An Introduction to Survival Analysis

Micky Liu and Yanqi Gao

Mentor: Xianwei Li

April 10, 2024

Table of contents

Basic Concepts

Survival Analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs*.

$$\boxed{\text{start follow-up}} \xRightarrow{\text{time}} \boxed{\text{event}}$$

- We often refer to the time variable as **survival time** and the event as **failure**
- Examples:
  - Heart transplant/time until death (months)
  - Parolees/time until rearrest (weeks)

Goals of Survival Analysis

1. Estimate and interpret survivor and/or hazard functions from survival data
2. Compare survival and/or hazard functions across groups
3. Assess the relationship of explanatory variables to survival time

Survival Function and Hazard Function

The **survival function** gives the probability that a person survives longer than some specified time t.

$$S(t) = \mathbb{P}\,(T > t)$$

The **hazard function** gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\,(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Properties

Properties of $S(t)$

- $S(t) \in [0, 1]$
- $S(t) = \exp\left[-\int_0^t h(u)du\right]$

Properties of $h(t)$

- $h(t) \in [0, \infty)$
- $h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$

Figure 1: Relationship between $S(t)$ and $h(t)$

Survival Curves

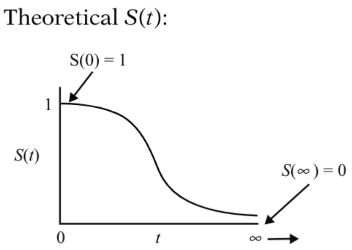

Figure 2: Theoretical survival curve

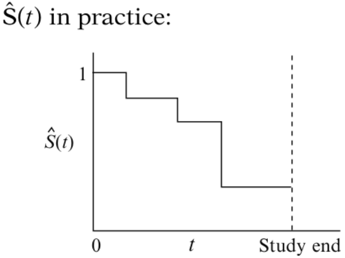

Figure 3: Survival curve in practice

Hazard Curves

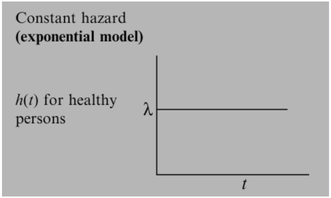

Figure 4: Hazard function ($T \sim Exponential$)

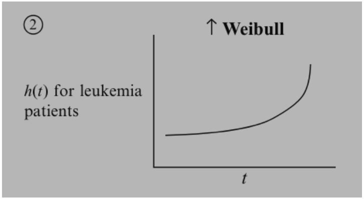

Figure 5: Hazard function ($T \sim Weibull$)

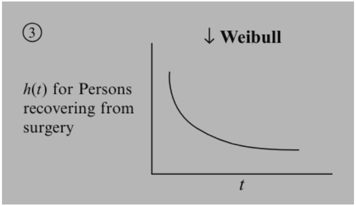

Figure 6: Hazard function ($T \sim Weibull$)

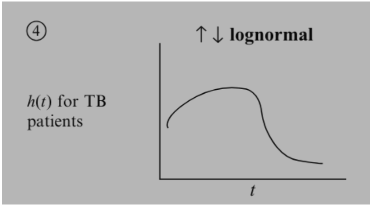

Figure 7: Hazard function ($T \sim Lognormal$)

Problem of Data Set in Practice

- What if the data set is not *complete*?
- What if the data set is not *100% accurate*?

Censored Data

For an individual, we say the survival time of such individual is **censored** when we don't know the exact survival time. Some reasons of data being censored are

- Person withdraws from the study
- Person does not experience a failure before the study ends.

Censored data can be classified into three groups

- Right-censored
- Left-censored
- Interval-censored

Censoring Assumptions

Three assumptions about censoring are

- **Random Censoring** - censored group at time t has the identical feature of remain at risk group at time t
- **Independent Censoring** - censored subgroup at time t has the identical feature of remain at risk group at time t
- **Non-informative Censoring** - the censoring time gives no information about the actual survival time

EM Algorithm

A more theoretical approach to fit censored data in a model, where we have two steps

- Expectation step

$$Q(\theta; \widehat{\theta}_{t-1}) := \mathbb{E}\left\{\ell(\theta; C, NC)\,\middle|\,NC, \widehat{\theta}_{t-1}\right\}$$

- Maximization step

$$\widehat{\theta}_t = \arg\max_\theta Q(\theta; \widehat{\theta}_{t-1})$$

- Main idea here: Given a good first guess, and use EM iteration to keep on updating the guess. Conclude the the answer when the sequence of estimators generated from EM converges.

