Survival Analysis

Tianhang Zheng
https://tianzheng4.github.io

Survival Analysis

Survival analysis concerns a special kind of outcome variable: the time until an event occurs.

For example, suppose that we have conducted a five-year medical study, in which patients have been treated for cancer.

We would like to fit a model to predict patient survival time, using features such as baseline health measurements or type of treatment.

An Example

The applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model churn, the event when customers cancel subscription to a service.

The company might collect data on customers over some time period, in order to predict each customer's time to cancellation.

However, presumably not all customers will have cancelled their subscription by the end of this time period; for such customers, the time to cancellation is censored.

Survival and Censoring Times

For each individual, we suppose that there is a true failure or event time T, as well as a true censoring time C.

The survival time represents the time at which the event of interest occurs (such as death).

By contrast, the censoring is the time at which censoring occurs: for example, the time at which the patient drops out of the study or the study ends.

Survival and Censoring Times

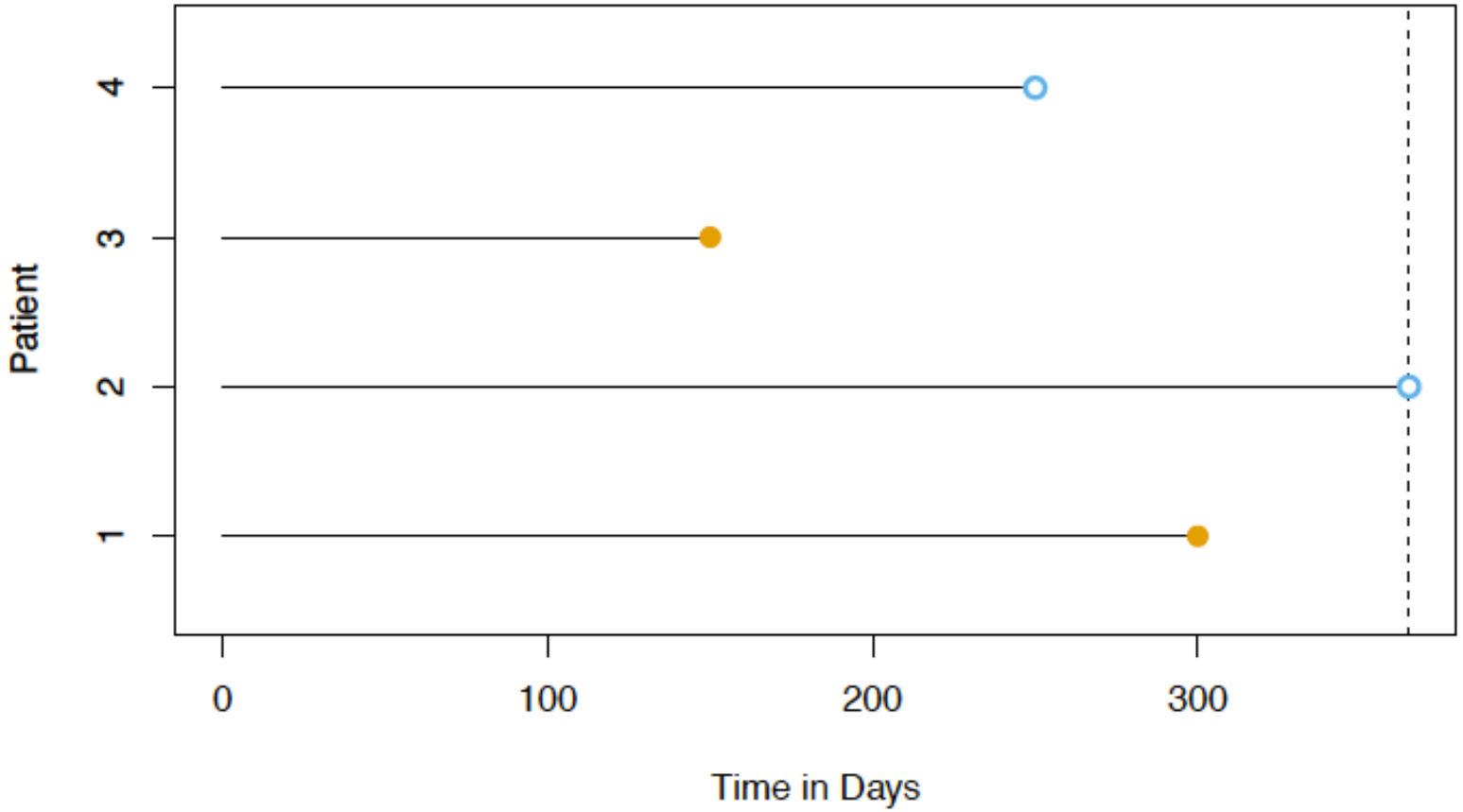
We observe either the survival time T or else the censoring time C. Specifically, we observe the random variable

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

If the event occurs before censoring (i.e. T < C) then we observe the true survival time T; if censoring occurs before the event (T > C) then we observe the censoring time.

