Part III

Part III: Survival data

Chapter 7

Working with time-to-event data

A data type that is commonly found in clinical trials is **time to event data**. This type of data captures the amount of time that elapses before a particular event happens. As a sub-field of statistics, survival analysis has been around for a long time, as people have thought about and worked with data like mortality records (most notably John Graunt, who used the 'Bills of Mortality' during the 1600s to better understand the plague and other causes of death). However, it developed rapidly during the many cancer related clinical trials of the 1960s and 1970s. In these cases, the event in question was very often death, and which is why this branch of statistics came to be known as **survival analysis**. However, the event can be many other things, and indeed can be a positive outcome (for example being cured of some condition). Time-to-event data also appears in other applications, such as engineering (eg. monitoring the reliability of a machine) and marketing (eg. thinking of the time-to-purchase). As well as the books already mentioned, this chapter makes use of Collett (2003b).

Usually, survival data is given in terms of time, but it can also be the number of times something happens (for example, the number of clinic appointments attended) before the event in question occurs.

Survival data is trickier to handle than the data types we have seen so far, for two main reasons. Firstly (and simply) survival data is very often skewed, so even though it is (usually) continuous, we can't just treat it as normally distributed. Secondly (and more complicatedly, if that's a word) with time-to-event data we don't usually observe the full dataset.

7.1 Censored times

If a trial monitors a sample of participants for some length of time, many will experience the event before the trial is finished. However, for some of the sample we likely won't observe the event. This could be because it doesn't happen within the lifetime of the trial, or it could be because the participant exits the trial prematurely for some reason (eg. withdrawal), or simply stops attending follow-up appointments after a certain times. For these participants, we do know that they had not experienced the event up to some time t, but we don't know what happened next. All we know is that their time-to-event or **survival time** is greater than that time t. These partial observations are known as **censored** times, and in particular as **right-censored** times, because the event happens *after* the censored time. It is possible (but less common) to have *left-censored* or *interval-censored* data, but in this course we will deal only with right-censoring.

If we were to treat censored times as observations, ie. as though the event had happened at time t, we would bias the results of the trial very seriously. The survival times reported would be systematically shorter than the true ones. For example, in the dataset shown in Figure 7.1, we would estimate the survival probability at time 10 as 0.2, since only two of the 10 participants were still in the trial after time 10. But it may well be that some of the participants whose observations were censored before $t = 10$ were still alive at $t = 10$.

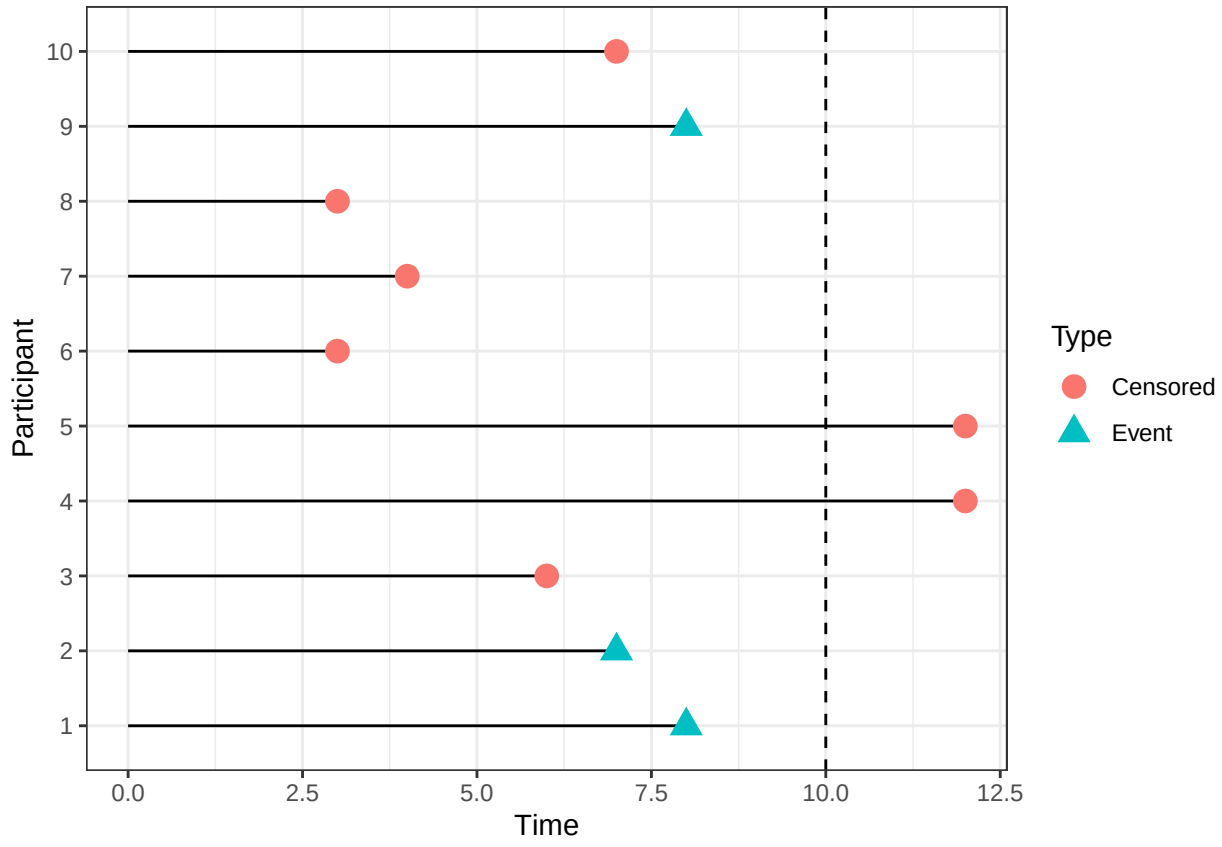


Figure 7.1: An example of some censored data. The dashed line at time 10 indicates the end of the trial period.

