Introduction to Computational BioStatistics with R: survival analysis

Erik Spence

SciNet HPC Consortium

6 November 2025

SciNet

Today's slides

To find today's slides, go to the "Introduction to Computational BioStatistics with R" page, under Lectures, "Survival analysis".

`https://scinet.courses/1391`

What is survival analysis?

Survival analysis refers to a collection of methods that are used for studying situations where the dependent variable is the time until a specific event occurs.

- time: can be any unit (hours, days, years),
- event: can be anything (death, occurrence of a disease, heart attack),
- Examples might be:
  - time until tumor recurrence,
  - time until cardiovascular death after some treatment,
  - time until a machine part fails.

The subjects of the study are followed after a specified time period to determine when the event of interest occurs.

SciNet

Why bother?

What is special about survival analysis? Can't we use an lm or glm?

- Survival times are usually continuous,
- Survival times are strictly positive numbers. Regular linear regression may not be the best choice,
- The event in question corresponds to a binary variable,
- Data are often missing, incomplete or approximate. These data are called 'censored'.
- Regular regression techniques can't handle missing data.
- For some subjects we may know their survival time inexactly, meaning the survival time is at least equal to some value,
- while for other subjects we know their survival time exactly.

The ability to handle the mixture of exact, approximate and missing data makes survival analysis techniques special.

SciNet

Why bother?, continued

If there is no censoring, standard regression techniques might be sufficient.

But even if the data isn't censored, regular regression techniques may be inadequate.

- Time is restricted to be positive. Regular regression techniques assume positive and negative values of the independent variable.
- Time's distribution is usually skewed, with many early events and fewer later ones,
- The probability of surviving past a certain time may be of more interest than the expected time of an event,
- We're often more interested in the 'hazard function', a product of survival analysis. This function can sometimes give more insight into the failure mechanism than regular regression.

At the end of the day, the unusual nature of the data calls for a different analytical approach.

SciNet

Censoring data

Censoring of data occurs when information about a particular data point is incomplete.

- An event might not happen during the duration of the study,
- An event might occur, but we don't have the exact time.
- The event might only be measured within a certain range.
- Outside of this range, the only available information is that its value is greater or smaller than a specific value, or that it lies between two values.
- Analysing a censored variable requires special techniques.
- For the methods we will discuss to be valid, the censoring mechanism must be independent of the survival mechanism.

Censoring data complicates matters quite a bit.

SciNet

Examples of censoring

Censoring shows up all the time in medical studies.

- A subject does not experience the event before the study ends.
- A subject is lost to follow-up during the study period.
- A subject withdraws from the study.
- A subject dies for an unrelated reason.

There are three main types of censoring:

- right censoring: the above examples,
- left censoring: the subject enters the study after the event occurs.
- interval censoring: the event occurs between two successive surveys within the study.

Interval censoring is quite common, as you might imagine.

SciNet

Censoring mitigation

There are some 'simple' approaches to censoring that some researchers use:

- set the censored observations to 'missing',
- replace the missing value with
  - zero,
  - a minimum, maximum, or mean value,
  - a random value within an acceptable range.

As you might expect, you must be careful! This is very dangerous!

- when there is 'minimal' censoring, using one of the above approaches might be acceptable.
- when it is not minimal, the approaches can
  - cause serious bias in the results,
  - discard potentially important information,
  - create a sample that is not representative of the population being studied.

Messing with your data is always a dangerous game.

SciNet

Survival analysis definitions

Some terminology, to start.

- time, $T \geq 0$, is our independent variable,
- $S(t) = P(T > t)$ is our survival function
  - P is the probability,
  - t is time.
  - the survival function gives the probability that a subject will survive past time t.
- the survival function is non-increasing,
- at $t = 0$, $S(t) = 1$. The probability of surviving past time 0 is 1.
- at $t = \infty$, $S(t) = 0$.

