UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Data Science for Biomedicine

Lab 3

Tim Boleslawsky

Preliminaries

First we want to look at the histogram and the density plot of the survival times in the data. For this, I read the dataset and filtered based on the condition that FSTAT equals 1. This histogram and density plot of this data can be seen in figure 1.

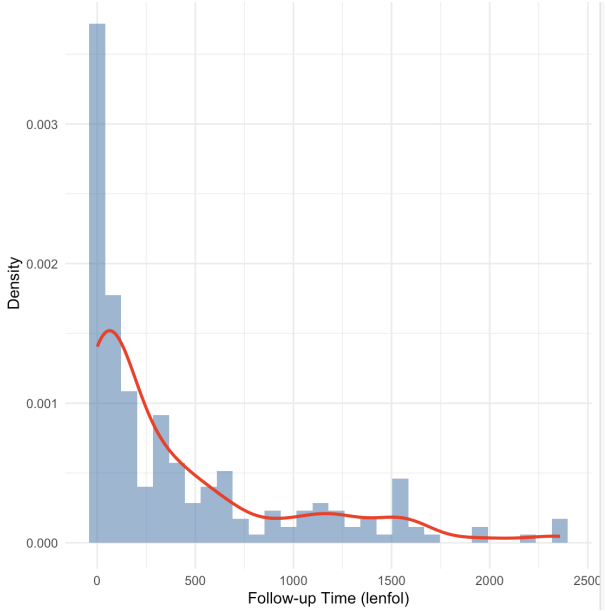

Figure 1: Histogram and Density of Survival Times

From figure 1, we can see that the distribution has a rapid decline with a heavy concentration near zero, leading me to believe it is approximately exponential.

Next I want to check the cdf of this data and estimate the probability of surviving up to 200 days. The empirical cdf of the survival times can be seen in figure 2.

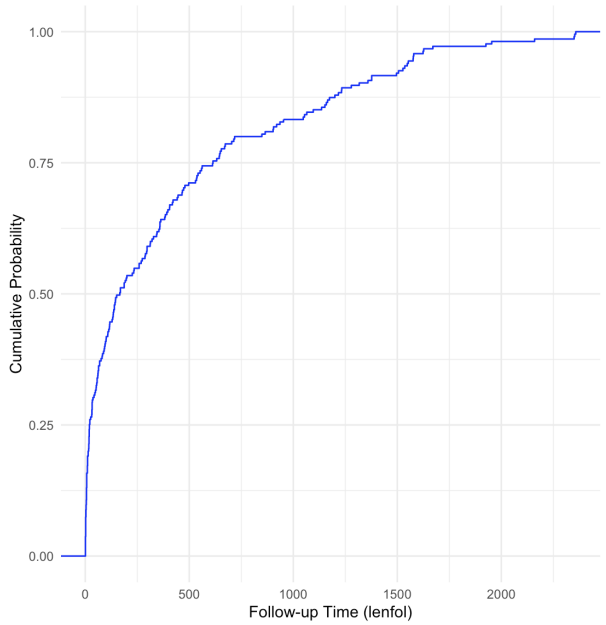

Figure 2: Empirical CDF of the Survival Times

The cdf further confirms the assumption that this data follows an exponential distribution. When looking at the probability of surviving up to 200 days, I get a value of roughly 0.535. Meaning we have a 53.5% chance of surviving up to 200 days.

When now looking at the estimation of the survival function (1-ecdf) we want to answer the question of what the probability of surviving beyond 1000 days is. In figure 3, we can see the estimated survival function.