An Illustration

For patients 1 and 3, the event was observed. Patient 2 was alive when the study ended. Patient 4 dropped out of the study.



Censoring

Suppose that a number of patients drop out of a cancer study early because they are very sick.

An analysis that does not take into consideration the reason why the patients dropped out will likely overestimate the true average survival time.

Censoring

Similarly, suppose that males who are very sick are more likely to drop out of the study than females who are very sick.

Then a comparison of male and female survival times may wrongly suggest that males survive longer than females.

The Survival Curve

The survival function (or curve) is defined as

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

This decreasing function quantifies the probability of surviving past time t.

The Survival Curve

Consider the **BrainCancer** dataset, which contains the survival times for patients with primary brain tumors undergoing treatment with stereotactic radiation methods.

The predictors are gtv (gross tumor volume, in cubic centimeters); sex (male or female); diagnosis (meningioma, LG glioma, HG glioma, or other); loc (the tumor location: either infratentorial or supratentorial); ki (Karnofsky index); and stereo (stereotactic method).

The Survival Curve

Only 53 of the 88 patients were still alive at the end of the study.

Suppose we'd like to estimate $S(20) = Pr(T > 20)$, the probability that a patient survives for at least 20 months

We can simply compute the proportion of patients who are known to have survived past 20 months.

The Survival Curve

But is it a right estimation?

17 of the 40 patients who did not survive to 20 months were actually censored

We cannot simply assume that they died, which may lead to an underestimation

The Kaplan-Meier Estimate

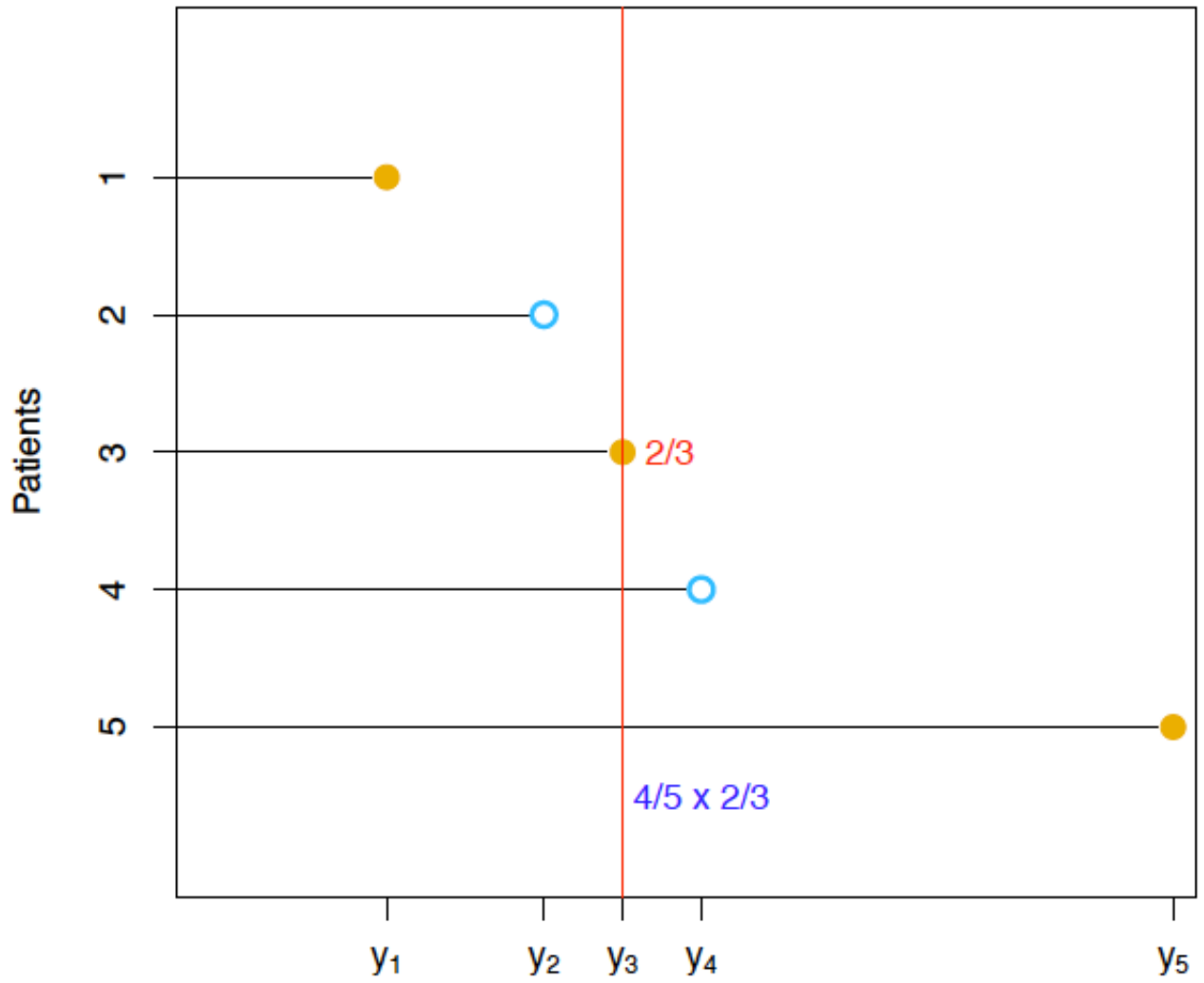
But is it a right estimation?