In theory the survival function is smooth. In practice we observe events on a discrete time scale (days, weeks, etc).

The hazard function

The hazard function, $h(t)$, is the instantaneous rate at which events occur (the derivative of the survival function). It corresponds to the probability that an subject who is under observation at time t has an event at that time.

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

The cumulative hazard function describes the accumulated risk up to time t:

$$H(t) = \int_0^t h(u)du$$

If we know any of the functions $S(t)$, $H(t)$ or $h(t)$ we can derive the other two:

$$h(t) = -\frac{d\log(S(t))}{dt} \qquad H(t) = -\log(S(t)) \qquad S(t) = \exp(-H(t))$$

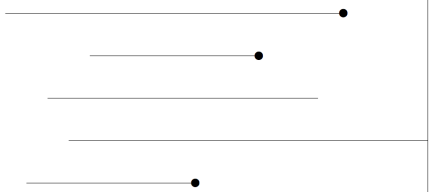
Data collection

Survival data is often represented this way:

- T_i is the event time for the subject,
- C_i is the censoring time for the subject,
- Y_i is the observed event, $\min(T_i, C_i)$,
- δ_i is the censoring status.

$$\delta_i = \begin{cases} 1 & \text{observation } (T_i \leq C_i) \\ 0 & \text{censoring } (T_i > C_i) \end{cases}$$

Note that the start times are not the same for these subjects.



T_i	C_i	Y_i	δ_i
80	100	80	1
40	80	40	1
74+	74	74	0
85+	85	85	0
40	95	40	1

Termination of study

Estimating the Survival function

Assuming that every subject follows the same survival function (there are no covariates or other important differences), it is possible to estimate $S(t)$. There are two families of techniques available:

- Non-parametric estimators
  - Kaplan-Meier estimator
- Parametric estimators
  - exponential
  - Weibull
  - Gamma
  - log-normal

As with the statistical techniques we've see before, non-parametric means we don't assume any functional form for the estimation (the data speaks for itself), while parametric means we do assume a functional form.

Non-parametric estimation

If we're not assuming any particular functional form to the survival function, we have two possible approaches:

- when no data are censored: $S(t) \simeq 1 - F(t)$, where $F(t)$ is the empirical cumulative distribution function.
- when some observations are censored: $S(t)$ can be estimated using the Kaplan-Meier product-limit estimator.

Both of these can be found in R packages.

Kaplan-Meier survival estimation

Kaplan-Meier survival estimation is the first non-parametric technique most people reach for. How does it work?

- Suppose that k patients have events in the period of followup, at distinct times: $t_1 < t_2 < t_3 < ... < t_k$.
- The probability of surviving from one interval to the next is multiplied together to give the cumulative survival probability.
- Thus, the probability of being alive at time t_j, $S(t_j)$, is calculated from $S(t_{j-1})$, n_j, the number of patients alive just before t_j, and d_j, the number of events at t_j:

$$S(t_j) = S(t_{j-1}) \left(1 - \frac{d_j}{n_j}\right)$$

where $S(0) = 1$. The value of $S(t)$ is a constant between times of events.

Creating a survival object

The standard package for doing survival analysis in R is the 'survival' package.

The 'ovarian' dataset contains survival data for two different treatments of ovarian cancer.

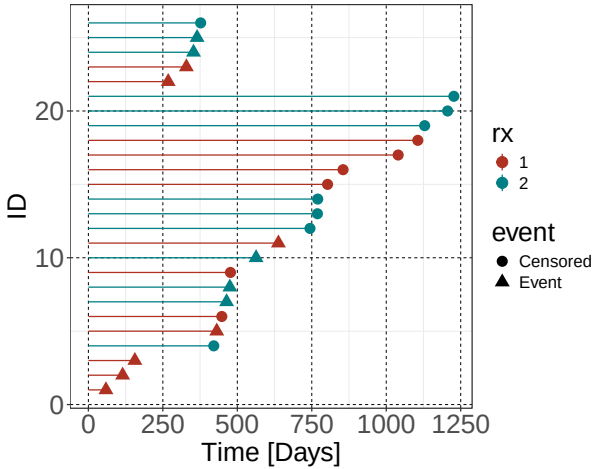
- The 'futime' column contains the survival or censoring time.
- The 'fustat' column contains the censoring status.

