

Exercises day 5 – Survival and event history analysis

Data

The analysis is based on the Nepal study data presented in the lecture. We look at the time to the next occurrence of pneumonia. The subjects were children recruited to the study immediately after they had recovered from their first case of pneumonia. The children were under observation over a certain period, some for about 3 months, others for about 6 months. The time variable `time_14` measures the time from inclusion to the next occurrence of pneumonia, or until the end of follow-up, whatever comes first. If the end of the follow-up comes first, the time observation is said to be censored. If pneumonia comes first the child is said to have had an event. We assume that the probability of being censored due to loss to follow-up or death from causes other than the one being studied is unrelated to the probability of pneumonia, i.e. independent censoring. The children were randomized to two treatments, either placebo or zinc.

Load the data file:

```
pneumonia <- read_csv("Pneumonia_time.csv")
```

There is a number of variables in the data set. The variables we will use are:

age: Age of the child (in months)

sex: Sex of the child (0 = boy, 1 = girl)

treat_1: Treatment by zinc (0 = placebo, 1 = zinc)

time_14: Time since last visit (or time to next pneumonia)

event: Event (0 = censored, 1 = pneumonia)

```
# Load packages  
library(tidyverse)  
library(ggthemes)  
library(survival)  
library(survminer)
```

Describe data

Questions 1) and 2) below give rough descriptive statistics of the data file. They do not appropriately account for censoring.

- 1) Get an overview of the variables in the data file. How many rows and columns? How many variables? How many observations? What is the mean age of the children? What is the percentage of boys? How many pneumonia cases? How many pneumonia cases were treated with zinc? What is the median time to next pneumonia?

```
glimpse(pneumonia)
```

```
View(pneumonia)
```

```
str(pneumonia)
```

```
summary(pneumonia)
```

```
head(pneumonia)
```

```
tail(pneumonia)
```

```
typeof(pneumonia)
```

```
class(pneumonia)
```

```
summary(pneumonia$age)
```

```
hist(pneumonia$age)
```

```
pneumonia |>
```

```
  count(sex) |>
```

```
  mutate(Percentage = round(100 * n / sum(n),1))
```

```
table(pneumonia$event, pneumonia$treat_1)
```

```
summary(pneumonia$time_14)
```

```
hist(pneumonia$time_14)
```

- 2) Look at the distribution of time to pneumonia separately by treatment and sex. Does it look like the distribution differ by treatment or sex?

```
by(pneumonia$time_14, pneumonia$treat_1, summary)
```

```
ggplot(
```

```
  data = pneumonia,
```

```

    mapping = aes(x = time_14)
  ) +
    geom_histogram() +
    facet_wrap(~ factor(treat_1, labels = c("Placebo",
"Zinc")))) +
    theme_minimal() +
    labs(x = "Time since last visit ", y = "Frequency",)

```

(similarly for sex)

Kaplan-Meier analyses

- 3) Create a single overall Kaplan-Meier survival curve. Estimate (roughly) the median survival time based on the Kaplan-Meier plot. What can you say about censoring based on the plot?

```
KM <- survfit(Surv(time_14, event) ~ 1, data = pneumonia)
```

```

ggsurvplot(
  KM,
  data = pneumonia
)

```

```

ggsurvplot(
  KM,
  data = pneumonia,
  risk.table = TRUE,
)

```

- 4) Create and compare Kaplan-Meier survival curves in the two groups with and without treatment of zinc. What is your conclusion?

```

KM.treat <- survfit(Surv(time_14, event) ~ treat_1, data =
pneumonia)
print(KM.treat)
KM.treat

```

```

ggsurvplot(
  KM.treat,
  data = pneumonia,
)

```

```
ggsurvplot(
  KM.treat,
  data = pneumonia,
  conf.int = TRUE,
  legend.title = "",
  legend.labs = c("Placebo", "Zinc")
)
```

```
ggsurvplot(
  KM.treat,
  data = pneumonia,
  conf.int = TRUE,
  legend.title = "",
  legend.labs = c("Placebo", "Zinc"),
  risk.table = TRUE,
  pval = TRUE
)
```

- 5) Create and compare Kaplan-Meier survival curves by sex. What is your conclusion?

```
KM.sex <- survfit(Surv(time_14, event) ~ sex, data =
pneumonia)
print(KM.sex)
KM.sex
```

```
ggsurvplot(
  KM.sex,
  data = pneumonia
)
```

```
ggsurvplot(
  KM.sex,
  data = pneumonia,
  conf.int = TRUE,
  legend.title = "",
  legend.labs = c("Boy", "Girl")
)
```

```
ggsurvplot(
  KM.sex,
  data = pneumonia,
```

```

    conf.int = TRUE,
    legend.title = "",
    legend.labs = c("Boy", "Girl"),
    risk.table = TRUE,
    pval = TRUE
)

```

- 6) Test for a significant difference between two or several survival curves using the log-rank and Wilcoxon (Gehan-Wilcoxon) tests. Explain the output. Is there any statistically significant difference between the two treatments? What test (logrank or Wilcoxon test) would you prefer here and why?

```

wilcox <- survdiff(
  Surv(time_14, event) ~ treat_1,
  data = pneumonia,
  rho = 1)

```

```
wilcox
```

```

logrank <- survdiff(
  Surv(time_14, event) ~ treat_1,
  data = pneumonia,
  rho = 0)

```

```
logrank
```

Cox regression

The survival times depend on other factors such as age, sex, etc. We can take these factors into account when estimating or testing for a difference between two survival curves by using Cox regression.

- 7) Estimate the effect of treatment (treat_1) on the risk of pneumonia. Compare this results with the results in exercise 4.

```

pneum.cox <- coxph(Surv(time_14, event) ~ treat_1,
  data=pneumonia)

```

```
summary(pneum.cox)
```

- 8) Create the survival curves based on the Cox regression. Compare these curves with the survival curves in exercise 5.

```
fit.cox <- survfit(  
  pneum.cox,  
  newdata = data.frame(treat_1 = c(0,1)  
)
```

```
ggsurvplot(  
  fit.cox,  
  data = pneumonia  
)
```

```
ggsurvplot(  
  fit.cox,  
  data = pneumonia,  
  legend.title = "",  
  legend.labs = c("Placebo", "Zinc")  
)
```

- 9) Taking into account age and sex in the Cox regression model, estimate the effect of treatment on the risk of pneumonia:

```
pneum.cox.adj <- coxph(Surv(time_14, event) ~ treat_1 + age +  
sex, data=pneumonia)
```

```
summary(pneum.cox.adj)
```

```
fit.cox.adj <- survfit(  
  pneum.cox.adj,  
  newdata = data.frame(  
    treat_1 = c(0,1)  
    age = mean(pneumonia$age, na.rm = TRUE),  
    sex = mean(pneumonia$sex, na.rm = TRUE)  
  )  
)
```

```
ggsurvplot(  
  fit.cox.adj,  
  data = pneumonia
```

```
)
```

```
ggsurvplot(  
  fit.cox.adj,  
  data = pneumonia,  
  legend.title = "",  
  legend.labs = c("Placebo", "Zinc")  
)
```

Note: In this file, age is age at time of inclusion. We might prefer to adjust for age as a time-dependent covariate, but we ignore that here.

- 10) An important assumption to check is the proportional hazard assumption. Use the following commands and give your opinion about the proportional hazard assumption.

```
cox.zph(pneum.cox.adj)
```

```
plot(cox.zph(pneum.cox.adj))
```