Survival Analysis: Kaplan–Meier and Cox Proportional Hazards Models

Gul Bulbul

Contents

1 Overview

Survival analysis focuses on modeling the time until an event of interest occurs, such as death, disease relapse, or disease progression. A key challenge in survival data is the presence of censoring, where the event has not yet occurred for some individuals at the end of follow-up.

This document presents two fundamental tools in survival analysis: the Kaplan–Meier estimator and the Cox proportional hazards model.

2 Kaplan–Meier Estimator

Definition

The Kaplan–Meier estimator is a non-parametric method used to estimate the survival function:

$$S(t) = P(T > t),$$

which represents the probability that an individual remains event-free beyond time t.

Real-World Example Consider a clinical study evaluating a new cancer treatment. Let the random variable T denote the time (in months) from the start of treatment until disease relapse or death. Some patients may leave the study early or may not experience the event by the end of follow-up, resulting in right-censored observations.

The Kaplan–Meier estimator is used to estimate the survival function

$$S(t) = P(T > t),$$

which represents the probability that a patient remains relapse-free beyond time t. For example, if the Kaplan–Meier estimate at $t = 12$ months is $S(12) = 0.70$, this means that approximately 70% of patients are expected to remain event-free at least 12 months after treatment initiation.

The stepwise decreases in the Kaplan–Meier curve occur at observed event times, while censored observations reduce the number of individuals at risk without causing a drop in the survival probability. This makes the Kaplan–Meier method particularly suitable for analyzing time-to-event data with incomplete follow-up, which is common in real-world clinical studies.

Estimator

The Kaplan–Meier estimator is given by:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where t_i denotes event times, d_i is the number of events at time t_i, and n_i is the number of individuals at risk just prior to t_i.

Real-World Example Consider a clinical study following patients after cancer surgery, where the event of interest is disease relapse. Let T denote the time (in months) from surgery to relapse. At each observed event time t_i, suppose d_i patients experience relapse and n_i patients are still at risk just prior to t_i.

The Kaplan–Meier estimator combines these observed proportions to estimate the survival function

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right),$$

which represents the estimated probability that a patient remains relapse-free beyond time t.

For example, if at month 6 there are $n_1 = 20$ patients at risk and $d_1 = 2$ relapses occur, the survival probability is multiplied by $1 - 2/20 = 0.9$. If another $d_2 = 3$ relapses occur at month 12 among $n_2 = 15$ patients at risk, the estimate is further multiplied by $1 - 3/15 = 0.8$. The Kaplan–Meier estimator therefore reflects the cumulative effect of observed events over time.

Patients who do not experience relapse during follow-up are treated as right-censored; they contribute to the risk set n_i up to their censoring time but do not directly reduce the survival probability. This makes the Kaplan–Meier estimator well suited for real-world clinical studies with incomplete follow-up.

Key Properties

- Accounts for right-censored observations
- Produces a stepwise survival curve
- Does not adjust for covariates

Interpretation

Kaplan–Meier curves are primarily used for descriptive analysis and visualization of survival probabilities over time. Group comparisons are commonly performed using the log-rank test.

3 Cox Proportional Hazards Model

Model Specification

The Cox proportional hazards model relates covariates to the hazard function:

$$h(t \mid X) = h_0(t) \exp(\beta^\top X),$$

where $h(t \mid X)$ is the hazard at time t, $h_0(t)$ is the baseline hazard, and β is a vector of regression coefficients.

Hazard Ratio

The exponentiated coefficients,

$$\text{HR} = \exp(\beta^\top X),$$

are interpreted as hazard ratios:

- HR = 1: no effect
- HR < 1: reduced risk
- HR > 1: increased risk

Proportional Hazards Assumption

The model assumes that hazard ratios are constant over time. This assumption is commonly assessed using Schoenfeld residuals and graphical diagnostics.

Interpretation

The Cox model enables inference on the effect of covariates while properly accounting for censoring, making it the most widely used regression model in clinical survival analysis. Censoring occurs when the event of interest has not been observed for an individual by the end of the follow-up period. Right censoring is the most common form in clinical studies.

