0)

# Presenting the dataset
kable(head(data,10), caption ="Survival dataset of petrochemical company ")
```

Theory before answering 3a)

The Kaplan Meier curve is a non-parametric estimation of a survival function $S(t) = P(T > t)$ which gives a curve, or rather a step function showing survival over time for a random variable. Better explained, the survival function makes it possible to calculating the probability of surviving a given interval. (Rich et al., 2010). The formula to estimating the curve is as follows:

$$\hat{S}(t) = \prod_{x:x<t} (1 - q_x) = \prod_{x:x<t} \left(1 - \frac{d_x}{r_x}\right)$$

However, in order to calculate confidence intervals and confidence bands, the variance must be estimated. This can be done with Greenwood's formula (Bartoszek, 2025, p.25):

$$v^2(t) := \text{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{x:x<t} \frac{d_x}{r_x(r_x - d_x)}$$

where:

- t is a timepoint.
- x is the observed event.
- d_x is the number of events.
- r_x are the number of individuals at risk, right before the time x.
- $1 - d_x/r_x$ is the probablity of surviving the interval.

Now it is possible to estimate a confidence intervals and confidence bands as follows (p. 25-26):

$$\text{CI: } \left[\hat{S}(t) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) v(t),\ \hat{S}(t) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) v(t)\right]$$

$$\text{Confidence bands: } \sqrt{n}\left(\hat{S}(t) - S(t)\right) \approx S(t)\, W(\sigma(t))$$

When analysing the confidence intervals, one must not make a statement about the entire survival curve. This is because the confidence intervalsa single time point estimation, meaning when analysing two time points the confidence is less than 95% (problem with multiple testing). However, the confidence bands are simultaneous. This means that when analysing the confidence bands, one can make a decision about the entire curev and not just a specific time point (Sachs et al., 2022).

To statistically determine whether two Kaplan-Meier curves actually differs, a log-rank test can be performed. This a χ^2-test (Rich et al., 2010). This is a hypothesis test where statistic is as follows:

$$Z = \sum_x \left(d_x^{(1)} - \frac{d_x\, r_x^{(1)}}{r_x}\right)$$

A p-value for the statistic can be caluclated to determine whether the nullhypothesis (that the groups do not differ) can be rejected. If the null hypothesis is rejected, then it can be stated that the survival distribution differ (Bartoszek, 2025, p.39).

Question 3a)

Plot the survival function for the groups. Include confidence intervals and confidence bands. Which are wider and more appropriate, the confidence intervals or confidence bands? Find the median and mean survival times (with confidence intervals) for the groups. How would you interpret the results? Can you provide some explanation for them?

In this survival analysis, ggplot and ggplotsurv was used instead of base R's plot function because it gives better graphs with confidence intervals and bands. In figure 3, the Kaplan Meier curves for the comparisons of retirement age are shown. In the figure, time is presented on the x-axis and survival probability after 65 years on the y-axis. The red graph curve is for the retirement age of 55 and blue for 65. In the graphs plus signs can be shown which indicate censored data. When analysing the curves, it can be noted that the for individuals retiring at 65 years over time are higher than for 55 years. This difference is not huge, however, note that the survival probability for 65 goes to zero at time $t = 8$ while 55 years are slightly over 0%. With that being said, at $t = 8$, we are 95% sure that the true survival probability is 0 for retirement age 65, and 95% sure that the true survival probability for retirement age 55 is more than 0%. It does not make much sense that people who retire early have a higher survival rate than those who retire later, because the once retiring early are probably doing so for a reason (i.e health problems). One may question the accuracy of the results.