If we were to remove the censored times, and only analyse the data in which the event was observed during the lifespan of the trial, we would be losing data and therefore valuable information. This approach may well also lead to bias, for example if some subset of patients died quite soon into the trial, but the remainder lived a long time (past the end of the trial). If our analysis ignores the survivors, we are likely to underestimate the general survival time. In the dataset in Figure 7.1 there are five participants (3,6,7,8,10) whom we are no longer able to observe at time 10, but of whom none had experienced the event by the point at which they were censored.

So we know that we need to somehow include these censored times in our analysis. How we do so will depend on our approach.

7.2 The Survival Curve and the Hazard function

The field of survival analysis is relatively unusual in statistics, in that it isn't treated predominantly parametrically. For most continuous data, it is overwhelmingly common to work with the normal distribution and its friends (eg. the student's t distribution). Similarly binary data is dominated by the binomial distribution. Inference is therefore often focussed on the parameters μ, σ or p, as an adequate summary of the truth given whatever parameteric assumption has been made.

However, in survival analysis, it is often the case that we focus on the whole shape of the data; there isn't an accepted dominating probability distribution. In order to be able to deal with time-to-event data, we need to introduce some key ways of working with such data.

The **survival time** (or time-to-event) t for a particular individual can be thought of as the value of a random variable T, which can take any non-negative value. We can think in terms of a probability distribution over the range of T. If T has a probability distribution with underlying *probability density function* $f(t)$, then the *cumulative distribution function* is given by

$$F(t) = P(T < t) = \int_0^t f(u)\, du,$$

and this gives us the probability that the survival time is less than t.

Definition 7.1. The **survival function**, $\mathrm{S}(t)$, is the probability that some individual (in our context a participant) survives longer than time t. Therefore $\mathrm{S}(t) = 1 - F(t)$. Conventionally we plot $\mathrm{S}(t)$ against t and this gives us a **survival curve**.

We can immediately say two things about survival curves:

1. Since all participants must be alive (or equivalent) at the start of the trial, $\mathrm{S}(0) = 1$.
2. Since it's impossible to survive past $t_2 > t_1$ but not past time t_1, we must have $\frac{d\,\mathrm{S}(t)}{dt} \leq 0$, ie. $\mathrm{S}(t)$ is non-increasing.

One summary that is often used is the **median survival time**, the time at which half of the participants have already experienced the event, and half haven't.

Figure 7.2 shows two survival curves, comparing different therapies. We see that the hormonal therapy reduces the survival time slightly compared to no hormonal therapy.

Following on from the survival function, we have another (slightly less intuitive) quantity: the **Hazard function** $h(t)$.

Definition 7.2. The **Hazard function** $h(t)$ is the probability that an individual who has survived up to time t fails just after time t; in other words, the instantaneous probability of death (or *experiencing the event*) at time t.