The 'Surv' function creates a 'survival' object.

```
>
> library(survival)
>
> S1 <- Surv(time = ovarian$futime,
+            event = ovarian$fustat)
>
> S1
 [1]  59 115 156 421+ 431 448+ 464 475 477+ 563
[11] 638 744+ 769+ 770+ 803+ 855+ 1040+ 1106+
[19] 1129+ 1206+ 1227+ 268 329 353 365 377+
>
```

Our censored data

```
>
> my.data <- ovarian
> my.data$rx <- as.factor(my.data$rx)
> my.data$ID <- 1:nrow(my.data)
> my.data$event <- ifelse(my.data$fustat,
+                "Event", "Censored")
>
> ggplot(my.data,
+        aes(x = ID, color = rx)) +
+    geom_linerange(aes(ymin = 0,
+                       ymax = futime)) +
+    coord_flip() + theme_bw() +
+    geom_point(aes(y = futime,
+                   shape = event)) +
+    labs(y = "Time [Days]")
>
```



Non-parametric estimation, example

```
> fit1 <- survfit(S1 ~ rx, data = my.data)
>
> summary(fit1)

Call:  survfit(formula = S1 ~ rx, data = my.data)

rx=1
 time   n.risk   n.event   survival   std.err   lower 95% CI   upper 95% CI
   59       13         1      0.923    0.0739          0.789          1.000
  115       12         1      0.846    0.1001          0.671          1.000
  156       11         1      0.769    0.1169          0.571          1.000
  268       10         1      0.692    0.1280          0.482          0.995
  329        9         1      0.615    0.1349          0.400          0.946
  431        8         1      0.538    0.1383          0.326          0.891
  638        5         1      0.431    0.1467          0.221          0.840
    ⋮
```

Non-parametric estimation, example, continued

```
    ⋮
rx=2
  time   n.risk   n.event   survival   std.err   lower 95% CI   upper 95% CI
   353       13         1      0.923    0.0739          0.789          1.000
   365       12         1      0.846    0.1001          0.671          1.000
   464        9         1      0.752    0.1256          0.542          1.000
   475        8         1      0.658    0.1407          0.433          1.000
   563        7         1      0.564    0.1488          0.336          0.946
>
```

The rx value indicates which treatment group is being considered. In this case there are two.

Non-parametric estimation, example, continued 2

```
> summary(fit1, times = 365 * 2)

Call:  survfit(formula = S1 ~ rx, data = my.data)

rx=1
    time   n.risk   n.event   survival   std.err   lower 95% CI   upper 95% CI
 730.000    4.000     7.000      0.431     0.147          0.221          0.840

rx=2
    time   n.risk   n.event   survival   std.err   lower 95% CI   upper 95% CI
 730.000    6.000     5.000      0.564     0.149          0.336          0.946
```

It's important to use the probability given above. If you calculate the "naive probability"

$$\left(1 - \frac{7}{13}\right) = 0.46$$

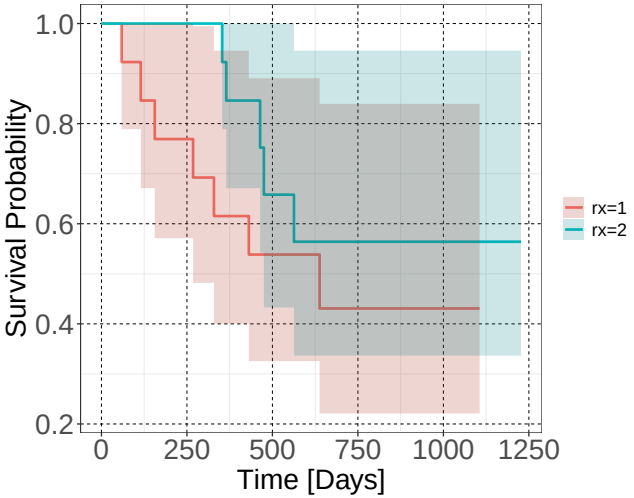
you will overestimate the survival. This is because those that are censored are considered event-free, which is not the case.