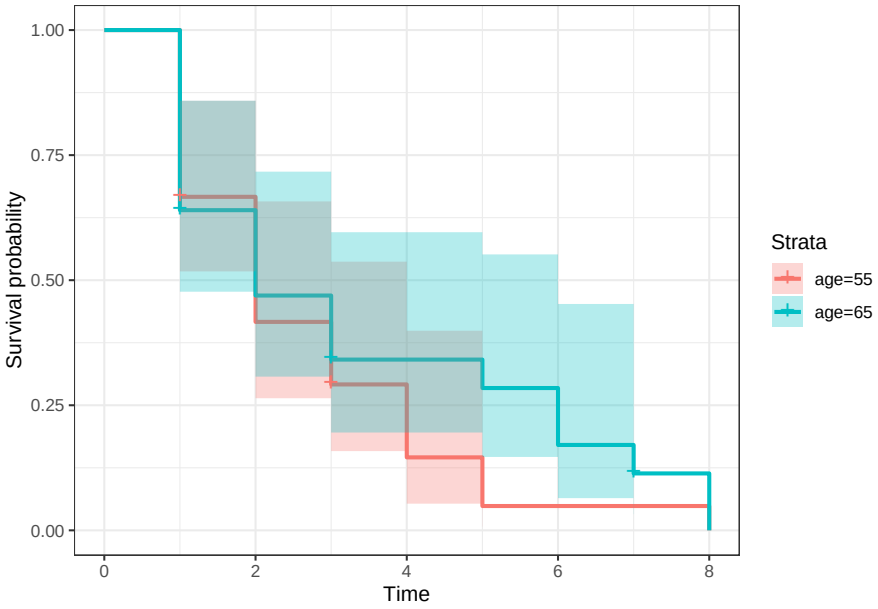


Figure 3: Kaplan-Meier curve with confidence intervals.

In figure 4, the Kaplan Meier curves are presented once again but now with confidence bands. Note that for the confidence intervals (in figure 3), the intervals are point by point estimates, which means that it is only possible to look at one timepoint at a time. If one were to check more than one timepoint the confidence would not equal to 95, because of it being multiple testing. This is not the case when doing confidence bands, since these are simultaneous. The confidence bands (figure 3) for the Kaplan Meier curves are wide and overalp each other. Thus indicating that there are a big uncertainty the estimates. Since there are a big chance of drawing the wrong conclusions by just looking at the confidence bands, one may want to draw a real conclusion based on a statistical test of whether the curves do differ. For this a log-rank test will be used.

Performing a Log-rank test

A log ranked test was performed at 0.05 significance level, with following hypothesis:

$$H_0 : S_{55}(t) = S_{65}(t) \text{ for all } t$$

$$H_a : S_{55}(t) \neq S_{65}(t) \text{ for at least one } t$$

The hypotheses state that if there is no difference between individuals who retired at 55 vs 65, the null hypothesis is rejected. Thus, indicating that at a 5% significance level there is a significant difference in survival.

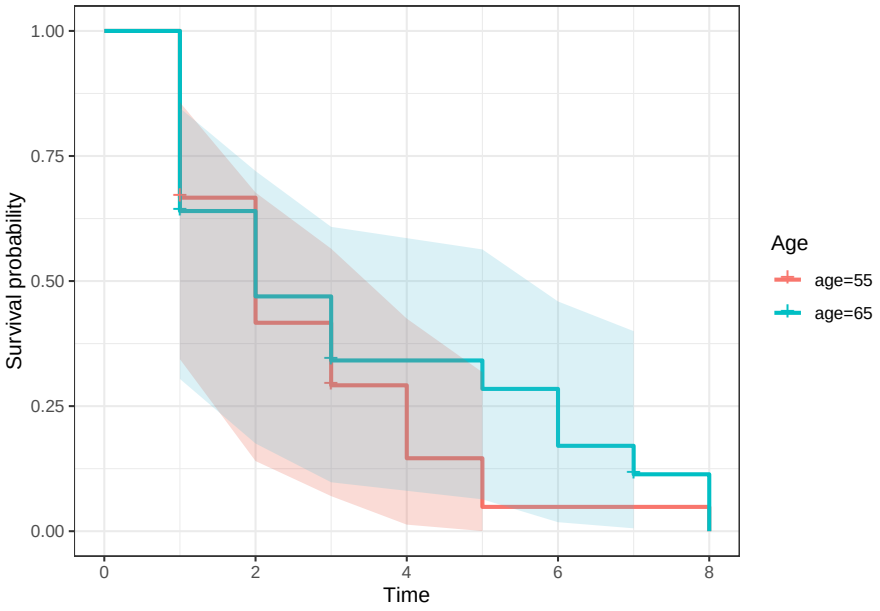


Figure 4: Kaplan-Meier curve with confidence bands.

In table 5 and 6, the calculations and results of the log-rank test are shown. Specifically in table 6, the results show a χ^2 statistic of 0.9 and p-value of 0.4. This means that the null hypothesis is not rejected, thus, there is no difference in survival when comparing individuals retiring at 55 vs 65. Fun fact, in Greek, χ is pronounced the same way as *hi* in *history*. This is why some greeks roll their eyes when they hear someone pronounce it in a "wrong" way.

Table 5: Calculations of the log-rank test

age	N	Observed	(O-E)ˆ2/E	(O-E)ˆ2/V
55	30	22.5	0.286	0.853
65	25	23.5	0.273	0.853

Table 6: Resluts of the statistic and p-value of the log-rank test.

chisq	df	p
0.9	1	0.4

In table 7, the survival models outputs are presented. The first column shows age of retiring (55 vs 65). The second column shows that in 30 individuals retired at 55, while 25 individuals retired at 65. The third column shows that from the 30 individuals retired at 55, 25 had the event. 21 out of the 25 who retired at 65 had the event. In the third column rmeans (restricted mean survival time) is shown, which is the mean survival time estimated up to the last timepoint (8). For group 5 the rmean is 2.67 years and 3.36 years for the 65 group. Hence, group 65 lived approximately 0.7 years more than group 55. The next column presents the standard error for the restricted mean which are 0.356 for group 55 and 0.542 for group 2. The standard error mesaures the precision of the mean estimate is. The SE estimates are relatively large compared to the mean estimates, indicating that the mean estimates are somewhat imprecise. This is also reflected in the 95% confidence intervals for the median survival, which are quite wide, especially for group 65. Hence, there is a big uncertainty around the estimates. The median for both groups are 2. The interval for group 55 is

[2,4], and [1,6] for group 65. The time interval is from 0-8, which means that group 65 covers a large part of the studied time interval.

Table 7: Coefficients from the model including: age, n, rmean, SE(rmean), median, and 95% confidence interval.

age	n	events	rmean	se_rmean	median	LCL	UCL
55	30	25	2.67	0.356	2	2	4
65	25	21	3.36	0.524	2	1	6