17 of the 40 patients who did not survive to 20 months were actually censored

We cannot simply assume that they died, which may lead to an underestimation

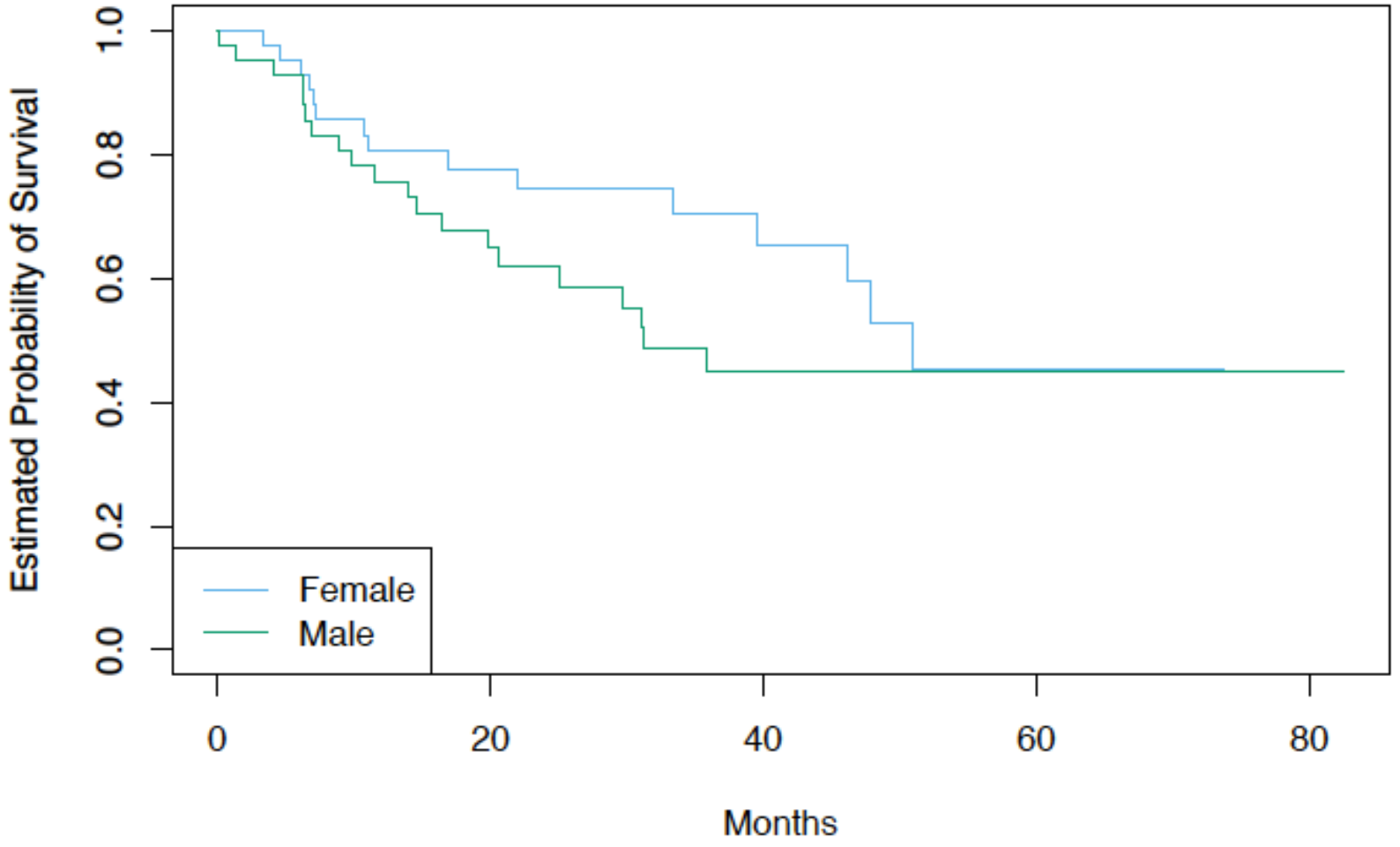
The Kaplan-Meier Estimate

$$\widehat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$



Log-Rank Test

We wish to compare the survival of males to that of females. Shown are the Kaplan-Meier survival curves for the two groups.



Log-Rank Test

A two-sample t-test seems like an obvious choice: but the presence of censoring again creates a complication.

Therefore, we use log-rank test here

Log-Rank Test

$d_1 < d_2 < \ldots\ldots < d_K$ are the unique death times among the non-censored patients, r_k is the number of patients at risk at time d_k, and q_k is the number of patients who died at time d_k.

We further define r_{1k} and r_{2k} to be the number of patients in groups 1 and 2, respectively, who are at risk at time d_k.

Similarly, we define q_{1k} and q_{2k} to be the number of patients in groups 1 and 2, respectively, who died at time d_k. Note that $r_{1k} + r_{2k} = r_k$ and $q_{1k} + q_{2k} = q_k$.

Log-Rank Test

At each death time dk, we construct a 2x2 table of counts of the form shown above.

	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

Log Rank Test: the Main Idea

To test H0 : E(X) = 0 for some random variable X, one approach is to construct a test statistic of the form

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

where E(X) and Var(X) are the expectation and variance, respectively, of X under H0.

Log Rank Test: the Main Idea

In order to construct the log-rank test statistic, we compute a quantity that takes exactly the form above, with $X = \sum_{k=1}^{K} q_{1k}$, where q_{1k} is given in the top left of the table above.

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

When the sample size is large, the log-rank test statistic W has approximately a standard normal distribution.

Regression Models with a Survival Response

We wish to predict the true survival time T. Since the observed quantity Y = min(T;C) is positive and may have a long right tail, we might be tempted to fit a linear regression of log(Y) on X. But censoring again creates a problem.