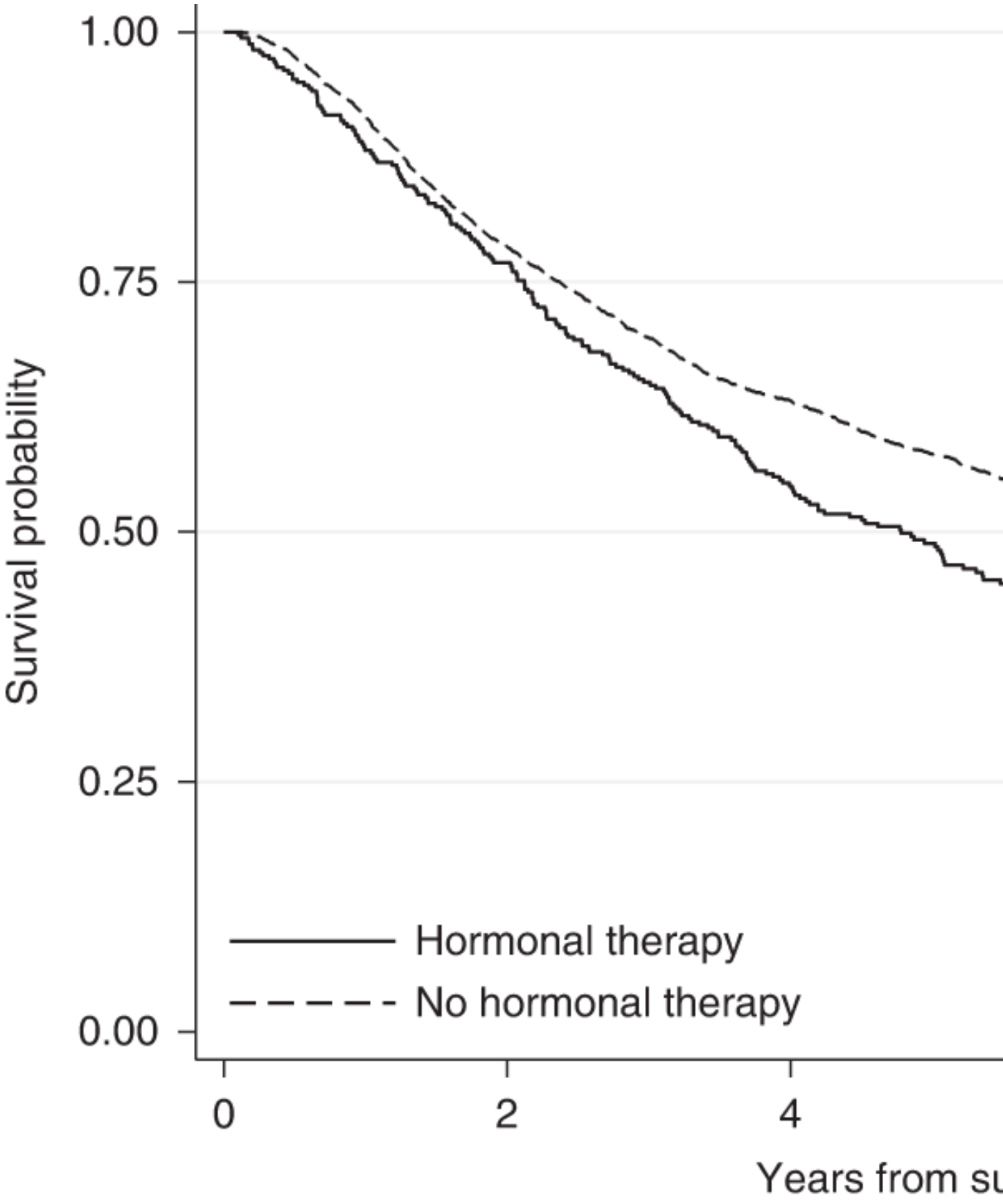


Figure 7.2: An example of two survival curves, taken from @syriopoulou2022standardised.

If we use T to denote the random variable of survival time (or time-to-event) then $\mathrm{S}(t)$ and $h(t)$ are defined by

$$\begin{aligned}\mathrm{S}(t) &= \Pr(T > t)\\ h(t) &= \lim_{s\to 0+} \frac{\Pr(t < T < t+s \mid T > t)}{s}.\end{aligned}$$

Using the definition of conditional probability, we can rewrite $h(t)$ as

$$\begin{aligned}h(t) &= \lim_{s\to 0+} \frac{\Pr(t < T < t+s \mid T > t)}{s}\\ &= \lim_{s\to 0+} \left[\frac{1}{\Pr(T > t)} \cdot \frac{\Pr((t < T < t+s) \cap (T > t))}{s}\right]\\ &= \lim_{s\to 0+} \left[\frac{1}{\Pr(T > t)} \cdot \frac{\Pr(t < T < t+s)}{s}\right]\\ &= \frac{f(t)}{\mathrm{S}(t)},\end{aligned}$$

where $f(\cdot)$ is the probability density of T. The hazard function can take any positive value (unlike the survival function), and for this reason $\log(h(t))$ is often used to transform it to the real line. The hazard function can also be called the 'hazard rate', the 'instantaneous death rate', the 'intensity rate' or the 'force of mortality'.

As we hinted before, there are fundamentally two ways to deal with survival data: we can go about things either parametrically or non-parametrically. Unusually for statistics in general, the non-parametric paradigm is prevalent in survival analysis. We will consider some methods from both paradigms.

7.2.1 The Kaplan-Meier estimator

The **Kaplan-Meier estimator** is a non-parametric estimate of $\mathrm{S}(t)$, originally presented in Kaplan and Meier (1958). The idea behind it is to divide the interval $[0,\ t]$ into many short consecutive intervals,

$$[0,\ t] = \bigcup_{k=0}^{K} [s_k,\ s_{k+1}],$$

where $s_k < s_{k+1}\ \forall k$, $s_0 = 0$ and $s_{K+1} = t$. We then estimate the probability of surviving past some time t by multiplying together the probabilities of surviving the successive intervals up to time t. No distributional assumptions are made, and the probability of surviving interval $[s_k,\ s_{k+1}]$ is estimated by $1 - Q$, where

$$Q = \frac{\text{Number who die in that interval}}{\text{Number at risk of death in that interval}}.$$

More precisely, let's say that deaths are observed at times $t_1 < t_2 < \ldots < t_n$, and that the number of deaths at time t_i is d_i out of a possible n_i. Then for some time $t \in [t_J,\ t_{J+1})$, the Kaplan-Meier estimate of $\mathrm{S}(t)$ is

$$\hat{\mathrm{S}}(t) = \prod_{j=0}^{J} \frac{(n_j - d_j)}{n_j}.$$

Notice that the number of people at risk at time t_{j+1}, denoted n_{j+1}, will be the number of people at risk at time t_j (which was n_j), minus any who died at time t_j (which we write as d_j) and any who were censored

Table 7.1: Ovarian cancer data. FU time gives the survival or censoring time, and FU status the type: 0 for a censored observation, 1 for death.

	FU_time	FU_status
1	59	1
2	115	1
3	156	1
22	268	1
23	329	1
24	353	1
25	365	1
26	377	0
4	421	0
5	431	1
6	448	0
7	464	1
8	475	1
9	477	0
10	563	1
11	638	1
12	744	0
13	769	0
14	770	0
15	803	0
16	855	0
17	1040	0
18	1106	0
19	1129	0
20	1206	0
21	1227	0

in the interval $[t_j, t_{j+1})$. In this way, the Kaplan-Meier estimator incorporates information from individuals with censored survival times up to the point they were censored.