Non-parametric estimation, example, plotted

```
>
> library(ggsurvfit)
>
> ggsurvfit(fit1) +
+       add_confidence_interval() +
+       labs(x = "Time [Days]",
+            y = "Survival Prob")
>
```

Because this is non-parametric, the curves have steps in them.

The 95% confidence interval is given by the shaded area.



Non-parametric estimation, example, continued 3

Sometimes we're interested in the median survival time. This is given by directly examining the fit.

```
>
> fit1

Call:  survfit(formula = S1 ~ rx, data = my.data)

        n  events  median  0.95LCL  0.95UCL
rx=1   13       7     638      268       NA
rx=2   13       5      NA      475       NA
>
```

Note that the median for rx = 2 is NA because more than half of the patients survived the end of the study.

Parametric survival functions

The Kaplan-Meier estimator is a useful tool for estimating survival functions.

Sometimes, we may want to make more assumptions that allow us to model the data in more detail. By making these additional assumptions, and specifying a parametric form for $h(t)$, is becomes possible to:

- compute selected quantiles of the distribution,
- estimate the expected failure time,
- derive a concise equation and smooth function for estimating $S(t)$, $H(t)$ and $h(t)$,
- estimate $S(t)$ more precisely than KM assuming the parametric form is correct!

This is where parametric survival functions come in.

SciNet

Estimation of parametric survival functions

Maximum likelihood estimation is generally used to determine (*estimate*) the unknown parameters of a parametric distribution.

In survival analysis, the parametric form is usually assigned to the hazard function, rather than the survival function.

Some popular distributions for estimating hazard functions are

- constant (healthy people),
- increasing Weibull (leukemia patients),
- decreasing Weibull (patients recovering from surgery),
- exponential,
- log-normal (tuberculosis patients),
- log-logistic.

SciNet

Estimation of parametric survival functions, continued

Assuming that the time-to-event is a constant, with an exponential dependence,

$$H(t) = \exp(-\beta_0 - \beta_1 rx),$$

then

$$S(t) = \exp(-H(t)t)$$

thus,

$$S(t) = \exp(-(\exp(-6.87 - 0.61 rx)t)$$

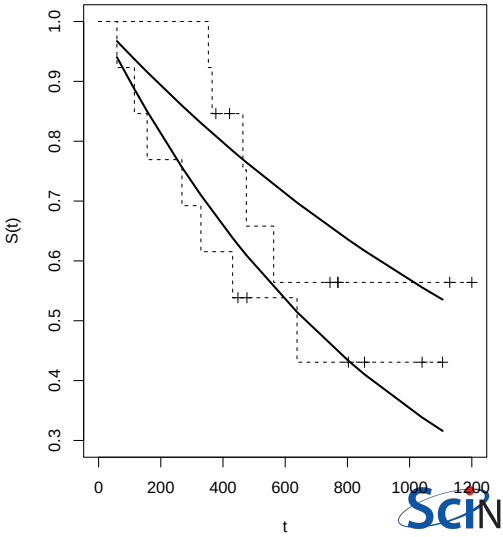
```
> s2 <- survreg(S1 ~ rx, data = my.data,
+     dist = 'exponential')
> summary(s2)
Call:
survreg(formula = S1 ~ rx, data = my.data,
                dist = "exponential")
             Value   Std. Error       z         p
 (Intercept)  6.868        0.378   18.17   <2e-16
 rx2          0.613        0.586    1.05      0.3

Scale fixed at 1

Exponential distribution
Loglik(model)= -97.5   Loglik(intercept only)= -98
Chisq= 1.11 on 1 degrees of freedom, p= 0.29
Number of Newton-Raphson Iterations:  4
n= 26
>
```

