6.9: Survival Analysis and Censored Data

Navona Calarco

The University of Toronto

Introduction

Survival analysis and censored data are related to the analysis of the outcome variable best described as "the time until and event occurs". Some examples include:

- Predicting survival time of patients diagnosed with cancer.
- Predicting the time a customer will cancel a subscription.
- Predicting the customer wait time for a call centre.

Survival and Censoring Times

For each individual we assume there exists:

- T: the true **survival time** which represents the time at which the event of interest occurs. *Ex: The time at which the patient dies.*
- C: the true **censoring time** which is when the censoring occurs. *Ex: The time at which the patient drops out of the study.*

However, we only ever observe the time that comes first:

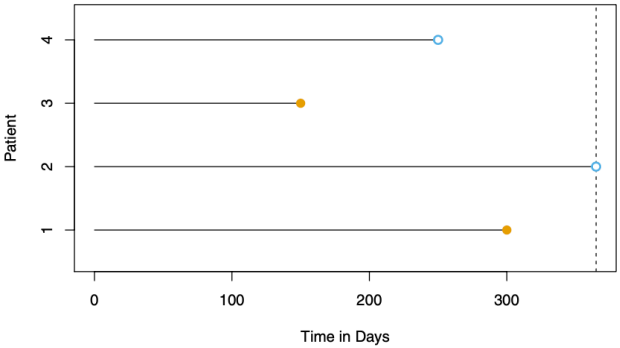
$$Y = \min(T, C)$$

In order to keep track of which time Y represents, we also have a status indicator:

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$$

Survival and Censoring Times

Suppose we have the following data from a medical study with $n = 4$ patients. We observe 4 pairs of (Y, δ), denoted as $(y_1, \delta_1), \ldots, (y_4, \delta_4)$.



- For patient 1 and 3, we observe the time to the event (i.e. death) during the study.
  - Thus, $\delta_1 = \delta_3 = 1$
- Patient 2 survives until the end of the study.
  - Thus, $\delta_2 = 0$
- Patient 4 drops out of the study early.
  - Thus, $\delta_4 = 0$

Independent Censoring

We assume that the censoring mechanism is independent. That is, the event time T is independent of the censoring time C.

Example of a violation:

- Some patients drop out of a cancer study because they are very sick.
- If we analyse the data without considering why patients drop out then we will overestimate the true average survival time.

The data collection process must be examined closely to check whether independent censoring is a reasonable assumption.

The Kaplan-Meier Survival Curve

The Kaplan-Meier survial curve (or survival function) is a decreasing function that quantifies the probability of surviving past time t. It is defined by

$$S(t) = \Pr(T > t)$$

Estimating $S(t)$ is complicated by the presence of censoring but we have an approach to overcome this challenge.

- $d_1 < d_2 < \cdots < d_K$ denote the K unique survival times among the non-censored individuals.
- q_k denotes the number of events that took place at time d_k. (i.e. the number of patients that died at time d_k)
- r_k denotes the number of individuals alive and in the study just before d_k (the **at risk** patients).
- The set of patients that are at risk are the **risk set**.

The Kaplan-Meier Survival Curve

So, the probability of surviving past time d_k is given by:

$$S(d_k) = \Pr(T > d_k) = \Pr(T > d_k \mid T > d_{k-1}) \Pr(T > d_{k-1}).$$

Note that $\Pr(T > d_{k-1})$ amounts to $S(d_{k-1})$. So,

$$S(d_k) = \Pr(T > d_k \mid T > d_{k-1}) S(d_{k-1}).$$

Thus,

$$S(d_k) = \Pr(T > d_k \mid T > d_{k-1}) \times \cdots \times \Pr(T > d_2 \mid T > d_1) \Pr(T > d_1)$$

Now we need the estimates for each of these terms.

The Kaplan-Meier Survival Curve

The estimator

$$\widehat{\Pr}\left(T > d_j \mid T > d_{j-1}\right) = \left(r_j - q_j\right)/r_j$$

is the fraction of the risk set at time d_j who survived past time d_j. So, the **Kaplan-Meier estimator** of the survival curve is

$$\widehat{S}\left(d_k\right) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j}\right)$$

- For times t between d_k and d_{k+1} we set $\widehat{S}(t) = \widehat{S}\left(d_k\right)$.
- This gives the Kaplan-Meier survival curve a step-like shape.

Exercise: The Kaplan-Meier Survival Curve

Open the Survival Analysis and Censored Data R Markdown or Jupyter Notebook file.

- Go over the "The Kaplan-Meier Survival Curve" section together as a class.

The Log-Rank Test

The log-rank test aims to test the null hypothesis that there is no difference between two survival curves. The statistic is computed with

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

where $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ are the expectation and variance under the assumption of the null hypothesis.

The Log-Rank Test}

Suppose we want to compare the survival of two groups of people by comparing the two survival curves.

Recall:

- $d_1 < d_2 < \cdots < d_K$ are the unique survival times among the non-censored individuals.
- r_k is the number of individuals at risk at time d_k.
- q_k denotes the number of patients that died at time d_k

Now we define:

- r_{1k}, r_{2k} are the number of individuals at risk at time d_k in group 1 and 2 respectively.
- q_{1k}, q_{2k} are the number of patients that died at time d_k in group 1 and 2 respectively.
- Note that $r_{1k} + r_{2k} = r_k$ and $q_{1k} + q_{2k} = q_k$.