Greenwood (1926) derived an approximation to the variance of the Kaplan-Meier estimate of survival curve, given by

$$\hat{V}(t) = \left(\hat{\mathrm{S}}(t)\right)^2 \sum_{t_i \leq t} \frac{d_i}{n_i (n_i - d_i)}$$

This uses the Delta method (which we've seen before in the binary outcome chapters) and makes the assumption that events at time t_i are independent binomial draws from a population of size n_i. We can use this to form confidence intervals for the survival curve, and indeed `ggsurvfit` can add these automatically to plots.

Example 7.1. Edmonson et al. (1979) conducted a trial on patients with advanced ovarian cancer, comparing cyclophosphamide (group C) with a mixture of cyclophosphamide and adriamycin (group T). Patients were monitored, and their time of death was recorded, or a censoring time if they were alive at their last observation. The data are shown in Table 7.1.

We see that there are 26 individuals, and we have the time of death for 12 of them. The remaining 14 observations are censored. We can use this data to calculate the Kaplan-Meier estimator of the survival curve, as shown in Table 7.2. The columns are (from left to right): time t_j; number at risk n_j; number of

Table 7.2: Kaplan-Meier estimator calculations for ovarian cancer dataset.

time	n_risk	n_event	n_cens	survival	SE_surv
59	26	1	0	**0.9615385**	0.0377146
115	25	1	0	**0.9230769**	0.0522589
156	24	1	0	**0.8846154**	0.0626563
268	23	1	0	**0.8461538**	0.0707589
329	22	1	0	**0.8076923**	0.0772920
353	21	1	0	**0.7692308**	0.0826286
365	20	1	0	**0.7307692**	0.0869893
431	17	1	2	**0.6877828**	0.0918815
464	15	1	1	**0.6419306**	0.0965213
475	14	1	0	**0.5960784**	0.0999261
563	12	1	1	**0.5464052**	0.1032094
638	11	1	0	**0.4967320**	0.1051027

events/deaths d_j; number of censorings in $[t_{j-1}, t_j)$; estimate of survival curve and standard error of the estimate (using Greenwood's formula).

Figure 7.3 shows the Kaplan-Meier survival curve estimate for the `ovarian` data. Using the package `ggsurvfit` we can add in a table below the x axis showing the number at risk and the number of events at some times points.

The Kaplan-Meier estimate may seem a bit dissatisfying, since it stops changing at $t = 638$ with a probability of 0.497. However, this is really (in a non-parametric setting) all we can say with the data available; 10 of the participants were definitely still alive at $t = 638$, and some of the other censored participants may also have been.

For a clinical trial, we want to plot the survival curves separately for the different treatment groups. This will give a first, visual, idea of whether there might be a difference, and also of the suitability of certain models (we'll talk about this later).

Example 7.2. Figure 7.4 shows the Kaplan Meier plots for the ovarian cancer data from Figure 7.3, this time split by treatment group.

The second dataset we will use throughout this chapter has been simulated based on a trial of acute myeloid leukemia (Le-Rademacher et al. (2018)) and is from the `survival` package Therneau (2024). The Kaplan-Meier estimate for the myeloid data is shown in Figure 7.5.

7.2.2 A parametric approach

In a parametric approach, we'll assume that the survival time T follows some probability distribution, up to unknown parameters which we will estimate from the data. The simplest distribution for time-to-event data is the *exponential distribution*, which has density