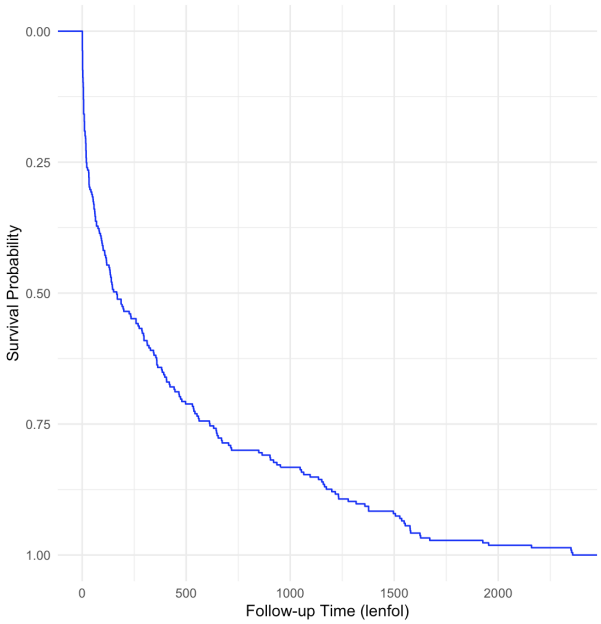

Figure 3: Estimated Survival Function

The probability of surviving beyond 1000 days is roughly 16.7%.

Code:

```
# Read the dataset
whas500 <- read.csv("whas500.csv", sep = ";")

# Preliminaries
deceased <- whas500 %>% filter(FSTAT == 1)

histogram <- ggplot(deceased, aes(x = LENFOL)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "steelblue",
                 alpha = 0.6) +
  geom_density(color = "red", size = 1) +
  labs(title = "Histogram and Density of Survival Times",
       x = "Follow-up Time (lenfol)", y = "Density") +
  theme_minimal()

print(histogram)

cdf <- ggplot(deceased, aes(x = LENFOL)) +
  stat_ecdf(geom = "step", color = "blue") +
  labs(title = "Empirical CDF of Survival Times",
       x = "Follow-up Time (lenfol)", y = "Cumulative Probability") +
  theme_minimal()

print(cdf)

prob_200 <- mean(deceased$LENFOL <= 200)
print(prob_200)

survival_function_estimate <- ggplot(deceased, aes(x = LENFOL)) +
  stat_ecdf(geom = "step", color = "blue") +
  scale_y_reverse() + # Reverse the y-axis for survival function
  labs(title = "Estimated Survival Function",
       x = "Follow-up Time (lenfol)", y = "Survival Probability") +
  theme_minimal()

print (survival_function_estimate)

prob_1000 <- mean(deceased$LENFOL > 1000)
print(prob_1000)
```

Survival analysis

Before we start with the survival analysis, we want to get familiar with the data. For that, we first look at the descriptive statistics of the columns "LENFOL", "AGE", "BMI", and "HR". The resulting table can be seen in figure 4.

Subjects	Mean_lenfol	SD_lenfol	Min_lenfol	Max_lenfol	Mean_age	SD_age	Min_age	Max_age	Mean_bmi	SD_bmi	Min_bmi	Max_bmi	Mean_hr	SD_hr	Min_hr	Max_hr
500	882.436	705.6651	1	2358	69.846	14.49146	30	104	26.61378	5.405655	13.04546	44.83886	87.018	23.58623	35	186

Figure 4: Descriptive Statistics of Columns "LENFOL", "AGE", "BMI", and "HR"

When analyzing this table, I see a few key takeaways. The most important are summarized below.

- The survival time varies greatly. Some subjects died within days, while others lived for over 6 years.
- The age range is wide, but the average is nearly 70, making this a primarily elderly population.
- The BMI varies significantly. Some subjects are very underweight, while others seem to be obese.
- The heart rate shows large differences, which could indicate different health conditions.

Now we also want to produce a "pairs" plot for the columns "LENFOL", "GENDER", "AGE", "BMI", and "HR" for survivors as well as non-survivors. This "pairs" plot can be seen in figure 5.

Figure 5: Pairs Plot of Columns "LENFOL", "GENDER", "AGE", "BMI", and "HR"

First let's explain the general structure of the "pairs" plot. The diagonal elements show density plots for each variable, split by survival status (red = non-survivors, blue = survivors). The upper triangle displays correlation coefficients between variables. The lower triangle shows scatter plots for numerical variables, with color-coded points based on survival status.

Now let's look at some key observations from the plot.

- Survival Time (LENFOL): First we can see that the density plot shows a multi-model distribution. This could suggest different subgroups with varying survival times. We also see a correlation with age and a tendency for older individuals to have shorter survival times. Contrary to age, with BMI we have slight a positive correlation, suggesting higher BMI might be associated with longer survival. Lastly, we have a slight negative correlation with hart rate. Higher heart rates are weakly associated with shorter survival.

