Group 1: Wildfires and Toxic Ash

Marc Los Huertos

2025-04-30

Survival Analysis

Introduction

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs. Because of censoring–the nonobservation of the event of interest after a period of follow-up–a proportion of the survival times of interest will often be unknown.

Survival analysis is widely used in various fields, including medicine, biology, and engineering. In this report, we will focus on survival analysis in the context of ecotoxicology, specifically examining the effects of pollutants on aquatic organisms.

Background

Survival analysis is a set of statistical methods used to analyze time-to-event data, where the event of interest is often death or failure. It is particularly useful in ecotoxicology, where researchers study the effects of pollutants on organisms over time.

Survival analysis allows researchers to estimate survival probabilities, compare survival curves between groups, and assess the impact of covariates on survival.

In this report, we will demonstrate survival analysis using a hypothetical dataset of Daphnia, a common model organism in ecotoxicology. We will use the Kaplan-Meier estimator to estimate survival curves and the Cox proportional hazards model to assess the impact of treatment conditions on survival.

Objectives

1. Introduce survival analysis and its relevance in ecotoxicology.
2. Demonstrate the use of the Kaplan-Meier estimator to estimate survival curves.
3. Perform a log-rank test to compare survival curves between treatment groups.
4. Fit a Cox proportional hazards model to assess the impact of treatment conditions on survival.
5. Discuss the implications of the results for ecotoxicological research and risk assessment.

Daphnia as a Test Species

Daphnia (commonly known as water fleas) are frequently used in ecological and toxicological studies due to their sensitivity to pollutants, short lifespans, and ease of culture.

In this report, we perform a survival analysis to assess the effects of different treatment conditions (e.g., pollutant doses) on the mortality of Daphnia over time. This approach helps quantify survival probabilities and assess the risk associated with various exposures.

Theoretical Background

Survival analysis is a statistical approach used to model time-to-event data. Here, the event of interest is death, and the main goals are:

The object of primary interest is the survival function, conventionally denoted S, which is defined as

[S(t)=Pr(T>t)]

where t is some time, T is a random variable denoting the time of death, and "Pr'' stands for probability. That is, the survival function is the probability that the time of death is later than some specified time t. The survival function is also called the survivor function or survivorship function in problems of biological survival, and the reliability function in mechanical survival problems. In the latter case, the reliability function is denoted R(t).

Usually one assumes S(0) = 1, although it could be less than 1 if there is the possibility of immediate death or failure.

To assess whether survival differs between groups (e.g., different treatments).

To quantify how treatment influences the hazard or risk of death.

Key concepts:

- Censoring: Some Daphnia may survive beyond the observation period; their survival time is considered right-censored.
- Kaplan-Meier Estimator: Non-parametric method for estimating survival functions.
- Log-Rank Test: Compares survival curves across groups.
- Cox Proportional Hazards Model: Evaluates the effect of covariates on the hazard function.

Load Required Packages

```
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

See Survival Analysis Packages for more information.

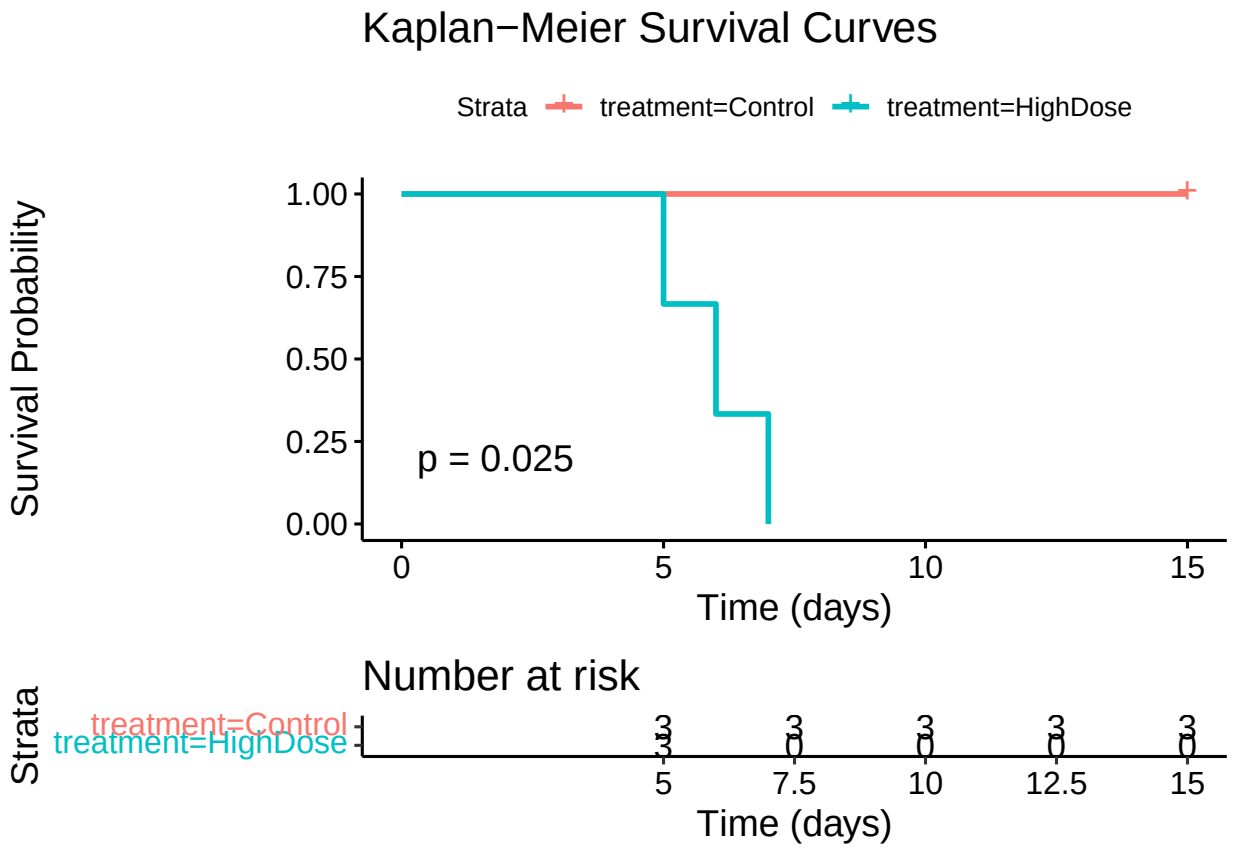
Load Data – Marc's Fake Data