$$f(t) = \lambda e^{-\lambda t} \text{ for } t > 0,$$

survival function

$$\mathrm{S}(t) = 1 - \int_0^t \lambda e^{-\lambda t} = e^{-\lambda t},$$

and mean survival time $\frac{1}{\lambda}$. The hazard function is therefore

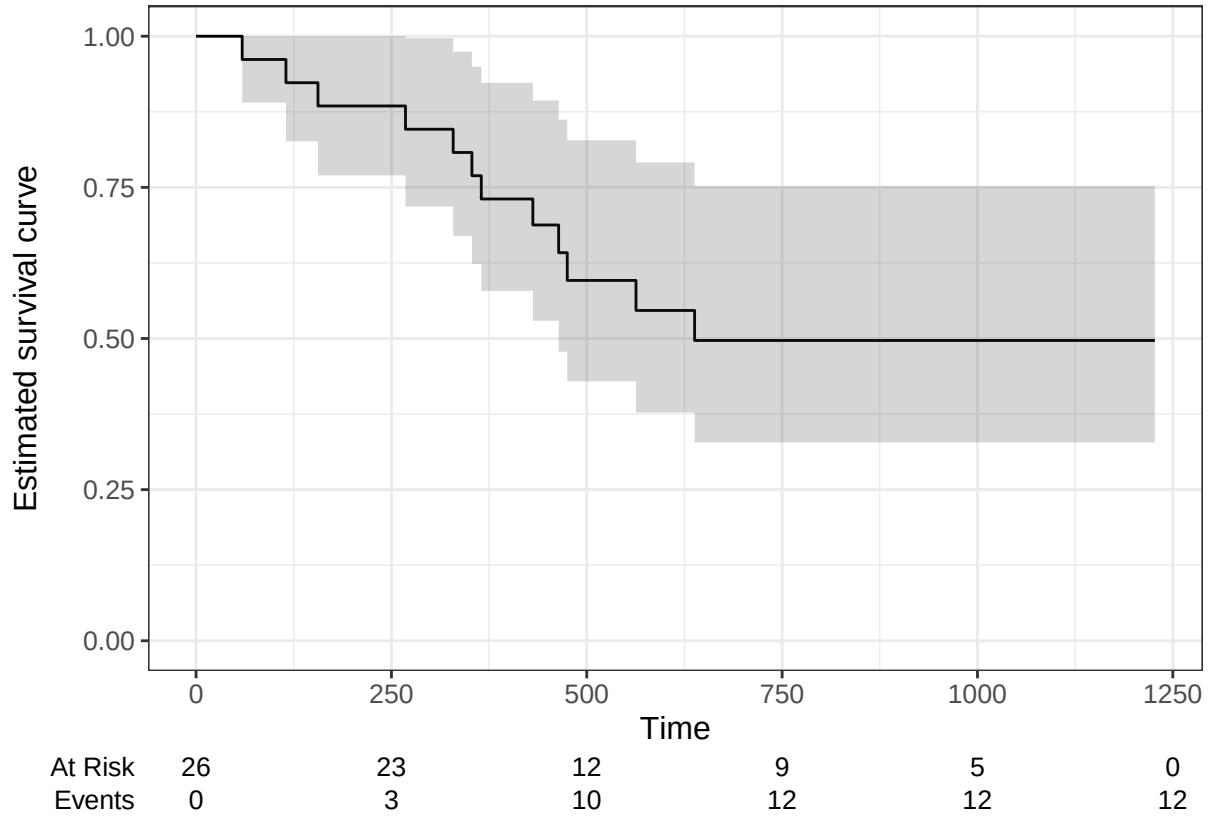


Figure 7.3: Kaplan-Meier estimate of survival curve for ovarian cancer data with a 95% confidence level.

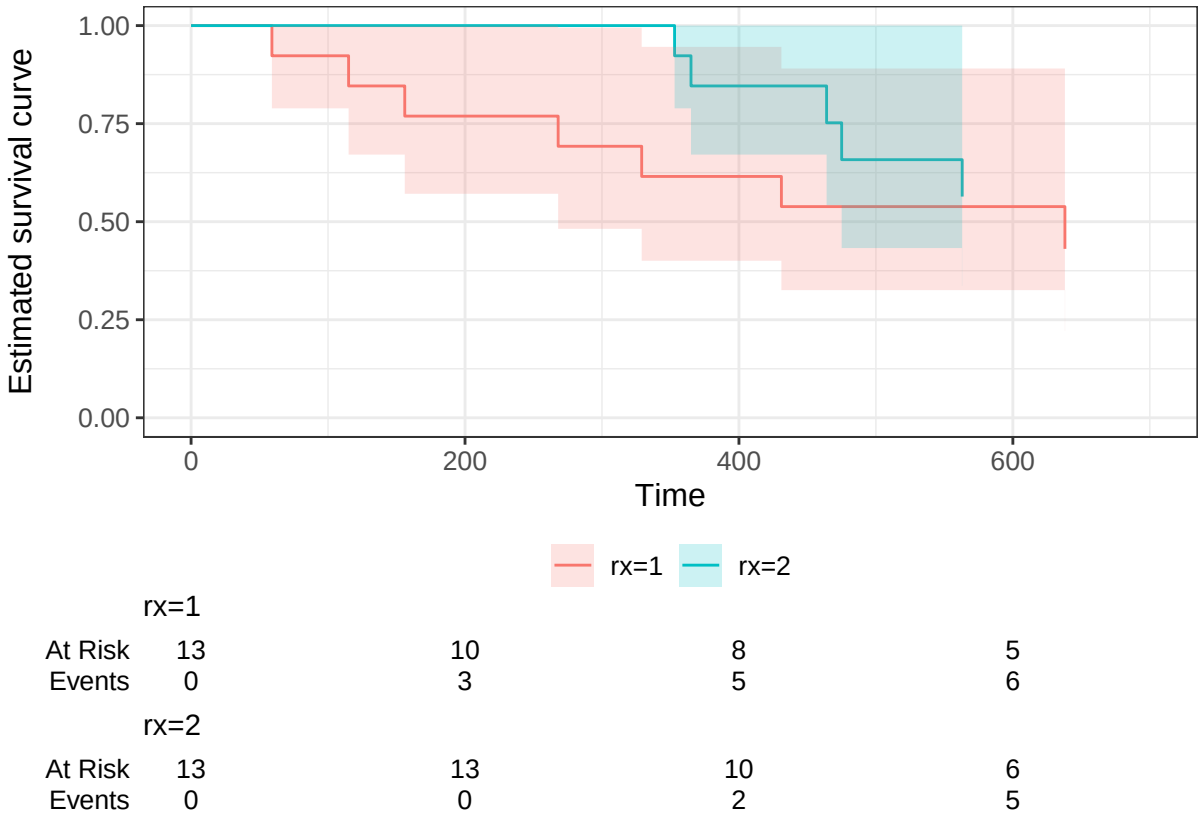


Figure 7.4: Kaplan-Meier curves for the ovarian cancer data, split by treatment group.