- Gender: There are two take-aways for gender in this plot. First, the boxplots for gender suggests that men (blue) have a wider range of survival times compared to women (red). And second, the gender distributions seems roughly balanced.

- Age: Again two take-aways. Older individuals tend to have lower BMI, possibly due to frailty or muscle loss. Older individuals tend to have slightly higher heart rates, which could indicate declining cardiovascular health.

- BMI: The only thing I can take-away from the plot for BMI is very surprising. The data shows a weak correlation with survival time, suggesting a protective effect of BMI in certain cases.

- Heart Rate: The only take-away for heart rate I have, is that higher heart rate is weakly associated with shorter survival times.

2.1 Non-Parametric Analysis

First we want to have a look at the output of the survfit() function with different options. For the output of the survfit() function with default parameters, I get the following results: n = 500, events = 215, median = 1627, 0.95LCL = 1527, 0.95UCL = NA. We can interpret these as having 500 subjects in the study with 215 occurring events and a median survival time of 1627. The confidence limits in this case are calculated using the default log-transformation. That the 0.95UCL is NA in this case could be because of high censoring or few events.

When looking at the same output with the conf.type set to "log-log" the CLs are calculated using the log-log transformation, which tends to produce more accurate and wider confidence intervals, especially when survival probabilities are close to 1. Now we get the following output: n = 500, events = 215, median = 1627, 0.95LCL = 1506, 0.95UCL = 2353. The number of subjects, events and median survival time unsurprisingly stay the same. But the confidence level change and we can see that now we have a wider (0.95LCL is decreased) confidence interval and a valid value for 0.95UCL.

We now want to look at the survival curve and the hazard curve. These can be seen in figures 6 and 7 respectively.

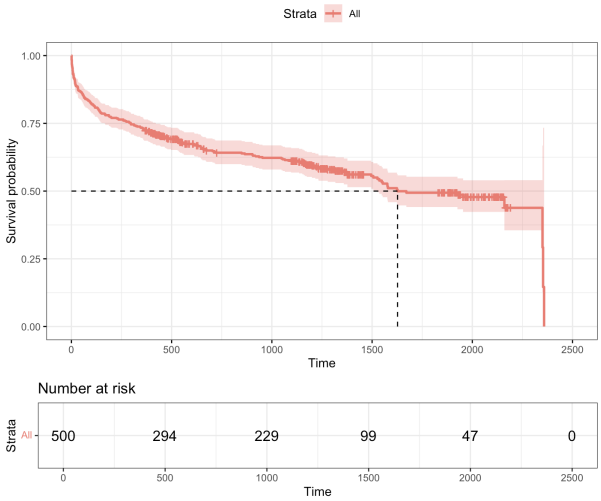

Figure 6: Survival Curve

The survival curve shows us that we have a relative steep decline in survival probability in the beginning and that this decline in survival probability, with some bumps, slows down over time. We can also see the median survival time (1627) in the dotted black line. Lastly, we see that the survival curve drops to zero at some event near time 2350. This means that there is no censoring at the last recorded time, meaning that every remaining subject experiences the event.

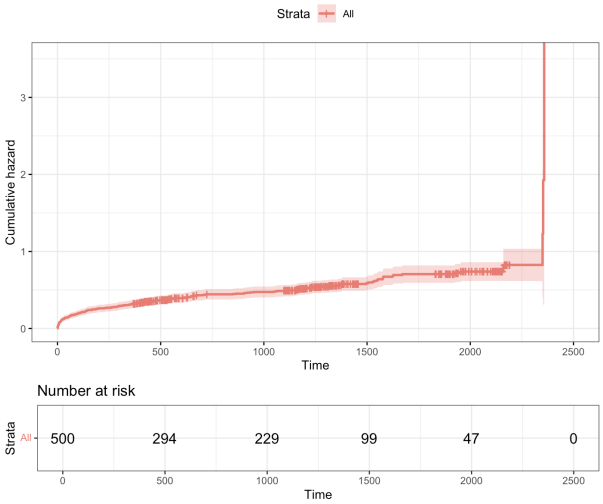

Figure 7: Hazard Curve

The hazard function measures instantaneous risk of an event at time t, given survival up to t. The hazard curve shows similarity to the survival curve in that it rises quickly and then flattens to a more steady rise. Although it is obvious that the rise flattens more quickly here than in the survival curve. We saw that the survival curve dropped to zero at the last recorded time because there is no censoring at the last recorded time. We can also see this in the hazard curve, where the curve spikes drastically at the last recorded time.