```
daphnia_data <- data.frame(
  id = 1:6,
  time = c(15, 15, 5, 6, 15, 7),
  status = c(0, 0, 1, 1, 0, 1), #event/censor =0
  treatment = c("Control", "Control", "HighDose", "HighDose", "Control", "HighDose")
```

```
)
knitr::kable(daphnia_data)
```

id	time	status	treatment
1	15	0	Control
2	15	0	Control
3	5	1	HighDose
4	6	1	HighDose
5	15	0	Control
6	7	1	HighDose

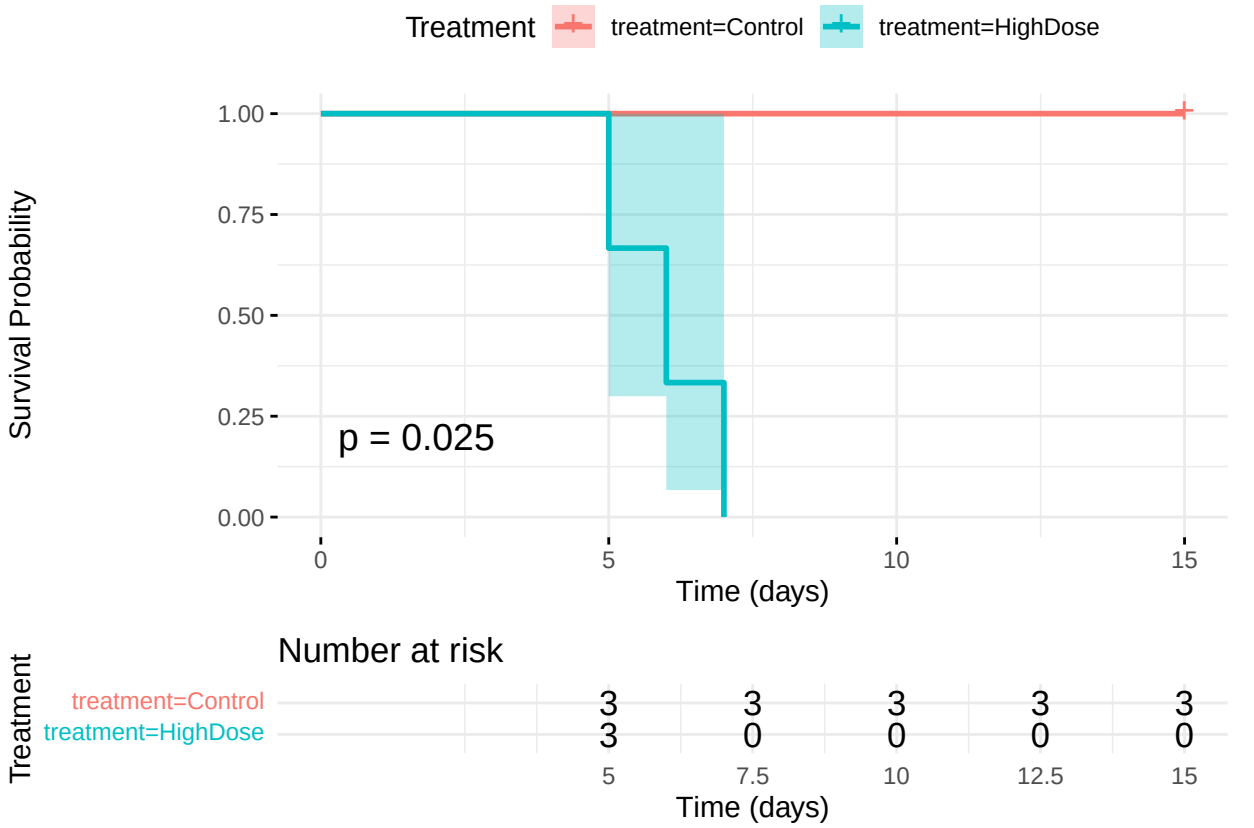
```
surv_obj <- Surv(time = daphnia_data$time, event = daphnia_data$status)
# Fit a Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ treatment, data = daphnia_data)
# Plot the survival curves
ggsurvplot(km_fit, data = daphnia_data,
           title = "Kaplan-Meier Survival Curves",
           xlab = "Time (days)",
           ylab = "Survival Probability",
           risk.table = TRUE,
           pval = TRUE)
```



Create a survival object