Estimation of parametric survival functions, plot

```
>
> plot(fit1, lt = 2, xlim = c(0, 1200),
+     ylim = c(0.3, 1), xlab = "t",
+     ylab = expression(hat(S)*"(t)"),
+     mark.time = TRUE)
>
> t <- sort(my.data$futime[my.data$rx == 1])
>
> lines(t, 1 - pexp(t, exp(-6.868)),
+     col = 'black', lwd = 2)
>
> lines(t, 1 - pexp(t, exp(-6.868 - 0.613)),
+     col = 'black', lwd = 2)
>
```



Comparing survival

Survival in two or more groups of subjects can be compared using a non-parametric test.

The most widely used test is the log-rank test.

- The approach calculates, for each time event and group, the number of events expected since the previous event if there was no difference between the two groups.
- The number of expected events for each time is then summed over all times, giving the total number of expected events in each group, E_i, for group i.
- The test then compares the observed number of events, O_i to the expected number:

$$\chi^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$$

This is compared to a χ^2 distribution with $g - 1$ degrees of freedom, where g is the number of groups. In this way a p-value can be calculated.

Comparing survival, continued

The 'survdiff' function calculates the log-rank test.

If there are only two groups, the null hypothesis is that the ratio of the hazard rates is equal to 1.

In this case there is no significant difference between the two groups. This is because the sample size is so small.

```
>
> survdiff(S1 ~ rx, data = my.data)
Call:  survdiff(formula = S1 ~ rx, data = my.data)

          N   Observed   Expected   (O-E)^2/E   (O-E)^2/V
 rx=1    13          7       5.23       0.596        1.06
 rx=2    13          5       6.77       0.461        1.06

Chisq= 1.1 on 1 degrees of freedom, p= 0.3
>
```

Cox regression

It may be that your survival curve depends upon multiple independent variables. Cox regression assumes the form for that dependence given at the bottom of the slide.

The quantity of interest is called the Hazards Ratio, which is the exponential of the coefficients. This represents the ratio of hazards between the two treatment groups. It is not a risk itself.

In this case 0.551 times as many members of rx=2 are dying as rx=1, at any given time.

```
>
> coxph(S1 ~ rx, data = my.data)
Call:
coxph(formula = S1 ~ rx, data = my.data)

          coef   exp(coef)   se(coef)        z      p
 rx2   -0.5964      0.5508     0.5870   -1.016   0.31

Likelihood ratio test=1.05 on 1 df, p=0.3052
n= 26, number of events= 12
>
```

$$h(t|\boldsymbol{X_i}) = h_0(t)\exp(\beta_1 \boldsymbol{X_{i1}} + ... + \beta_p \boldsymbol{X_{ip}})$$

Many different types of estimators

There are many different ways of estimating the survival function. Others to consider include

- Multistate Models
- Relative Survival
- Multivariate Survival
- Bayesian Models
- Higher-Dimension Models
- Time-varying covariates
- Time-dependent effects
- Poisson regression model
- Parametric Proportional Hazards Model
- Accelerated Failure Time Models
- Additive Models
- Buckley-James Models
- ...

There are many many approaches to consider.

Summary

Survival analysis is a collection of techniques which are used to model the time until an event occurs.

- These techniques are needed due to 'censored' data: data that are incomplete, approximate, or missing.
- The probability of an event occuring is given by the Survival function.
- Kaplan-Meier is a non-parametric technique used to estimate the Survival function.
- Non-parametric techiques for getting the Survival function also exist.
- Cox regression allows independent variables, other than time, to be added to the Survival function modelling.

SciNet