SURVIVAL ANALYSIS WITH THE SURVIVAL PACKAGE

Survival analysis with the survival package

In this instruction you will:

- work with **time-to-event** data,
- use core functions from the `survival` package,
- visualise survival curves,
- fit and interpret a **Cox proportional hazards model**.

We will use a built-in dataset (`lung`) that ships with the `survival` package.

1. Setup and basic concepts

Install and load the package:

```
install.packages("survival")
library(survival)
```

Key ideas:

- **Survival time**: time from a defined origin (e.g. diagnosis, training start) to an event (e.g. death, failure, dropout).
- **Event indicator**: `1` if event occurred, `0` if right-censored (we only know the event did not occur up to a certain time).
- **Survival function** $S(t)$: probability of "surviving" beyond time t.
- **Hazard**: instantaneous risk of event at time t, given survival up to t.

2. The lung dataset and Surv objects

Inspect the built-in `lung` dataset:

```
data(lung)
?lung
head(lung)
str(lung)
summary(lung)
```

We will use:

- `time` – survival time in days,
- `status` – event indicator (`1` = censored, `2` = death),
- `sex` – `1 = male`, `2 = female`,
- `age` – age in years.

Create a `Surv` object:

```
lung$event <- ifelse(lung$status == 2, 1, 0)  # 1 = death, 0 = censored

surv_obj <- Surv(time = lung$time, event = lung$event)
head(surv_obj)
```

Tasks 2.x

Task 2.1
Check how many censored vs event observations there are (`table(lung$event)`).

Task 2.2
Compute the median and quartiles of `time` separately for men and women.

3. Kaplan–Meier survival curves and visualisation

3.1 Overall survival curve

Fit a Kaplan–Meier estimator with `survfit`:

```
km_all <- survfit(surv_obj ~ 1, data = lung)
summary(km_all)
```

Plot:

```
plot(km_all,
     xlab = "Days",
     ylab = "Survival probability",
     main = "Kaplan-Meier curve (all patients)",
     col = "steelblue",
     lwd = 2,
     conf.int = TRUE)  # add confidence bands
```

3.2 Survival curves by group (sex)

Fit separate curves by `sex`:

```
lung$sex_factor <- factor(lung$sex, levels = c(1, 2), labels = c("Male",

km_sex <- survfit(Surv(time, event) ~ sex_factor, data = lung)
summary(km_sex)
```

Plot both curves:

```
plot(km_sex,
     col = c("red", "blue"),
     lwd = 2,
     xlab = "Days",
     ylab = "Survival probability",
     main = "Kaplan-Meier curves by sex")
legend("topright",
       legend = levels(lung$sex_factor),
       col = c("red", "blue"),
       lwd = 2,
       bty = "n")
```

Tasks 3.x

Task 3.1
Read off from the plot (or `summary(km_sex)`) the median survival time for men and for women.

Task 3.2
Add a grid to the plot (e.g. using `grid()` after `plot`) and comment briefly on which group appears to have better survival.

4. Cox proportional hazards model

The Cox model relates the **hazard** to covariates:

$$h(t \mid x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots),$$

where:

- $h_0(t)$ is the baseline hazard,
- x_i are covariates (e.g. sex, age),
- $\exp(\beta)$ are **hazard ratios**.

4.1 Fit a simple Cox model

Model with sex only:

```
cox_sex <- coxph(Surv(time, event) ~ sex_factor, data = lung)
summary(cox_sex)
```

Key outputs:

- coefficient for `sex_factorFemale`,
- `exp(coef)` (hazard ratio),
- confidence intervals,
- Wald test p-value.

4.2 Add age as a continuous covariate

```
cox_sex_age <- coxph(Surv(time, event) ~ sex_factor + age, data = lung)
summary(cox_sex_age)
```

Interpretation examples:

- `exp(coef)[ "sex_factorFemale" ]` – relative hazard for females vs males after adjusting for age.
- `exp(coef)[ "age" ]` – multiplicative change in hazard for a 1-year increase in age.

4.3 Visualising Cox model results

You can plot **adjusted survival curves** for typical covariate values:

```
newdata <- data.frame(
  sex_factor = factor(c("Male", "Female"), levels = levels(lung$sex_fact
  age = rep(median(lung$age, na.rm = TRUE), 2)
)

fit_surv <- survfit(cox_sex_age, newdata = newdata)

plot(fit_surv,
     col = c("red", "blue"),
     lwd = 2,
     xlab = "Days",
     ylab = "Survival probability",
     main = "Adjusted survival curves (median age)")
legend("topright",
       legend = c("Male", "Female"),
       col = c("red", "blue"),
       lwd = 2,
       bty = "n")
```

Tasks 4.x

Task 4.1
From `summary(cox_sex_age)`, extract:

1. the hazard ratio for females vs males,
2. the hazard ratio for a 10-year age increase (hint: `exp(10 * coef["age"])`).

Task 4.2
Write 3–4 sentences interpreting these hazard ratios in plain language.

5. Checking the proportional hazards assumption

The Cox model assumes that hazard ratios are **constant over time**. We can check this with `cox.zph()`:

```
ph_test <- cox.zph(cox_sex_age)
ph_test
plot(ph_test)
```

Interpretation:

- p-values in `cox.zph` test for deviations from proportional hazards,
- plots show scaled Schoenfeld residuals over time.

Tasks 5.x

Task 5.1
Based on `cox.zph(cox_sex_age)`, comment whether the proportional hazards assumption appears reasonable for:

- `sex_factor`,
- `age`.

Task 5.2 (optional)
Try fitting a model with an additional covariate (e.g. `ph.ecog` performance status) and repeat the `cox.zph` check.

6. Optional: Synthetic survival data

To better understand model behaviour, you can simulate your own survival data.

Example with one binary covariate (`treatment`) and exponential baseline hazard:

```
set.seed(101)
n <- 300
treatment <- rbinom(n, size = 1, prob = 0.5)

lambda0 <- 0.002          # baseline rate
beta_trt <- log(0.6)      # hazard ratio ~ 0.6 for treatment vs control

u <- runif(n)
time_sim <- -log(u) / (lambda0 * exp(beta_trt * treatment))

censor_time <- rexp(n, rate = 0.0008)
time_obs <- pmin(time_sim, censor_time)
event_sim <- as.integer(time_sim <= censor_time)

sim_df <- data.frame(
  time = time_obs,
  event = event_sim,
  treatment = factor(treatment, levels = c(0, 1), labels = c("Control",
)

sim_cox <- coxph(Surv(time, event) ~ treatment, data = sim_df)
summary(sim_cox)
exp(coef(sim_cox))
```

Task 6.1 (optional)

1. Compare the estimated hazard ratio from `sim_cox` with the true value `exp(beta_trt)` $\simeq$ `0.6`.
2. Create Kaplan–Meier curves by `treatment` and comment on how well they reflect the hazard ratio.

7. Wrap-up

In this lab you:

- created and used `Surv` objects,
- computed and visualised Kaplan–Meier survival curves,
- fitted and interpreted a Cox proportional hazards model,
- briefly checked the proportional hazards assumption.

These tools are the foundation of practical survival analysis in R using the `survival` package.