```
km_fit <- survfit(surv_obj ~ treatment, data = daphnia_data)
```

```
ggsurvplot(km_fit, data = daphnia_data,
           pval = TRUE,
           conf.int = TRUE,
           risk.table = TRUE,
           legend.title = "Treatment",
           xlab = "Time (days)",
           ylab = "Survival Probability",
           ggtheme = theme_minimal())
```



Log-Rank Test

We test whether there is a statistically significant difference in survival distributions between treatment groups.

```
log_rank_test <- survdiff(surv_obj ~ treatment, data = daphnia_data)
log_rank_test
```

```
## Call:
## survdiff(formula = surv_obj ~ treatment, data = daphnia_data)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=Control   3        0     1.85      1.85      5.05
## treatment=HighDose 3        3     1.15      2.98      5.05
##
##  Chisq= 5.1  on 1 degrees of freedom, p= 0.02
```

The log-rank test compares the survival curves of different groups. A significant p-value indicates that the

survival distributions differ between treatment groups.

Cox Proportional Hazards Model

This model estimates how treatment affects the hazard (death rate). A hazard ratio > 1 means increased risk.

```
survdiff(surv_obj ~ treatment, data = daphnia_data)
```

```
## Call:
## survdiff(formula = surv_obj ~ treatment, data = daphnia_data)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment=Control   3        0     1.85      1.85      5.05
## treatment=HighDose 3        3     1.15      2.98      5.05
##
##  Chisq= 5.1  on 1 degrees of freedom, p= 0.02
```

```
cox_model <- coxph(surv_obj ~ treatment, data = daphnia_data)
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
```

```
summary(cox_model)
```

```
## Call:
## coxph(formula = surv_obj ~ treatment, data = daphnia_data)
##
##   n= 6, number of events= 3
##
##                          coef exp(coef)  se(coef)     z Pr(>|z|)
## treatmentHighDose 2.194e+01 3.381e+09 2.480e+04 0.001    0.999
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## treatmentHighDose 3.381e+09  2.957e-10         0       Inf
##
## Concordance= 0.875  (se = 0.077 )
## Likelihood ratio test= 5.99  on 1 df,   p=0.01
## Wald test            = 0  on 1 df,   p=1
## Score (logrank) test = 5.05  on 1 df,   p=0.02
```

Coefficients: Log hazard ratios

exp(coef): Hazard ratios (HR)

p-values: Significance of the treatment effect

Conclusion

Using survival analysis, we assessed the impact of treatments on Daphnia mortality:

Kaplan-Meier curves suggest differences in survival between treatment groups.

The log-rank test evaluates the statistical significance of these differences.

The Cox model quantifies the relative risk associated with each treatment.

This framework is suitable for analyzing ecotoxicological data and can be expanded to include more covariates such as temperature, age, or replicate ID. In real datasets, always check proportional hazards assumptions before interpreting the Cox model.

Fake Data

```
group1.csv = "/home/mwl04747/RTricks/00_Project_Group_Demos/Group1_FakeData.csv"
group1 = read.csv(group1.csv)
head(group1)
```

```
##           Sample.location Daphnia....dead.30.total. Concentration.of.PFAS
## 1   Santa Monica Upstream                         17              1490 ppt
## 2 Santa Monica Mid-stream                         20              1640 ppt
## 3 Santa Monica Downstream                         23              2600 ppt
## 4            Eaton stream                         15              1784 ppt
## 5    San Antonio Upstream                         10               790 ppt
## 6  San Antonio Downstream                         12               640 ppt
##   Heavy.metal.concentration
## 1                     1 ppb
## 2                    10 ppb
## 3                   10 ppm
## 4                    10 ppb
## 5                   > 1 ppt
## 6                   > 1 ppt
```

```
names(group1) = c("Location", "Survival", "PFAS", "Metals")
```

```
toxity.lm = lm(Survival ~ PFAS, data = group1)
summary(toxity.lm)
```

```
##
## Call:
## lm(formula = Survival ~ PFAS, data = group1)
##
## Residuals:
## ALL 6 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17        NaN     NaN      NaN
## PFAS1640 ppt        3        NaN     NaN      NaN
## PFAS1784 ppt       -2        NaN     NaN      NaN
## PFAS2600 ppt        6        NaN     NaN      NaN
## PFAS640 ppt        -5        NaN     NaN      NaN
## PFAS790 ppt        -7        NaN     NaN      NaN
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 5 and 0 DF,  p-value: NA
```

Hypotheses

Data Set

Summary Stats

Hypothesis Tests

Plots