Medical Statistics – Lab 9

R version

Part 1: Analysis of overall survival in the Worcester Heart Attack study

In this section, we are going to continue analyzing the Worcester Heart Attack Study (WHAS) dataset (file `whas500.sav`). The outcome of interest for today's analysis is overall survival, defined as the time from hospital admission to death from any cause. This information is captured in the variable `lenfol` (length of follow-up in days) and the variable `fstat` (follow-up status; dead or censored).

```
library(haven)   # for reading SPSS files
library(dplyr)   # for data manipulation
library(survival)  # for performing survival analysis

# Load the dataset
whas500 <- read_sav("datasets/whas500.sav")

# Convert labeled variables to factors
whas500 <- whas500 %>%
  mutate(across(where(is.labelled), as_factor))

# Create a numerical variable for follow-up status (1 = dead, 0 = censored)
whas500$fstat_numeric <- ifelse(whas500$fstat=="dead", 1, 0)
```

Association between MI order and overall survival

The variable `miord` represents the sequence of myocardial infarction (MI) events, categorized as either a first MI or a recurrent MI. Our aim is to analyze the relationship between MI order and overall survival outcomes.

Kaplan-Meier survival curves and logrank test

We start by constructing a Kaplan-Meier survival curve to compare the survival probabilities between patients with a first MI and those with a recurrent MI:

```
# Kaplan-Meier by MI order
group_fit <- survfit(Surv(lenfol, fstat_numeric) ~ miord, data = whas500)
summary(group_fit)

plot(group_fit, col = c("blue", "red"), xlab = "Follow-up Time (days)", ylab = "Survival Prob
legend("topleft", legend = c("First MI", "Recurrent MI"), col = c("blue", "red"), lty = 1)
```

Explanation of the code:

- The `survfit()` function is used to fit a Kaplan-Meier survival curve for the two groups defined by the `miord` variable.
- The formula `Surv(lenfol, fstat) ~ miord` specifies the survival time and event status as the response variables and the MI order as the predictor.
- `summary(group_fit)` provides the survival probabilities at different time points for each group.
- The `plot()` function is used to visualize the Kaplan-Meier curves for the two groups.

> Question 1
>
> Based on the Kaplan-Meier table, what are the estimated survival probabilities at 3 years for patients with a first MI and those with a recurrent MI?

> Question 2
>
> Based on the Kaplan-Meier curves, do you observe any differences in survival times between patients with a first MI and those with a recurrent MI?

To formally test the difference in survival between the two groups, we can use the logrank test:

```
# Logrank test
survdiff(Surv(lenfol, fstat_numeric) ~ miord, data = whas500)
```

> Question 3
>
> Based on the results of the logrank test, is there a significant difference in overall survival between patients with a first MI and those with a recurrent MI?

Cox regression

Next, we will perform a Cox proportional hazards regression analysis to assess the association between MI order and overall survival while adjusting for potential confounders. We will start with the unadjusted model:

```
# Unadjusted Cox regression
coxph_model_unadj <- coxph(Surv(lenfol, fstat_numeric) ~ miord, data = whas500)
summary(coxph_model_unadj)
```

Explanation of the code:

- The `coxph()` function is used to fit a Cox proportional hazards regression model with the survival time and event status as the response variables and the MI order as the predictor.
- The `summary()` function provides the estimated hazard ratio (HR) and its significance.

> Question 4
>
> What is the hazard ratio (HR) for patients with a recurrent MI compared to those with a first MI based on the unadjusted Cox regression model?

> Question 5
>
> Does the result of the Cox regression model support the findings from the logrank test regarding the association between MI order and overall survival?

Now, let's adjust the Cox regression model using `age` and `gender` as potential confounders:

```
# Adjusted Cox regression
coxph_model_adj <- coxph(Surv(lenfol, fstat_numeric) ~ miord + age + gender, data = whas500)
summary(coxph_model_adj)
```

> Question 6
>
> After adjusting for age and gender, what is the hazard ratio (HR) for patients with a recurrent MI compared to those with a first MI? How does this compare to the unadjusted HR? Can you explain the change in the HR after adjusting for these variables?

Part 2: Unguided exercises

Excercise 1

File `Ex9_1.sav` (available from the Datasets menu) contains data from a small experiment concerning motion sickness at sea (Burns, 1984). Subjects were placed in a cabin subjected to vertical motion for two hours. The outcome variable was the waiting time to emesis (vomiting). Some subjects requested an early stop to the experiment although they had not vomited, yielding censored observations, while others successfully survived two hours. The experiment was carried out with two "treatments": two combinations of movement accelerations and frequency. One combination was used for a group of 21 subjects, the other in a different group of 28 subjects.

(a) Calculate and plot Kaplan-Meier estimates of survival probabilities in the two groups.

(b) Calculate the 95% CI for the difference between survival probabilities of the two groups after 60 minutes.

(c) Compare the two survival curves by logrank test.

(d) Use Cox regression to compare the two treatments; compare the result to that of the logrank test; calculate the hazard ratio and its 95% CI.

(e) Two different persons undergo the experiment with different treatments. Estimate the probability that the waiting time until emesis under one of the treatments exceeds that under the other treatment.

Exercise 2

Subfertile women with a child wish may receive an in-vitro fertilization (IVF) treatment. In an observational study the waiting time until pregnancy was recorded. The women undergoing the IVF treatment were categorized (prior to the start of the treatment) in four groups, A, B, C and D, with respect to the type of infertility. The waiting times – some of them censored - of the groups were compared by means of Cox regression with the group variable entered as a categorical variable. The P-value of the Wald test with 3 df was 0.095. The dummy-variables v1, v2 and v3 were defined as follows:

	v1	v2	v3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

The hazard ratios and P-values for the dummy variables v1, v2 and v3 were reported as 1.20 (P=0.30), 2.80 (P=0.02) and 1.05 (P=0.48).

(a) Which group is estimated to have the longest waiting time for pregnancy?

(b) A researcher compared all pairs of groups by separate tests. She used the Bonferroni method to keep the type-I error of the entire procedure below 10%. Which differences were found to be significant?

(c) Which assumptions are needed for the validity of Cox regression in this case?

Exercise 3

Relation between survival and a number of variables was studied in 37 patients having a bone marrow transplant. Cox regression analysis using the occurrence of acute graft-versus-host disease (GvHD=1 if present and GvHD=0 if absent), diagnosis, recipient's age and sex, donor's age and sex, whether the donor had been pregnant and the type of the leukemia (CML=1 if chronic myeloid leukemia and CML=0 otherwise) yielded the following model:

Variable	Regression coefficient	Standard error
GvHD(0 = No, 1= Yes)	2.306	0.5898
CML (0 = No, 1 = Yes)	-2.508	0.8095

(a) What is the interpretation of the opposite signs for the regression coefficients?

(b) Calculate the relative risks of dying (hazard ratio) for the following patients relative to non-GvHD non-CML patients (i) with GvHD but not CML, (ii) CML but without GvHD, (iii) CML and GvHD.