```
# Loading required packages
require(survival)
require(survminer)

# Basically what width does here is that it
# marks whether an observation is censored
# or if the event happened
# if + then the observation is censored,
# otherwise the event occured!
time_with <- with(data, Surv(time,status))

# Now we fit the model: the Kaplan Meier estimation
model.fit <- survfit(time_with~age,
                     data = data
)

#  Base R plot function is not very good visualised so
# I prefer using ggsurvplot

# obs this is only for CI for the curves
plot1 <-ggsurvplot(
  model.fit, # The model
  data= data, # the data
  conf.int = T # the CI
)

# just to add theme to the plot
plot1 <- plot1$plot + theme_bw()

# need this packages for the bands
require(km.ci)

# Using km.ci to get the individually
band.55 <- km.ci(survfit(Surv(time,status) ~ 1,
                 data = subset(data, age == 55)),
                 conf.level = 0.95,
                 method = "logep"
                 )

band.65 <- km.ci(survfit(Surv(time,status) ~ 1,
                 data = subset(data, age == 65)),
                 conf.level = 0.95,
                 method = "logep"
```

```
                )

# Now I need to manually construct the data frames
# because ggplot requires DF to work (buh huu)
df55 <- data.frame(
  time = band.55$time,
  lower = band.55$lower,
  upper = band.55$upper
)

df65 <- data.frame(
  time = band.65$time,
  lower = band.65$lower,
  upper = band.65$upper
)

# Now I redo the plot without confidence intervals
# I save it and I need to use that to add the confidence
# bands.

plot2 <- ggsurvplot(
  model.fit,
  data = data,
  conf.int = FALSE, # removing CI to show the bands
  legend.title = "Age"
)$plot

plot2_2 <- plot2 +
  geom_ribbon(
    data = df55,
    inherit.aes = FALSE,
    aes(x = time, ymin = lower, ymax = upper),
    fill = "#E64B35", alpha = 0.20
  ) +
  geom_ribbon(
    data = df65,
    inherit.aes = FALSE,
    aes(x = time, ymin = lower, ymax = upper),
    fill = "#4DBBD5", alpha = 0.20
  ) +
  theme_bw()

# Doing a log rank test of difference between the Kaplan-Meier curves
  # This test is a log-rank test!
      #survdiff(time_with ~age,data=data)

# Saving resluts in DF
df_diff <- data.frame(
  age = c(55, 65),
  N = c(30, 25),
  Observed = c(22.5, 23.5),
  "(O-E)^2/E" = c(0.286, 0.273),
  "(O-E)^2/V" = c(0.853, 0.853),
```

```
  check.names = FALSE # IMPORTANT otherwise the formula wont show
 )

# Saving resluts in DF
test <- data.frame(
  chisq = 0.9,
  df = 1,
  p = 0.4
)

kable(df_diff, caption ="Calculations of the log-rank test")
kable(test, caption ="Resluts of the statistic and p-value of the log-rank test.")
# Taking the coefs and making a DF to present in a table
  #table_coefs <- print(model.fit,  print.rmean = TRUE)
table_coefs <- data.frame(
  age     = c(55, 65), #taking information from print above
  n       = c(30, 25),
  events  = c(25, 21),
  rmean   = c(2.67, 3.36),
  se_rmean = c(0.356, 0.524),
  median  = c(2, 2),
  LCL     = c(2, 1),
  UCL     = c(4, 6)
)