We now want to compare the survival time between genders. We use both the Kaplan-Meier estimators as well as the log-rank test to investigate this. The results of these can

be seen in figure 8.

```
                n events median 0.95LCL 0.95UCL
GENDER=0 300    111   2160    1624      NA
GENDER=1 200    104   1317     865    1579
Call:
survdiff(formula = surv.data ~ GENDER, data = whas500)

                N Observed Expected (O-E)^2/E (O-E)^2/V
GENDER=0 300      111    130.7      2.98      7.79
GENDER=1 200      104     84.3      4.62      7.79

 Chisq= 7.8  on 1 degrees of freedom, p= 0.005
```

Figure 8: Comparing Survival between Genders

When looking at the Kaplan-Meier estimators of both gender 0 (male) and gender 1 (female), we can see a slight uneven distribution and a big discrepancy between median survival times. We can see that male subjects tend to survive much longer than female subjects. The confidence intervals show that again we have a lower 0.95LCL for gender 1, showing that women don't survive as long, and that we have a NA 0.95UCL for gender 0, meaning many males survived past the study period.

The log-rank test confirms these findings. The p-value (0.005) is statistically significant, meaning survival differs significantly between genders. Also, the expected survival is worse for gender 1 than for gender 0.

To verify that directly optimizing the log-partial likelihood function and using the Cox model, give the same estimate for beta, we implement this function in R: $l(\beta) = \beta - \log(3e^{\beta} + 3) - \log(3e^{\beta} + 1) - \log(2e^{\beta} + 1)$. We first want to plot this function to get an estimate of where β should be. This plot can be seen in figure 9.

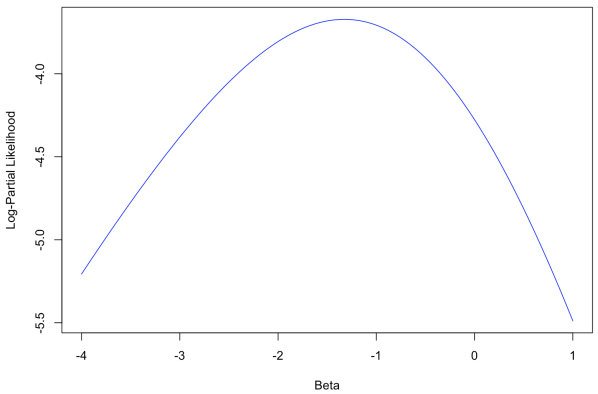

Figure 9: Log-Partial Likelihood Function

When optimizing this function, we get the estimated β is -1.3261, meaning that being in group T (treatment) is associated with a lower hazard (risk of failure) compared to group C (control). We also get the log-partial likelihood estimate of -3.672, which tells us how well this parameter explains the data. When we do the same now with the model, the coxph function also estimates β = -1.3261 and get the same exact log-partial likelihood estimate, confirming our manual calculation.

To understand the behavior of the covariates age and gender, we use the Cox proportional hazards model. The summary of the fit of this model can be seen in figure 10.

```
             coef exp(coef)  se(coef)      z Pr(>|z|)
GENDER1 -0.066285  0.935864  0.140585 -0.471    0.637
AGE      0.066928  1.069218  0.006196 10.802   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
GENDER1    0.9359     1.0685    0.7105     1.233
AGE        1.0692     0.9353    1.0563     1.082

Concordance= 0.731  (se = 0.018 )
Likelihood ratio test= 142.4  on 2 df,   p=<2e-16
Wald test            = 119.7  on 2 df,   p=<2e-16
Score (logrank) test = 126.9  on 2 df,   p=<2e-16
```

Figure 10: Summary of Cox Model

Let's first look at the question, if there are differences in survival by gender or age. For gender we have a p-value of 0.637, meaning gender does not significantly affect survival in this model. For age, we have a highly significant p-value of lower than 2e-16, meaning age significantly impacts survival.