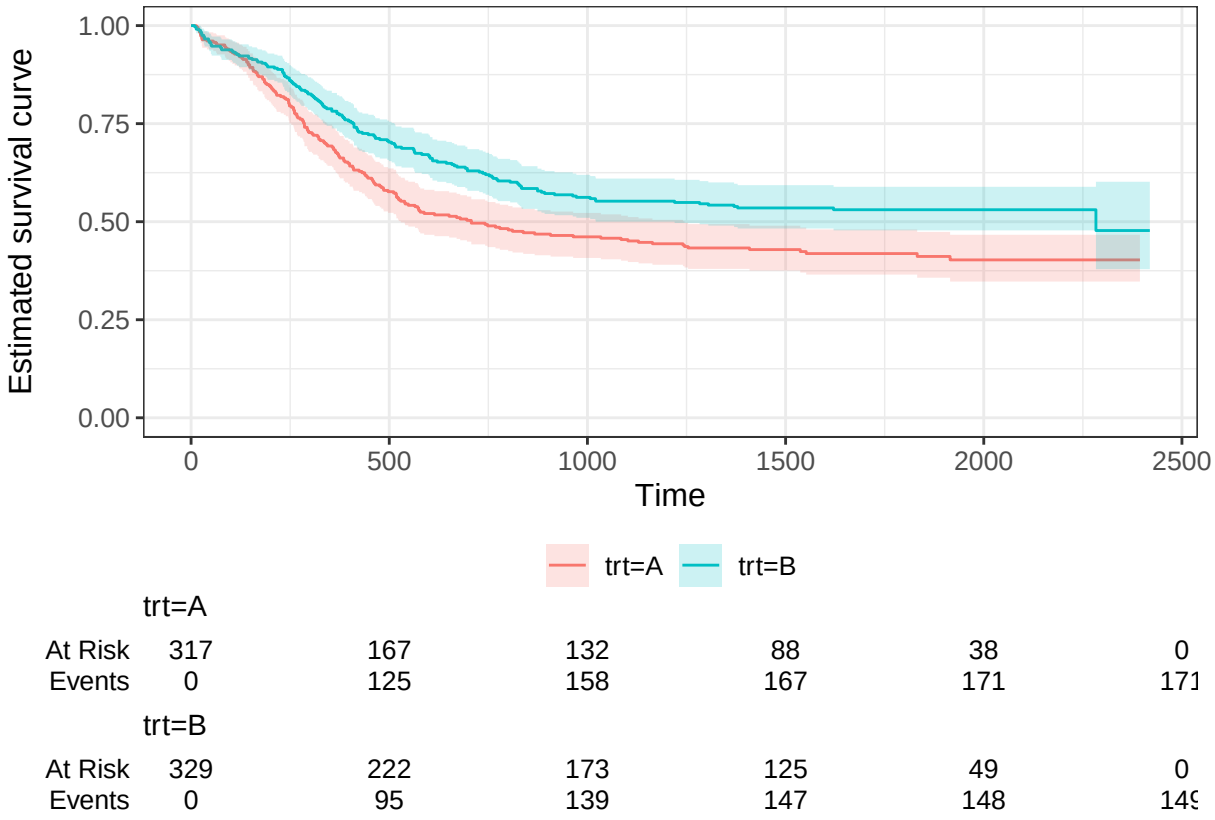


Figure 7.5: Kaplan Meier curves for the Myeloid data, split by treatment.

$$h(t) = \frac{f(t)}{S(t)} = \lambda,$$

that is, the hazard is constant.

Given some dataset, we want to be able to find an estimate for λ (or the parameters of our distribution of choice).

7.2.2.1 Maximum likelihood for time-to-event data

Suppose our dataset has n times $t_1, t_2, \ldots, t_n$. Of these, m are fully observed and $n - m$ are censored. We can create a set of indicators $\delta_1, \ldots, \delta_n$, where $\delta_i = 1$ if observation i is fully observed and $\delta_i = 0$ if it is censored.

Usually, the likelihood function is computed by multiplying the density function evaluated at each data point, $f(t_i \mid \text{params})$. However, this won't work for survival data, because for our censored times (those for which $\delta_i = 0$) we only know that the time-to-event is greater than t_i. For these observations, it is the survival function (remember that this is $p(T > t)$) that contributes what we need to the likelihood function.

Therefore (for any probability distribution) we have

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{7.1}$$

If we have $T \sim Exp(\lambda)$ then the log-likelihood is

$$\begin{aligned}\ell(\lambda \mid \text{data}) &= \sum_{i=1}^{n} \delta_i (\log \lambda - \lambda t_i) - \sum_{i=1}^{n} (1 - \delta_i) \lambda t_i \\ &= m \log \lambda - \lambda \sum_{i=1}^{n} t_i.\end{aligned}$$

From this we can find the maximum likelihood estimator (MLE)

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}.$$

The variance of the MLE is

$$\text{var}\left(\hat{\lambda}\right) = \frac{\lambda^2}{m}, \tag{7.2}$$

which we can approximate by

$$\text{var}\left(\hat{\lambda}\right) \approx \frac{m}{\left(\sum_{i=1}^{n} t_i\right)^2}.$$

Notice that the numerator in Equation (7.2) is m, the number of complete observations (rather than n the total number including censored observations). This shows that there is a limit to the amount we can learn if a lot of the data is censored.