The Log-Rank Test

At each death time d_k, we have the following table:

	Group 1	Group 2	Total
Died	q_{1k}	q_{2k}	q_k
Survived	$r_{1k} - q_{1k}$	$r_{2k} - q_{2k}$	$r_k - q_k$
Total	r_{1k}	r_{2k}	r_k

- $\frac{r_{1k}}{r_k}$ is the proportion of at risk individuals that are in group 1.
- If there is no difference in the survival rate between the two groups, we would expect $\frac{r_{1k}}{r_k} q_k$ individuals in group 1 to die at time d_k.
- So, $\mathrm{E}\left(q\_1k\right) = \frac{r_{1k}}{r_k} q\_k$ is the expected number of deaths at time d_k in group 1.

The Log-Rank Test

In our case, the log rank test statistic is computed with

$$W = \frac{\sum_{k=1}^{K} \left(q_{1k} - \mathrm{E}\left(q_{1k}\right)\right)}{\sqrt{\sum_{k=1}^{K} \mathrm{Var}\left(q_{1k}\right)}}$$

where

$$\mathrm{Var}\left(q_{1k}\right) = \frac{q_k \left(r_{1k}/r_k\right)\left(1 - r_{1k}/r_k\right)\left(r_k - q_k\right)}{r_k - 1}.$$

- When the sample size is large, W has approximately a standard normal distribution.
- Then, a p-value can be used to test the null hypothesis that there is no difference between the survival curves of the two groups.

Exercise: The Log-Rank Test

Open the Survival Analysis and Censored Data R Markdown or Jupyter Notebook file.

- Go over the "The Log-Rank Test" section together as a class.

Regression with a Survival Response

Suppose we would like to fit a regression model to survival data with the following properties.

- n observations (Y, δ)
  - $Y = \min(T, C)$.
  - δ equals 1 if $Y = T$ and 0 otherwise.
- $X \in \mathbb{R}^p$ is a vector of p features.
- We want to predict the true survival time T.

Note that we want to predict T, not Y. Censoring makes this difficult so we make use of the **hazard function**.

The Hazard Function

The hazard function $h(t)$ is the death rate in the instant after time t, given survival past that time. That is,

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}$$

where T is the unobserved survival time. We take the limit as Δt approached zero so we can think of Δt as an extremely small number.

Now we want to use the hazard function to model the survival time as a function of the covariates x_{ij}.

The Proportional Hazards Assumption

The **proportional hazards assumption** states

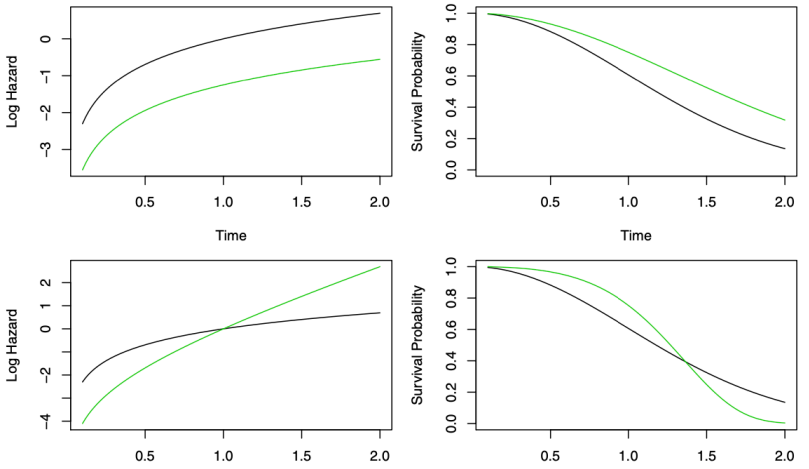
$$h\left(t \mid x_i\right) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$

- $h_0(t) \geq 0$ is an unspecified function known as the **baseline hazard** which is the hazard function for an individual with features $x_{i1} = \cdots = x_{ip} = 0$.
- $\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$ is called the **relative risk** for the feature vector $x_i = \left(x_{i1}, \ldots, x_{ip}\right)$ relative to $x_i = (0, \ldots, 0)$.

-$\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$ are the parameters that we need to estimate using the likelihood (we will not present the likelihood function).

The Proportional Hazards Assumption

Since the baseline hazard function $h_0(t)$ is unspecified the only assumption we are making is that a one-unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$.



- Two models with a binary covariate x_i.
- Left plots show the log hazards and right plots show the survival functions.
- Green is for $x_i = 0$ and black is for $x_i = 1$.
- Top model satisfies the assumption.
- Bottom model doesn't satisfy the assumption.

Cox Proportional Hazards Model

Because $h_0(t)$ in the proportional hazards assumption is unknown we cannot plug $h(t|x_i)$ into the likelihood to get estimates for $\beta = (\beta_1, \ldots, \beta_p)$.

Cox's proportional hazards model says that it it possible to estimate β without specifying the form of $h_0(t)$.

- Use the **partial likelihood** which is valid regardless of the the value of $h_0(t)$.
- Maximize the partial likelihood with respect to β.
- We can obtain p-values corresponding to null hypotesis such as $H_0 : \beta_j = 0$.
- We can obtain confidence intervals associated with the estimated coefficients.

Cox Proportional Hazards Model

Suppose we have a single binary predictor $x_i \in \{0, 1\}$ and we want to determine whether there is a difference between the survival times of the observations in the two groups.

- *Approach #1:* Fit a Cox proportional hazards model and test the null hypothesis $H_0 : \beta = 0$.
- *Approach #2:* Perform a log-rank test to compare the two groups.

In the case of a binary predictor both methods will yield the same result!

Exercise: The Cox Proportional Hazards Model

Open the Survival Analysis and Censored Data R Markdown or Jupyter Notebook file.

- Go over the "The Cox Proportional Hazards Model" section together as a class.

References

Chapter 11 of the ISLR2 and ISLP books:

James, Gareth, et al. "Survival Analysis and Censored Data." An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.

James, Gareth, et al. "Survival Analysis and Censored Data." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.