The different statistical tests that are available in the summary are the likelihood ratio test, the Wald test, and the score or log-rank test. We see for all that the p-value is lower than 2e-16, meaning the model is statistically significant. The likelihood ratio test tests this by comparing two models, one with all specified covariates (full model) and one without (null model). Our p-value shows that the given model significantly improves survival prediction compared to a model with no covariates. The Wald test tests each covariate individually by checking if its coefficient is significantly different from zero. Because we have an individual covariate that is statistically significant (age) the Wald test is significant. The score test tests the overall model without estimating full model parameters, making it computationally more efficient. The p-value gives us the same overall information as the p-value of the LRT test but is based on how the likelihood would change if a variable were included.

We now want to plot the Cox survival curve adjusted by age. The Kaplan-Meier survival curve gave us a "pure" survival time based on the observed data (including any variations in age). The Cox regression survival curve now gives us the survival time adjusted for the covariate age. Since we are holding age constant at its mean value, we get the effect of gender on survival while controlling for the influence of age. The Cox survival curve adjusted by age can be seen in figure 11.

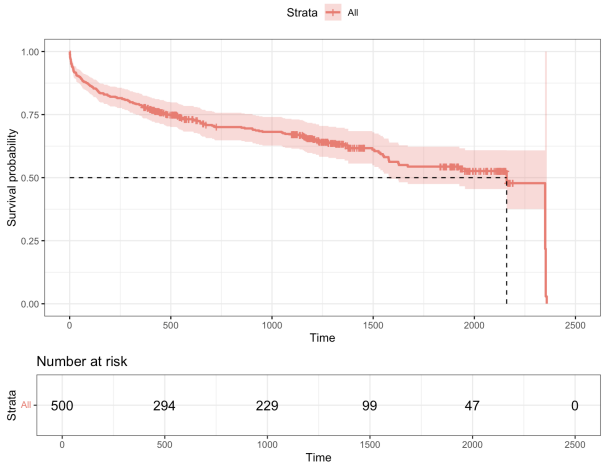

Figure 11: Cox Survival Curve adjusted by Age

In the Cox survival curve adjusted for age, we see that the actual mean survival probability, when assuming all subjects are the mean age, is actually a lot higher than the one we saw in the Kaplan-Meier survival curve. This confirms that age is a major risk factor for shorter survival.

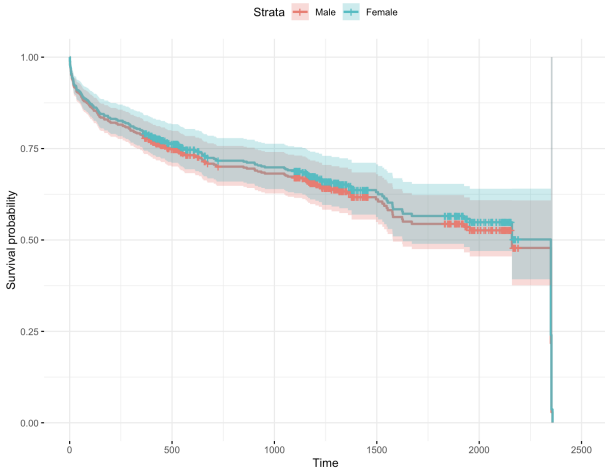

Figure 12: Cox Survival Curve adjusted by Age by Gender

In figure 12 we can see a visualization of the Cox survival curve adjusted by age by gender. The interesting thing here is that when adjusting for age, the female survival probability is actually higher than the male survival probability. If we recall, the analysis on the Kaplan-Meier estimators with the log-rank test showed us the opposite. Here the test showed us that male subjects tended to survive much longer than female subjects. After adjusting for age, we can now say that that is not the case.