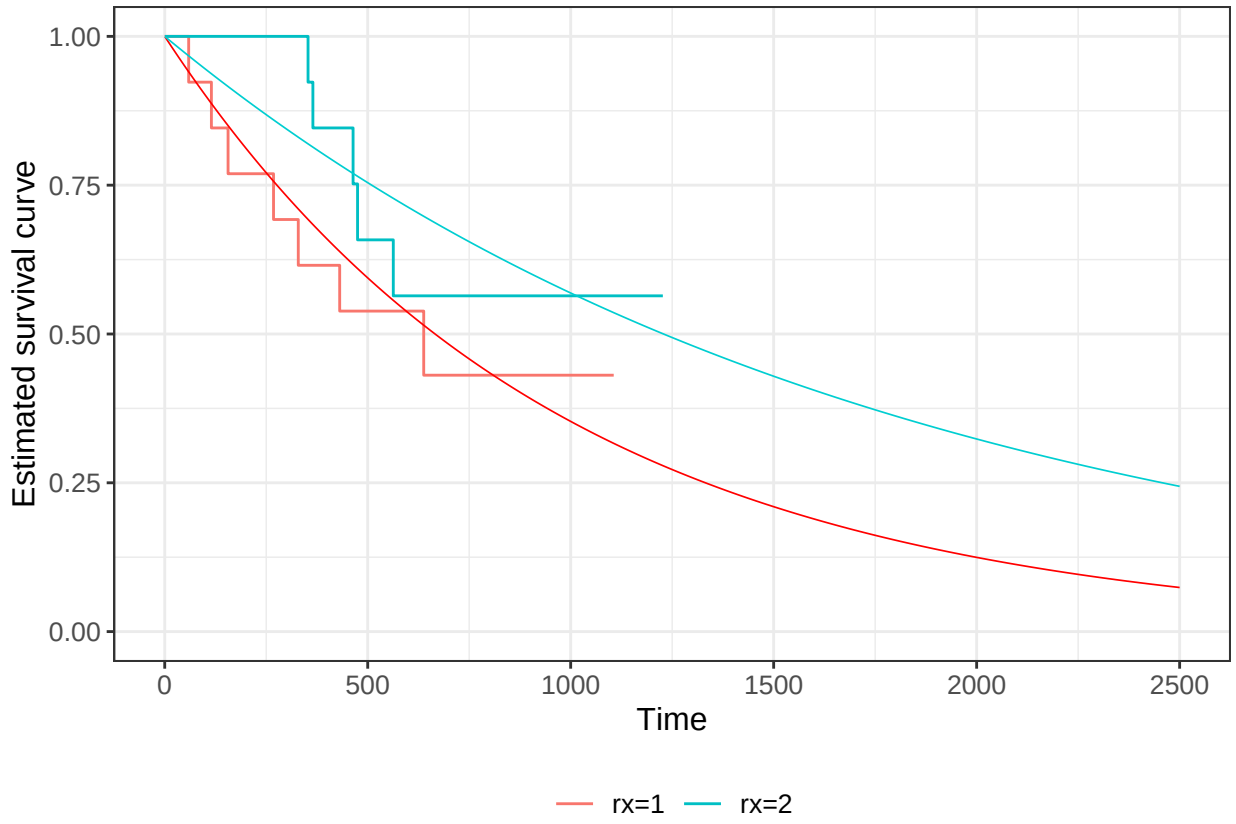
Example 7.3. Returning to the dataset from Example 7.1, we can fit an exponential distribution to the data simply by estimating the MLE

$$\begin{aligned}\hat{\lambda}_C &= \frac{m_C}{\sum_{i=1}^{n_C} t_i} \\ &= \frac{7}{6725} \\ &= 0.00104\end{aligned}$$

and

$$\begin{aligned}\hat{\lambda}_T &= \frac{m_T}{\sum_{i=1}^{n_T} t_i} \\ &= \frac{5}{8863} \\ &= 0.00056\end{aligned}$$

```
mC_ov = sum((ovarian$fustat==1)&(ovarian$rx==1))
mT_ov = sum((ovarian$fustat==1)&(ovarian$rx==2))
tsum_ov_C = sum(ovarian$futime[ovarian$rx==1])
tsum_ov_T = sum(ovarian$futime[ovarian$rx==2])
m_ov = mT_ov + mC_ov
tsum_ov = tsum_ov_C + tsum_ov_T
lamhat_ov_C = mC_ov / tsum_ov_C
lamhat_ov_T = mT_ov / tsum_ov_T
```



We can do the same for the `myeloid` data. Figure 7.6 shows the fitted curves, using $S(t) = \exp\left[-\hat{\lambda}_X t\right]$ for group X.

7.2.3 The Weibull distribution

Having only one parameter, the exponential distribution is not very flexible, and often doesn't fit data at all well. A related, but more suitable distribution is the **Weibull distribution**.

Definition 7.3. The probability density function of a **Weibull** random variable is

$$f(t \mid \lambda, \gamma) = \begin{cases} \lambda\gamma t^{\gamma-1} \exp\left[-\lambda t^{\gamma}\right] & \text{for } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here, γ is the *shape* parameter, and λ is the *scale* parameter. If $\gamma = 1$ then this reduces to an exponential distribution. You can read more about it, should you choose to, in Collett (2003b).