# Presenting the resluts
kable(table_coefs, caption ="Coefficients from the model including: age, n, rmean, SE(rmean), median, a
```

Theory before answering 3b and 3c)

Cox regression is a method to perform regression on survival data (Walters, 2009, s. 1). The model is formulated as follows (Duerden, 2009, p. 1):

$$h(t) = h_0(t) \exp(\beta_1 x_1 + ... + \beta_p x_p)$$

where :

- $h_0(t)$ is the baseline hazard.
- x_p are the covariates.
- β_p are the cofficients which describe how the covariates proportionally decreases/increases.

The coefficients in the Cox model can be interpreted as:

- Negative coefficient: a lower hazard, meaning a better survival.
- Positive coefficient: a higher hazard, meaning a worse survival.

The Cox model has a assumption which needs to be met, called the proprtional hazards assumption. This assumption says that effect over time needs to be constant (Walters, 2009, p. 4-5). To see if this assumption is met for each covariate the Schoenfeld residuals of the model can be calculated and visualised. The formula for the Schoenfeld residuals are (Bartoszek, 2025, p.56):

$$r_{ci} = x_{ci} - \sum_{k \in R_i} x_{ck}\, p_{ck}$$

$$p_{ck} = \frac{\exp(\beta^T X_k)}{\sum_{r \in R_i} \exp(\beta^T X_r)}$$

where:

- r_{ci} is the Schoenfeld residual for covariate c for an individual i.
- x_{ci} is the observed value of the covariate c for an individual i.
- p_{ck} is the probablity that an individual k would fail at the given time.

When looking at the plot the residuals over time should be constant (independent of time) (Bartoszek, 2025, p.58).

When comparing multiple Cox models, ΔAIC can be used. The formula is as follows:

$$\Delta\text{AIC} = AICci - AIC_{cmin}$$

where the AIC is calculated with the following formula:

$$AIC = -2L + 2K$$

and

$$AIC_c = AIC + \frac{2K(K+1)}{n-k-1}$$

Where L is the likliehood, n are the amount of observations, and K are free parameters. To note, one should not look at the AIC alone. Rather the difference in AIC should be checked (ΔAIC). When comparing the ΔAIC, if the difference in models are very low (i.e ≤ 4) one should not choose a mdoel over the other. However, if the difference is bigger (i.e > 14) then the other model is a better fit (Bartoszek, 2025, p.53-54).

Uppgift 3b)

Perform a Cox regression analysis to formally test if there is a significant difference in survival times between the two groups. How do you interpret the output of a Cox regression analysis?

In this task a cox regression model is made to see whether there is a significant difference in survival times between the two groups studied. The model, with one covariate can be defined as follow:

$$h(t) = h_0 \exp(\beta_1 \cdot x_1)$$

Where,

- $h(t)$ = the hazard rate for an individual time t-
- $h_0(t)$ is the baseline hazard.
- x_1 is the covariate, tetirement age.
- β_1 is the coefficient which modifies the baseline hazard.

The model is estimated, and it's parametercoefficients are presented in table 8. The coefficient, $\beta_1 = -0.02618$, which is the log hazard ratio. The exp() of the log hazard ratio, is the hazard ratio ($\exp(\beta_1) = 0.9741$), with a standard error of 0.030. The z-statistic is -865 with a p-value < 0.05. This can be interpreted as, the hazard ratio of 97.4% means that the individuals who retired at 65 have a 2.6% lower hazard of dying after the age of 65 in comparison to those who retire at 55. However, the p-value is insignificant, which means that no conclusion can be drawn whether this is true.

Table 8: The summary of the Cox model with age 65 vs 55

coef	exp(coef)	SE(coef)	P(>|z|)	p-value
-0.02618	0.97416	0.03026	-865	0.3855

When estimating the Cox model, a LRT test was provided. A LRT test is a Likelihood Ratio Test which compares a the model with one covariate to a model with no covariates. In table 9, the LRT of 0.75, df, p-value, n, and number of events are provided. The p-value is nonsignificant, which means that the nullhypotehesis that retirement age has no effect on the hazard cannot be rejected.

Table 9: Likelihood Ratio Test for the Cox model.

LRT	df	p-value	n	n_events
0.75	1	0.3855	55	46

In table 10, the models concordance is presented. The concordance can be thought of as the R^2 in linear regression. The concordance measures how well the model can rank the individuals by risk. The concordance in this case is 0.52, indicating that the model is no better than random guessing whether a person in group 55 will die before a person in group 65.

Table 10: The Cox models concordance value.

concordance
0.5159

```r
# To perform the cox regression I reuse the
# already made time_with

# Fitting the cox model
cox.fit <- coxph(time_with ~ age,
                 data = data
                 )

# Making a DF with the cox modell
df_cox <- data.frame(
  coef = c(-0.02618),
  "exp(coef)" = c(0.97416),
  "SE(coef)"  = c(0.03026),
  "P(>|z|)"         = c(-0865),
  "p-value"   = c(0.3855),
  check.names = FALSE
)

df_test <- data.frame(
  LRT = c(0.75),
  df = c(1),
  "p-value" = 0.3855,
  n = 55,
  n_events = 46,
  check.names = FALSE
)

# Checking for Proportional Hazards assumption
ph_check <- cox.zph(cox.fit)

df_ph <- data.frame(
  "-" = c("age", "GLOBAL"),
  "chisq" = c(0.81,0.81),
  df = c(1,1),
  "p-value" = c(0.37, 0.37),
  check.names = FALSE
)