Kaplan-Meier (KM) Survival Curves

A great technique to generate survival curve. General idea is

$$S(t_n) = S(t_{n-1}) \cdot \mathbb{P}\left(T > t_n \,\middle|\, T > t_{n-1}\right)$$

And we also have an example

Remission time for a group of 13 leukemia patients:
1, 1, 1, 2, 3, 4, 5, 2+, 3+, 3+, 4+, 4+, 5+

where $t+$ represents censored at time t

$t_{(f)}$	n_f	m_f	q_f	$\widehat{S}(t_{(f)})$
0	13	0	0	1
1	13	3	0	$10/13 = 0.7692$
2	10	1	1	$(0.7692)(9/10) = 0.6154$
3	8	1	2	$(0.6154)(7/8) = 0.5385$
4	5	1	2	$(0.5385)(4/5) = 0.4308$
5	2	1	1	$(0.4308)(1/2) = 0.2154$

KM Survival Curves Continued

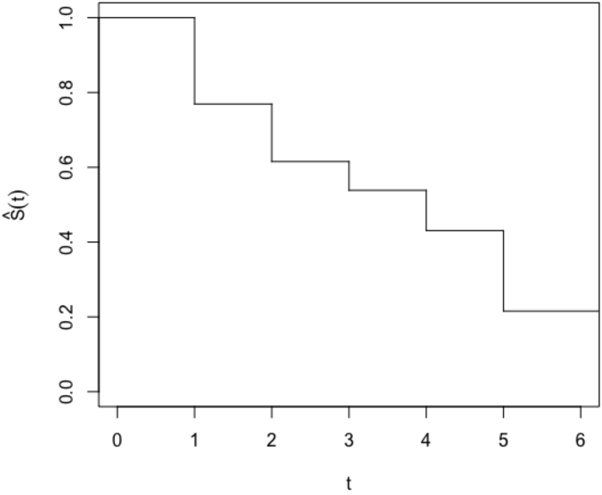

Figure 8: The Kaplan-Meier Survival curve based on the data example

Cox Proportional Hazard (PH) Model

Formula for the Cox PH model:

$$h(t, \mathbf{X}) = h_0(t) \exp\left\{\sum_{i=1}^{p} \beta_i X_i\right\}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

- $h_0(t)$: baseline hazard (unspecified)
- Measure of association: hazard ratio
- Primary quantities:
  - Estimated hazard ratios
  - Estimated survival curves

Hazard Ratio

The **hazard ratio** is the hazard for one individual divided by the hazard for a different individual.

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \exp\left\{\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right\}$$

- **Proportional hazard assumption**: Hazard ratio is constant over time

Checking PH Assumption

Approaches

- Graphical
  - Log-log survivor curves $-\ln(-\ln \hat{S})$
  - Observed vs expected survivor curves
- Goodness-of-fit test
- Time-dependent variables

Remark

- $\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp(\sum_{i=1}^{p} \hat{\beta}_i X_i)}$
- $\ln[-\ln \hat{S}(t, \mathbf{X})] - (\ln[-\ln \hat{S}(t, \mathbf{X}^*)]) = \sum_{i=1}^{p} \hat{\beta}_i (X_i - X_i^*)$

Log-log Survival Curves

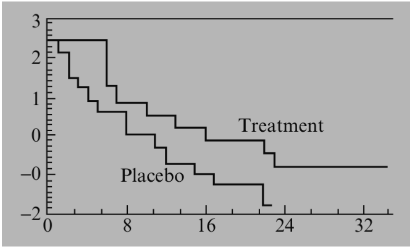

Figure 9: PH assumption met

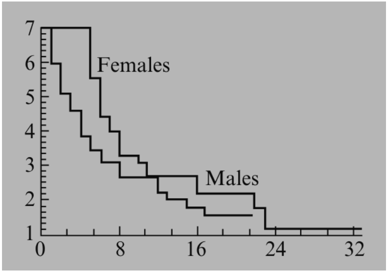

Figure 10: PH assumption violated

Paper

- Evaluation of Time-Varying Biomarkers in Mortality Outcome in COVID-19: an Application of Extended Cox Regression Model

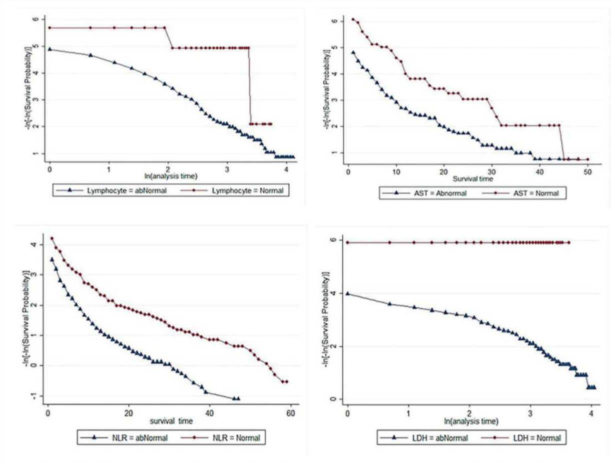

Figure 11: Log-log plots for biomarkers

References

- Kleinbaum, D. G.(2011) *Survival Analysis: A Self-Learning Text*, Third Edition (3rd ed.). Springer Nature

- Geraili, Z., Hajian-Tilaki, K., Bayani, M., Hosseini, S. R., Khafri, S., Ebrahimpour, S., Javanian, M., Babazadeh, A., Shokri, M. (2022). *Evaluation of Time-Varying Biomarkers in Mortality Outcome in COVID-19: an Application of Extended Cox Regression Model*. Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia Herzegovina : casopis Drustva za medicinsku informatiku BiH, 30(4), 295–301. `https://doi.org/10.5455/aim.2022.30.295-301`

- Haugh, M (2015), *The EM Algorithm*, `https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf`