Introduction to Survival Analysis

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- **Survival analysis** is employed when the variable of interest is the *time until an event occurs*.
- Time can be measured in many units including years, months, days, hours, etc.
- The event can be death, disease incidence, recovery, or any interesting event that can be considered in this context.

Definitions and Notation

Definitions

- **Survival time** is the time variable as it gives the time that an individual has *survived* over some follow-up period.
- The *event* is often referred to as **failure**.
- **Censoring** occurs when the value of measurement or the variable is only partially known.

Right Censoring

- **Right censoring** occurs when the object has not reached the event (survival time is partially realized).

- This can occur in a few ways:
  - The study ends before the object/person reaches the event.
  - The object/person is removed from or leaves the study.
  - The data are poorly collected or an object is missed.

Right Censoring Illustration

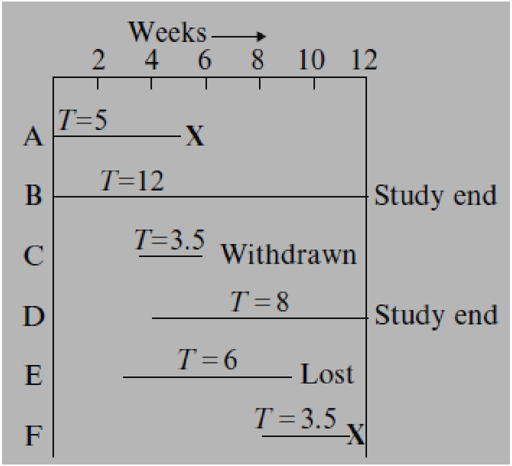

Figure: Source: (1)

Notations

- T random variable used for survival time.
- t specific value of interest related to T.
- d indicator variable ($d = 1$ for failure; $d = 0$ no failure/censoring)
- $S(t)$ the probability that a person survives longer than t (**survivor function**).
- $h(t)$ the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t (**hazard function**).

Modelling Approaches

- Kaplan-Meier survival curves.
- **Cox proportional hazards model.**
- Parametric survival models.
- Recurrent event survival analysis
- Note: *This list is not exhaustive and Survival Analysis is often offered as an entire university course.*

Example 1

1. Import the *diabetic* dataset from the *survival* package.
2. Take some time to understand the data.
3. Run the provided code to examine the Kaplan-Meier plot.

Cox Proportional Hazards Model

Cox Proportional Hazards Model

- The **Cox proportional hazards model** models probability of the event of dying/failing in the interval $(t,\ t+dt]$ conditioned on survival until t.
- The model allows for the inclusion of explanatory variables.
- It is a semi-parametric model.

Cox Proportional Hazards Model Notation

$$h(t) = h_0(t) \cdot exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

- $h(t)$ represents the hazard function.
- $h_0(t)$ represents the baseline hazard if all of the x_i's are equal to 0.
- t represents the survival time.
- $x_1, x_2, ..., x_p$ represent the set of explanatory variables.
- $\beta_1, \beta_2, ..., \beta_p$ represents the measure the effect sizes of the explanatory variables.

- Matrix notation:

$$h(t) = h_0(t) e^{\mathbb{X}^{\mathbb{T}} \beta}$$

Cox Proportional Hazards Model in R

- Using the *survival* package:

```
model <- coxph(Surv(time, status) ~ variable.1 + ... +
variable.p, cluster = id, data=data)
```

- The `summary()` function gives coefficient estimates and the significance of the overall model.

Example 2

1. Using the diabetic dataset, estimate Cox PH models with the following sets of explanatory variables:
   - Model 1: *age* and *trt*
   - Model 2: *laser*, *age*, *eye*, *trt*, and *risk*.
   - Model 3: *age*, *eye*, *trt*, *risk*, and *age:trt*.
2. Are the models significant?
3. What do the coefficient estimates imply?

Interpreting the Results

- The null hypothesis of the Likelihood ratio test, the Wald test, and the Score (logrank) test is that the model is not significant.
- Coefficient Interpretation:
  - $\beta_i = 0$ indicates no effect.
  - $\beta_i < 0$ indicates a reduction in the hazard.
  - $\beta_i > 0$ indicates an increase in hazard.

Example 3

1. Run the example code to check the differences in the survival curve of the treatment (trt) classes.
2. What did you notice?

Assumptions and Diagnostics

Assumptions

1. Independent observations.
2. Non-informative or independent censoring.
   - The censoring is not related to the event (independent).
3. **Proportional hazard ratios:** The Cox PH model assumes that the hazard ratio comparing any two specifications of explanatory variables is constant over time.

Testing Proportional Hazards Ratios

- We can use the `cox.zph(model)` function from the *survival* package to test this assumption.
  - The null hypothesis is that the hazard ratios remain proportional in time (satisfies the assumption).
  - Note: *This function tests the individual variables and the model as a whole.*
- We can visualize the proportions using the `ggcoxzph()` function.
  - If the assumption is satisfied, the plots will be randomly scattered around 0 (no patterns).
- *Other methods exist to test this assumption.*

Example 4

1. Estimate a Cox PH model using the significant variables from Example 2.

2. Does your model pass the proportional hazard ratios assumption?

Comments on Predictions

- `predict(model, new.data, type = "risk")` will predict the hazards ratio for *failure*.
- Lower values indicate a lower chance of failure (longer predicted survival time).

Exercise 1

- Import the *cancer* dataset from the survival package.
- Practice estimating some Cox proportional hazards models from this data.
- Does your best model pass the proportional hazards ratios assumption?

References & Resources

1. Kleinbaum, D. G., & Klein, M. (2013). *Survival analysis a self-learning text.* Springer.

- survfit2()
- survival
- cox.zph()