# Calculating the concordance

  # Need the package survcomp for this
  require(survcomp)

#concordance(cox.fit)
df_concordance <- data.frame(
  concordance = c(0.5159)
)
```

Uppgift 3c)

How would you validate the model? Discuss and perform model validation.

When validating the Cox model, one must check whether the model assumptions are met. In Cox regression, the hazard ratio summarises the effect (the relative risk between groups) in a single number. The p-values and confidence intervals assume that this effect is constant over time as well, meaning that if time is not constant then one cannot rely on what the p-values and CI are saying. This assumption is called the proportional hazards assumption. To check whether this assumption is met, one can study the Choenfeld residuals over time. In figure 5, the Schoenfeld residual plot is presented where the dots are the Schoenfeld residuals , the solid line can be interpreted as checking the effect change over time. In short, the line should be as flat as possible, indicating that the effect does not change over time. The dotted lines are 95% confidence bands. However, the plot can be hard to interpret, therefore, a test can be performed with a χ^2 statistic. The resluts are shown in table 11. For both the covariate age and GLOBAL the χ^2 value is 0.81 with a p-value of 0.37. This is a hypothesis test where the nullhypothesis is that the covariate is constant over time. Since the p-value are non-significant, the nullhypothesis is not rejected. Consequently, effect over time is constant. The GLOBAL variable is the overall test, but since there is only one covariate global = age.

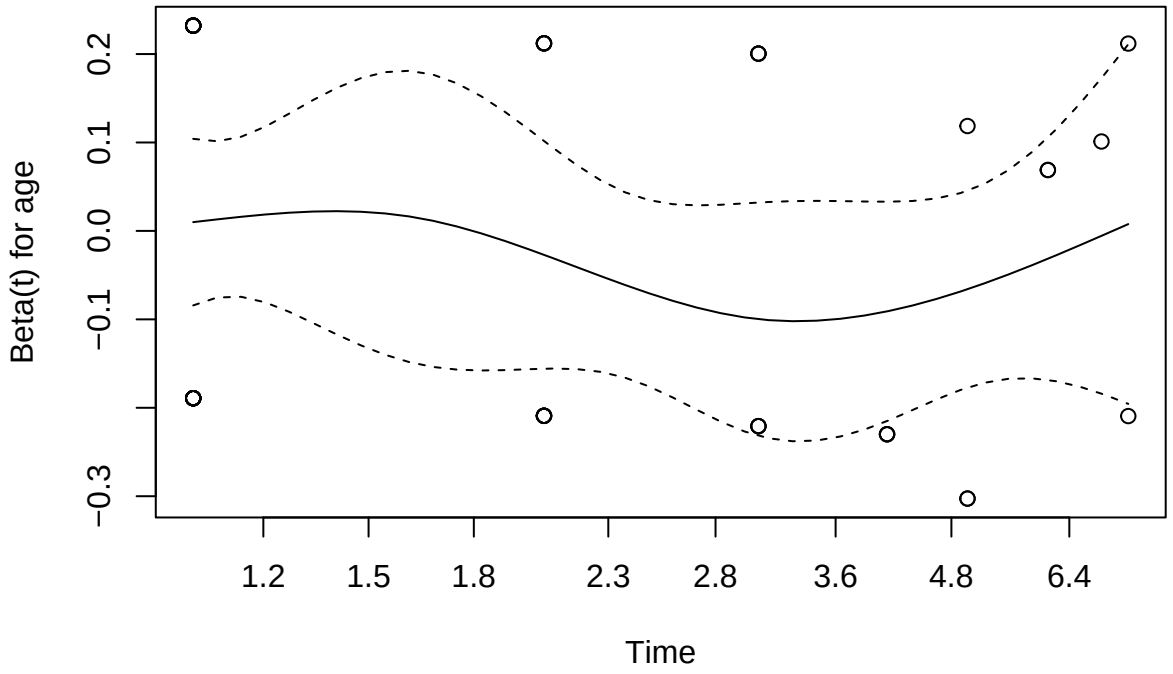


Figure 5: Model validation: Schoenfeld residuals over time for the covariate: age

Table 11: Checking of the poportional hazards assumption of the Cox model.

X.	chisq	pvalue
age	0.81	0.37
GLOBAL	0.81	0.37

```
# Plotting the Schoenfeld residuals for model evaluation
plot(ph_check, var=1)
```

```r
# Checking for Proportional Hazards assumption
ph_check <- cox.zph(cox.fit)
df_assum <- data.frame(
  "-" = c("age","GLOBAL"),
  chisq = c(0.81, 0.81),
  pvalue = c(0.37,0.37)
)
```

Question 4

It is of interest to study which variables best explain patient's survival after a heart attack. It is suggested to start with the variables age, gender, bmi and cvd. Study what effects do the variables have on survival time. Using model selection techniques see if you can significantly improve the model fit by adding variables. Evaluate the fit for the chosen model and interpret the results with words. Do not forget to include plots. If you plot the survival curves include confidence intervals and confidence bands. Which are wider and more appropriate, the confidence intervals or confidence bands

In question 4 I begin by making the suggested base model with age, gender, bmi, and cvd as covariates. This resluts in a model looking something like this:

$$h(t) = h_0(t) \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \beta x_3 + beta_4 x_4)$$

The following cox model is estimated:

```
## Call:
## coxph(formula = data_heart_with ~ age + gender + bmi + cvd, data = data_heart)
##
##   n= 500, number of events= 215
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## age     0.060374  1.062234  0.006571  9.187  < 2e-16 ***
## gender -0.094966  0.909404  0.141049 -0.673  0.50076
## bmi    -0.043409  0.957520  0.015598 -2.783  0.00539 **
## cvd     0.098355  1.103354  0.170390  0.577  0.56378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       1.0622     0.9414    1.0486    1.0760
## gender    0.9094     1.0996    0.6898    1.1990
## bmi       0.9575     1.0444    0.9287    0.9872
## cvd       1.1034     0.9063    0.7901    1.5408
##
## Concordance= 0.74  (se = 0.017 )
## Likelihood ratio test= 150.4  on 4 df,   p=<2e-16
## Wald test            = 128.3  on 4 df,   p=<2e-16
## Score (logrank) test = 135.4  on 4 df,   p=<2e-16