Code:

```
# Data exploration
summary_table <- whas500 %>%
  summarise(
    Subjects = n(),
    Mean_lenfol = mean(LENFOL),
    SD_lenfol = sd(LENFOL),
    Min_lenfol = min(LENFOL),
    Max_lenfol = max(LENFOL),

    Mean_age = mean(AGE),
    SD_age = sd(AGE),
    Min_age = min(AGE),
    Max_age = max(AGE),

    Mean_bmi = mean(BMI),
    SD_bmi = sd(BMI),
    Min_bmi = min(BMI),
    Max_bmi = max(BMI),

    Mean_hr = mean(HR),
    SD_hr = sd(HR),
    Min_hr = min(HR),
    Max_hr = max(HR)
  )

print(summary_table)

whas500$GENDER <- as.factor(whas500$GENDER)

pairs_plot <- ggpairs(
                  whas500,
                  columns = c("LENFOL", "GENDER", "AGE", "BMI", "HR"),
                  aes(color = as.factor(FSTAT)),
                  title = "Pairs Plot: Survivors vs Non-Survivors"
                )

print(pairs_plot)

# Survival analysis
# Non-parametric analysis
surv.data <- with (whas500 , Surv ( LENFOL , FSTAT ))
fit <- survfit (surv.data~1 , data = whas500)
# print(fit)
fit_log_log <- survfit (surv.data~1 ,
                        data = whas500,
                        conf.type="log-log")
# print(fit_log_log)
```

```
survival_function <- ggsurvplot (fit ,
                         risk.table = TRUE,
                         surv.median.line = "hv",
                         ggtheme = theme_bw())
print(survival_function)

hazard_function <- ggsurvplot (fit ,
                       fun = "cumhaz",
                       risk.table = TRUE,
                       ggtheme = theme_bw())
print(hazard_function)

fit_gender <- survfit (surv.data~GENDER,
                         data = whas500,
                         conf.type="log-log")
print(fit_gender)
surv.test <- survdiff(surv.data~GENDER, data = whas500)
print(surv.test)

# Parametric analysis with Cox regression
# Estimating beta
pll <- function(beta) {
  return(beta - log(3 * exp(beta) + 3) -
        log(3 * exp(beta) + 1) -
        log(2 * exp(beta) + 1))
}

beta_values <- seq(-4, 1, length.out = 100)

pll_values <- sapply(beta_values, pll)

plot(beta_values, pll_values, type = "l", col = "blue",
     xlab = "Beta", ylab = "Log-Partial-Likelihood",
     main = "Log-Partial-Likelihood-Function")

result <- optim(par = 0, fn = pll, method = "L-BFGS-B",
                control = list(fnscale = -1), # Maximization
                lower = -3, upper = 1)

beta_hat <- result$par
log_likelihood_max <- result$value
print(beta_hat)
print(log_likelihood_max)

data <- data.frame(
  Survtime = c(6, 7, 10, 15, 19, 25),
  Censor = c(1, 0, 1, 1, 0, 1),
```

```
  Group = c("C", "C", "T", "C", "T", "T")
)

cox_model <- coxph(Surv(Survtime, Censor) ~ Group, data = data)
print(summary(cox_model))
print(logLik(cox_model))

# Back to the whas500 dataset
fit_cox <- coxph(surv.data~GENDER+AGE, data=whas500)
print(summary(fit_cox))

surv.fit <- survfit(fit_cox, data=whas500)
survival_plot <- ggsurvplot(surv.fit,
            risk.table = TRUE,
            surv.median.line = "hv",
            ggtheme = theme_bw())
print(survival_plot)

sex.dt <- with(whas500,
# I changed the next line from c("Male", "Female") to ensure structure
                data.frame(GENDER = c(0, 1),
                           AGE = rep(mean(AGE, na.rm = TRUE), 2)))
sex.dt$GENDER <- factor(sex.dt$GENDER, levels = levels(whas500$GENDER))
surv.fit.gender <- survfit(fit_cox, data=whas500, newdata=sex.dt)
survival_plot2 <- ggsurvplot(surv.fit.gender,
                     legend.labs = c("Male", "Female"),
                     conf.int = TRUE,
                     ggtheme = theme_minimal())
print(survival_plot2)
```