4 Kaplan–Meier vs Cox Model

- Kaplan–Meier provides a non-parametric estimate of survival probabilities.
- Cox regression quantifies covariate effects through hazard ratios.
- Kaplan–Meier is descriptive; Cox is inferential.

5 Time-to-Event Endpoints: RFS and PFS

In oncology and clinical research, survival analysis is often defined with respect to specific time-to-event endpoints. Two commonly used endpoints are relapse-free survival (RFS) and progression-free survival (PFS). Both endpoints are analyzed using Kaplan–Meier curves and regression models such as Cox proportional hazards.

5.1 Relapse-Free Survival (RFS)

Definition. Relapse-free survival is the time from a defined baseline (e.g., surgery, achieving remission, or start of follow-up after curative-intent therapy) to the first occurrence of **relapse/recurrence** or **death**.

Event indicator. Let T_i denote event time and δ_i the event indicator. For RFS,

$$\delta_i = \begin{cases} 1, & \text{if relapse/recurrence occurs or the patient dies (whichever occurs first)} \\ 0, & \text{if censored (no relapse and alive at last follow-up).} \end{cases}$$

Censoring. Patients without relapse and alive at last contact are right-censored at their last follow-up time.

Typical use. RFS is commonly used in adjuvant settings and curative-intent treatments where patients may become disease-free, and the main question is whether treatment delays or prevents recurrence.

5.2 Progression-Free Survival (PFS)

Definition. Progression-free survival is the time from baseline (often treatment initiation) to the first occurrence of **disease progression** or **death**.

Event indicator.

$$\delta_i = \begin{cases} 1, & \text{if progression occurs or the patient dies (whichever occurs first)} \\ 0, & \text{if censored (no progression and alive at last follow-up).} \end{cases}$$

Censoring. Patients who are alive without documented progression at last follow-up are right-censored at their last assessment.

Typical use. PFS is widely used in advanced or metastatic disease settings where patients have measurable disease and the primary goal is to delay progression.

5.3 Key difference between RFS and PFS

- **Relapse** typically refers to disease returning after remission or disease-free status.
- **Progression** refers to worsening of existing disease (e.g., tumor growth or new lesions).

6 Cox Proportional Hazards Model for RFS/PFS

6.1 Model specification

The Cox proportional hazards model relates covariates to the hazard of experiencing the event defining the endpoint (RFS or PFS):

$$h(t \mid X) = h_0(t) \exp(X\beta), \tag{1}$$

where $h(t \mid X)$ is the instantaneous event risk at time t given covariates X, $h_0(t)$ is the unspecified baseline hazard, and β is a vector of regression coefficients.

6.2 Hazard ratio interpretation

For a one-unit increase in a covariate x_k (or comparing two groups), the hazard ratio is:

$$\text{HR} = \exp(\beta_k).$$

- HR < 1 indicates **lower** event risk (longer event-free time).
- HR > 1 indicates **higher** event risk (shorter event-free time).

Clinical interpretation. For RFS/PFS, the "event" is defined as relapse/progression or death. Thus, a treatment HR of 0.70 can be interpreted as approximately a 30% lower instantaneous risk of experiencing relapse/progression (or death) at any given time, assuming proportional hazards.

6.3 Proportional hazards (PH) assumption

The Cox model assumes that hazard ratios are constant over time:

$$\frac{h(t \mid X_1)}{h(t \mid X_2)} \text{ is constant in } t.$$

This assumption can be assessed using Schoenfeld residuals and graphical diagnostics.

6.4 Practical usage in clinical studies

- **Kaplan–Meier** curves provide descriptive comparisons of RFS/PFS across groups.
- **Cox regression** provides covariate-adjusted inference via hazard ratios.
- Both methods naturally handle **right censoring**.

6.5 Interpretation and Scope

The Cox model estimates *associations* between covariates and the hazard of the event. Causal interpretation requires additional assumptions or study designs, such as randomization or explicit causal modeling.

Survival analysis methods are widely used in clinical research to evaluate endpoints such as overall survival, relapse-free survival, and progression-free survival.