##        chisq df    p
## age    0.146  1 0.70
## gender 0.514  1 0.47
## bmi    1.057  1 0.30
## cvd    0.122  1 0.73
## GLOBAL 2.267  4 0.69
```

In the print of the Cox regression model, it is shown that the hazard ratios are $age = 1.06$, $gender = 0.91$, $bmi = 0.96$, $cvd = 1.10$. Age and BMI are significant, however, gender and cvd are non-significant. It should be noted that BMI is very close to being non-significant, but it's ok for now. Since gender is non-significant, it cannot be said whether men and women differ in survival time. The same goes for cvd. The hazard ratio for BMI is 0.96, whic means that in every one unit increase the hazard of death decreases by 4%, given the

other variables are constant. The hazard ratio for age is 1.06, so in every one unit increase the hazard of death increases by 6%. In conclusion, the model indicates that higher age decreases survival time and higher BMI increases survival time. In the second print, the χ^2 test for the proportional hazards assumption are shown. All of the p-values are non-significant meaning that the nullhypothesis cannot be rejected. Thus, the proportional hazards assumption is met.

To see if the model can be improved by adding additional covariates, the functioin step() is used. This function is an algorithm which calculates ΔAIC to find which covariates are worth adding. Following reslut is obtained:

In table 12, each variable is shown with either a plus or a minus sign. A plus sign before the variable name indicates that the variable was added and improved the AIC. A variable with a minus before the name indicates that the variable was added, but removed because the AIC did not improve. AIC is shown for each model when a variable is added, and lastly ΔAIC was calculated (manually in R because it wasn't possible to extract it) and if the model then based on a decision if the AIC improved significantly over a previous model, the model with a better AIC was chosen. The first delta AIC is not possible to be calculated because there is no model to compare it with.

Table 12: The AIC selection where variable, AIC, and delta AIC is presented. If there is a plus before the variable it means that the algorithm added the variable and AIC was improved. If there is a minus the variable was removed and the AIC improved

Variable	AIC	DeltaAIC
	2312.280	NA
+ chf	2276.050	-36.230367
+ sho	2267.641	-8.408812
+ hr	2258.459	-9.182331
+ diasbp	2248.751	-9.707779
+ year	2244.407	-4.344088
- cvd	2242.512	-1.895522

It was possible to improve the model further by adding covariates. The best model is:

```
## coxph(formula = data_heart_with ~ age + gender + bmi + chf +
##     sho + hr + diasbp + year, data = data_heart)
```

In the print below the it can be noted that all variables are significant. Even gender which was non-signficiant in the previous model became significant. The following hazard ratios for each variable is:

- $age = 1.05$
- $gender = 0.73$
- $bmi = 0.95$
- $chf = 2.04$
- $sho = 3.20$
- $hr = 1.01$
- $diasbp = 0.99$
- $year = 1.28$

This can be interpreted as in every increase in unit for age the hazard of death increases by 5% given that all variables are held constant, for the gender=1 there is a 27% decrease in hazard of death, for every increase in unit for BMI the death of hazard decreases by 5% given that all variables are held constant. chf = 1 has worser survival by double the hazard, cho = 1 gives a increase of hazard of death by 3.2 times. For

every incrase in hr the hazard of death increase by 1%, given that all variables are held constant. For every increase by one unit in diasbp the hazard of death decreases by 1%. Lastly for every one unit increase in year the hazard of death increases by 28%. With that being said the highest risk factor is sho=1 with a hazard ratio of 3.20.