To overcome this difficulty, we instead make use of a sequential construction, similar to the idea used for the Kaplan-Meier survival curve.

Hazard Function

The hazard function or hazard rate, also known as the force of mortality is formally defined as

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

where T is the (true) survival time.

It is the death rate in the instant after time t, given survival up to that time.

The Proportional Hazards Model

The proportional hazards assumption states that

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

where $h_0(t) \leq 0$ is an unspecified function, known as the baseline hazard. It is the hazard function for an individual with features $x_{i1}, x_{i2}, \ldots$

Proportional Hazards Model

The name proportional hazards arises from the fact that the hazard function for an individual with feature vector x_i is some unknown function $h_0(t)$ times the factor.

Because the form of the baseline hazard is unknown, we cannot simply plug $h(t|x)$ into the likelihood and then estimate $\beta_1, \beta_2, \ldots \beta_p$ by maximum likelihood.

Proportional Hazards Model

Therefore, the probability that the ith observation is the one to fail at time y_i (as opposed to one of the other observations in the risk set) is

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

Relative Risk Functions at each Failure Time

$$RR_1(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{1j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_1} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

$$RR_3(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{3j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_3} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

$$RR_5(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{5j}\beta_j\right)}{\sum_{i':y_{i'}\geq y_5} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

Partial Likelihood

To estimate β, we simply maximize the partial likelihood with respect to β. As is the case for logistic regression, no closed-form solution is available, and so iterative algorithms are required.

For example, we can obtain p-values corresponding to particular null hypotheses (e.g., $H_0: \beta_j$), as well as estimated standard errors and confidence intervals associated with the coefficients.