```
## Call:
## coxph(formula = data_heart_with ~ age + gender + bmi + chf +
##     sho + hr + diasbp + year, data = data_heart)
##
##             coef exp(coef)  se(coef)      z        p
## age      0.048777  1.049987  0.006590  7.402 1.34e-13
## gender -0.310662  0.732962  0.144468 -2.150 0.031525
## bmi    -0.051328  0.949967  0.016580 -3.096 0.001963
## chf      0.711200  2.036433  0.148769  4.781 1.75e-06
## sho      1.162817  3.198933  0.265866  4.374 1.22e-05
## hr       0.011718  1.011787  0.002910  4.027 5.64e-05
## diasbp -0.011918  0.988153  0.003471 -3.433 0.000596
## year     0.247974  1.281426  0.099540  2.491 0.012731
##
## Likelihood ratio test=228.1  on 8 df, p=< 2.2e-16
## n= 500, number of events= 215
```

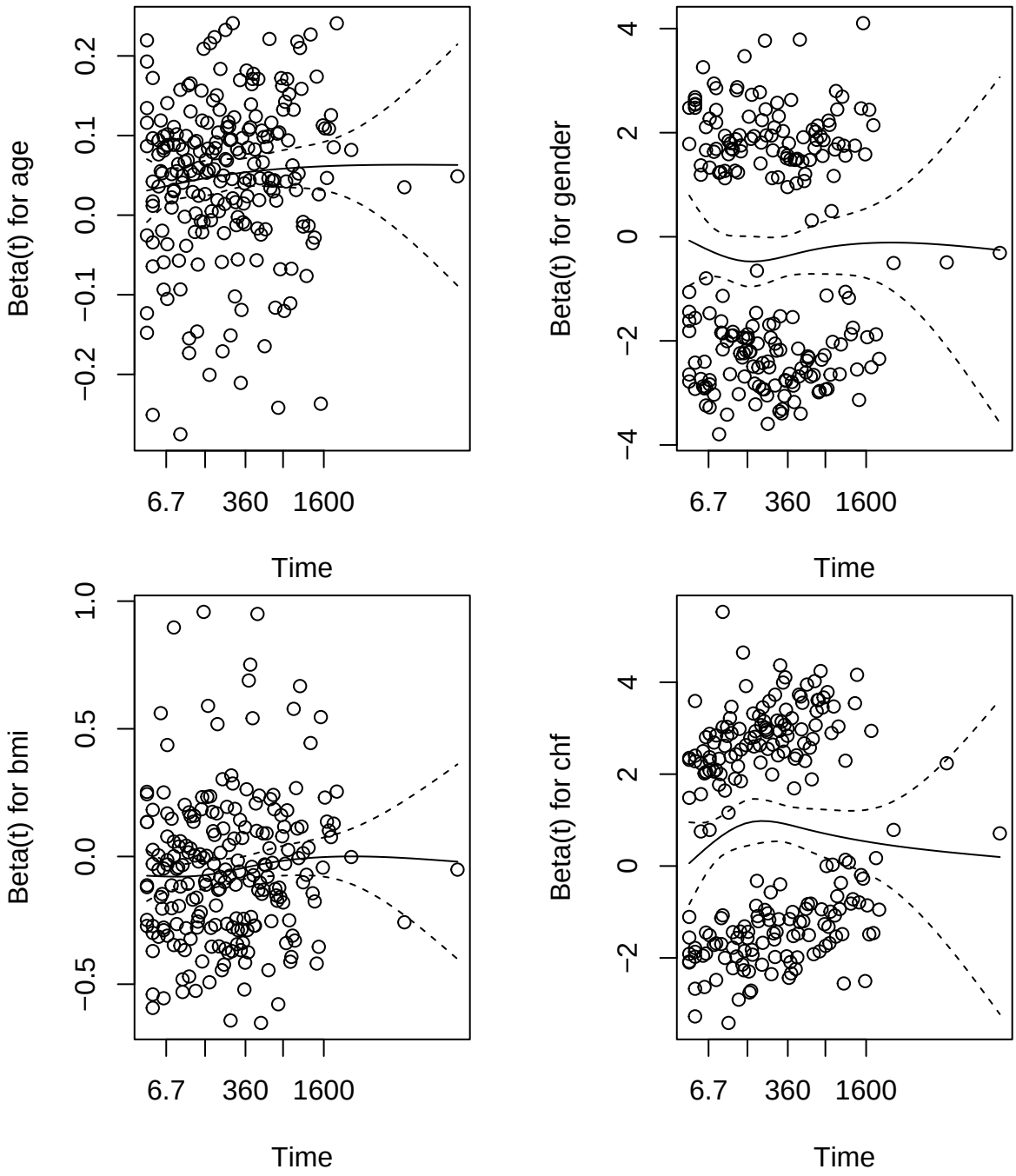
In the print out below the estimated concordance, the test for the proportional hazards assumption, and Schoenfeld residual plots can be seen. The concordande is 0.7872 which is very good. This means that the model is not random guessing. The table for the PH assumption shows that all p-values are non-significant, indicating that the nullhypothesis that the effect over time is constant. This can also be seen below in the eight Schoenfeld residual-plots. To conclude, the model is appropriate to draw conclusion with.

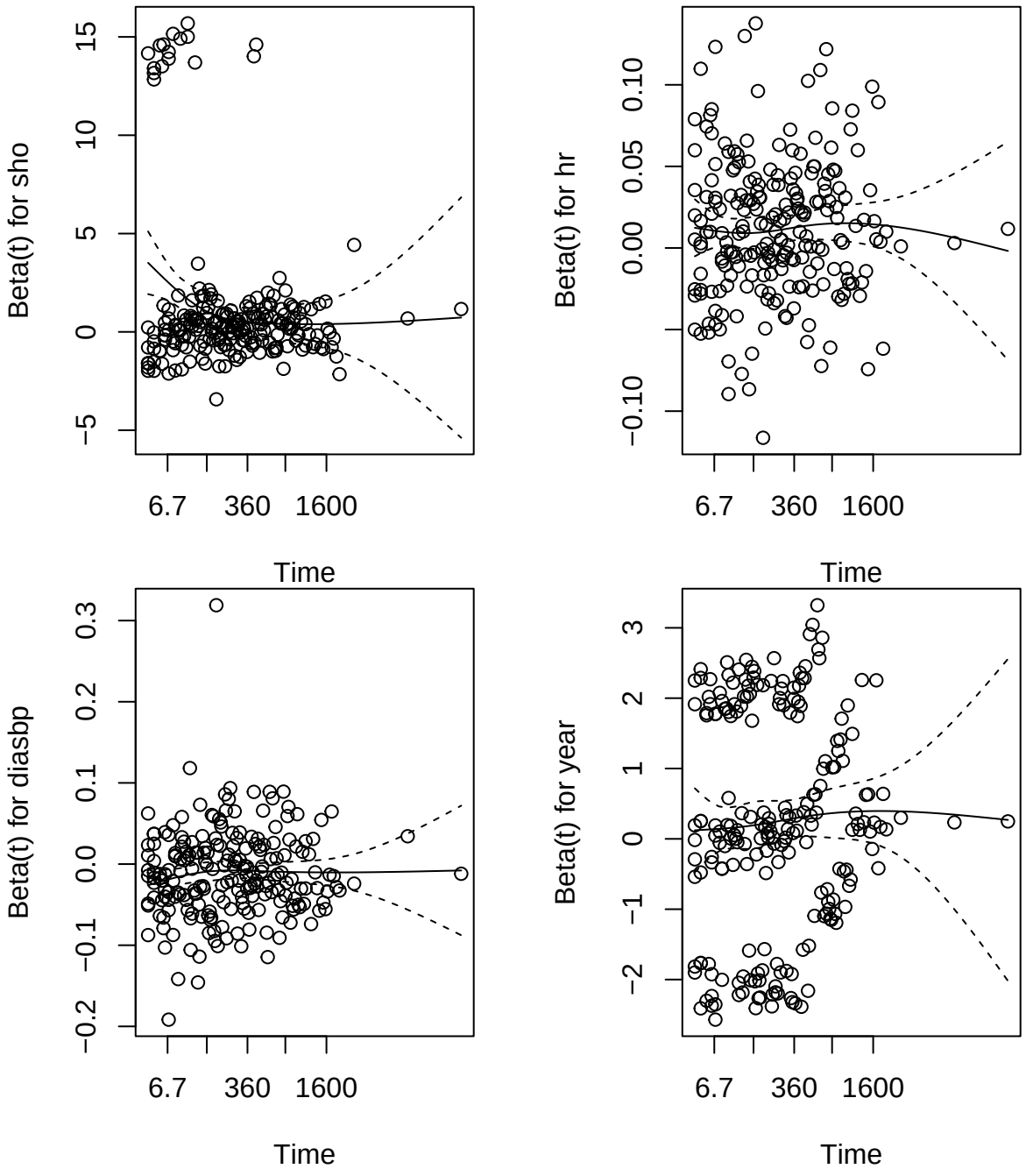
```
# Printing the concordance
round(model_selection$concordance,4)
```

```
##  concordant  discordant      tied.x      tied.y     tied.xy concordance
##  59156.0000  15993.0000      0.0000    119.0000      0.0000      0.7872
##         std
##      0.0153
```

```
cox.zph(model_selection)
```

```
##         chisq df      p
## age     1.284  1 0.2571
## gender  0.124  1 0.7246
## bmi     1.133  1 0.2872
## chf     0.348  1 0.5555
## sho     7.084  1 0.0078
## hr      1.980  1 0.1594
## diasbp  0.544  1 0.4609
## year    3.023  1 0.0821
## GLOBAL 15.196  8 0.0554
```


Beta(t) for age
0.2
0.1
0.0
−0.1
−0.2
6.7
360
1600
Time
Beta(t) for gender
4
2
0
−2
−4
6.7
360
1600
Time
Beta(t) for bmi
1.0
0.5
0.0
−0.5
6.7
360
1600
Time
Beta(t) for chf
4
2
0
−2
6.7
360
1600
Time




```
# Reading the excel file
require(readxl)
data_heart <- read_excel("\\\\filur03.it.liu.se\\students\\hambe399\\Downloads\\heart.xls")

# I begin by making a survival object with with
data_heart_with <- with(data_heart, Surv(time = lenfol, event = fstat))

# I begin with making the first model
# which is recommended in the
# question with following variables:
```

```r
cox.fit.1 <- coxph(data_heart_with ~ age + gender+bmi+cvd,
                   data= data_heart)
summary(cox.fit.1, caption ="Summary of the first model (base).")
#Checking the assumption
cox.zph(cox.fit.1)

# The function step() does automatic AIC selection from the
# base model defined above cox.fit.1.
# step() comes from the base package stats

model_selection <- step(cox.fit.1, scope = ~ age + gender + bmi + cvd +
                     hr + sysbp + diasbp + afb +
                     sho + chf + av3 + miord +
                     mitype + year)

# To explain what just happened was that I defined a base model
# consist of age, gender, bmi, and cvd. I then added remaning
# variables. The algortihm in the step() function basically
# calculated delta AIC. It then added every variable that
# improved the model, consquently, resluting in a lower
# AIC. If the AIC did not improve the variable was
# excluded from the mdoel, and next variable was tested

# I now extract the models that were tried
models_tried <- model_selection$anova

df_selection <- data.frame(
  # I show which variables were added/removed
  Variable = models_tried$Step,
  # I show the AIC
  AIC = models_tried$AIC,
  # I calculate the Delta AIC
  DeltaAIC = c(NA, diff(models_tried$AIC))
)

model_selection$call

# Now I print the best model according to delta AIC!
print(model_selection)

# Printing the concordance
round(model_selection$concordance,4)
cox.zph(model_selection)

# Now I can do residual analysis on the Schoenfeld residuals of the models
prophazard <- cox.zph(model_selection)
par(mfrow=c(1,2))
plot(prophazard, var ="age")
plot(prophazard, var = "gender")
```

```r
plot(prophazard, var = "bmi")
plot(prophazard, var = "chf")
```

```r
plot(prophazard, var = "sho")
plot(prophazard, var = "hr")
```

```r
plot(prophazard, var = "diasbp")
plot(prophazard, var = "year")
```

References

Bartoszek, K. (2025). Survival Analysis: Lecture slides 1–2 [Lecture slides]. Department of Computer and Information Science, Linköping University.

Duerden, M. (2009). What are hazard ratios? Hayward Medical Communications.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan–Meier curves.

Sachs, M. C., Brand, A., & Gabriel, E. E. (2022). Confidence bands in survival analysis. British Journal of Cancer, 127, 1636–1641.

Walters, S. J. (2009). What is a Cox model? Hayward Medical Communications.