For the Weibull distribution, we have

$$\mathrm{S}(t) = \exp\left(-\lambda t^{\gamma}\right).$$

As with the exponential distribution, we can use Equation (7.1) for the likelihood. For the Weibull distribution this becomes

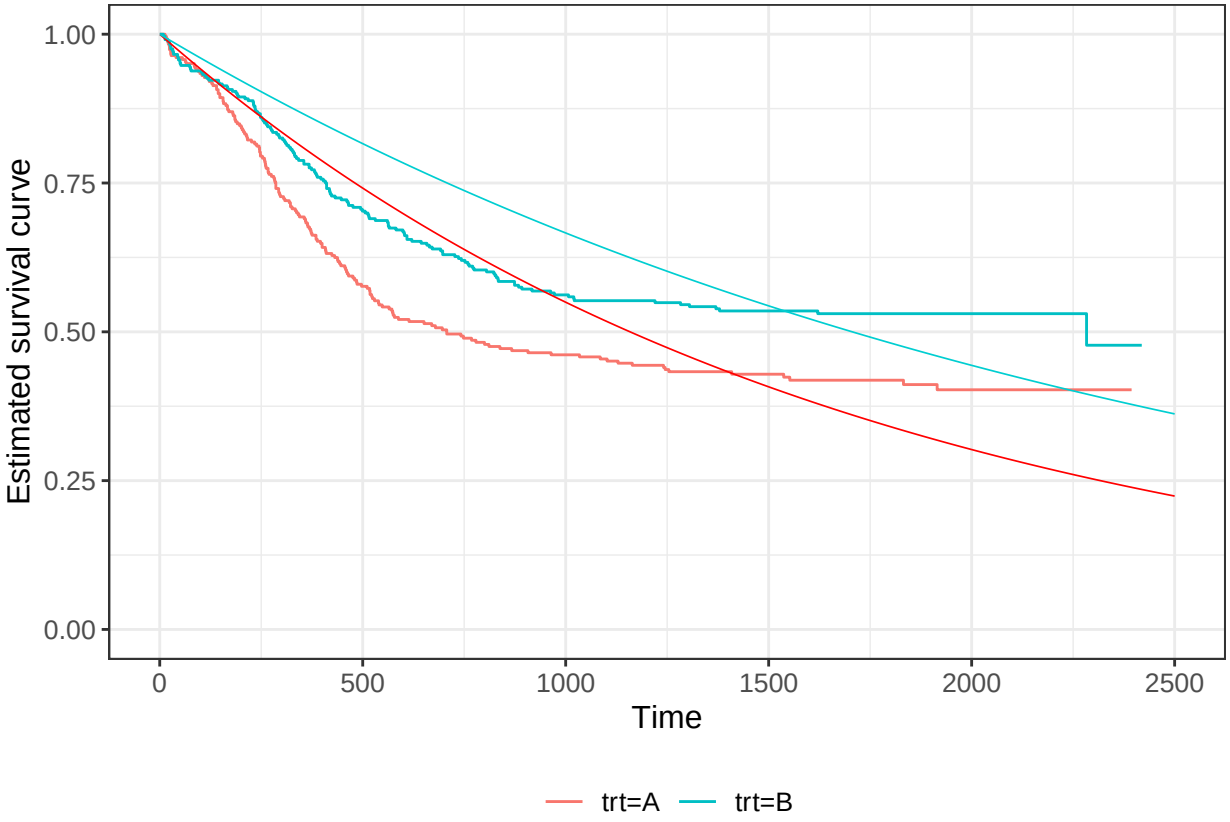


Figure 7.6: Kaplan Meier estimates of survival curves for the Myeloid data (solid lines), with the fitted exponential S(t) shown in dashed lines (red = group C, blue = group T).

$$\begin{aligned} L(\lambda, \gamma \mid \text{data}) &= \prod_{i=1}^{n} \left\{ \lambda\gamma t_i^{\gamma-1} \exp\left(-\lambda t_i^{\gamma}\right) \right\}^{\delta_i} \left\{ \exp\left[-\lambda t_i^{\gamma}\right] \right\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \lambda\gamma t_i^{\gamma-1} \right\}^{\delta_i} \exp\left(-\lambda t_i^{\gamma}\right) \end{aligned}$$

and therefore

$$\begin{aligned} \ell(\lambda, \gamma \mid \text{data}) &= \sum_{i=1}^{n} \delta_i \log(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma} \\ &= m \log(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma}. \end{aligned}$$

For the maximum likelihood estimators, we differentiate (separately) with respect to λ and γ and equate to zero, to solve for the estimators $\hat{\lambda}$ and $\hat{\gamma}$.

The equations we end up with are

$$\frac{m}{\hat{\lambda}} - \sum_{i=1}^{n} t_i^{\hat{\gamma}} = 0 \tag{7.3}$$

$$\frac{m}{\hat{\gamma}} + \sum_{i=1}^{n} \delta_i \log t_i - \hat{\lambda} \sum_{i=1}^{n} t_i^{\hat{\gamma}} \log t_i = 0. \tag{7.4}$$

We can rearrange Equation (7.3) to

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}},$$

and substitute this into Equation (7.4) to find

$$\frac{m}{\hat{\gamma}} + \sum_{i=1}^{n} \delta_i \log t_i - \frac{m}{\sum_{i=1}^{n} t_i^{\gamma}} \sum_{i=1}^{n} t_i^{\hat{\gamma}} \log t_i = 0.$$

This second equation is analytically intractable, so numerical methods are used to find $\hat{\gamma}$, and then this value can be used to find $\hat{\lambda}$.

Example 7.4. We can fit Weibull distributions to our `myeloid` dataset, as shown in Figure 7.7.

$$S(t) = \exp\left(-\lambda t^{\gamma}\right).$$

We see that there is some improvement compared to the exponential fit in Figure 7.6, but it still seems not to capture the fundamental shape.

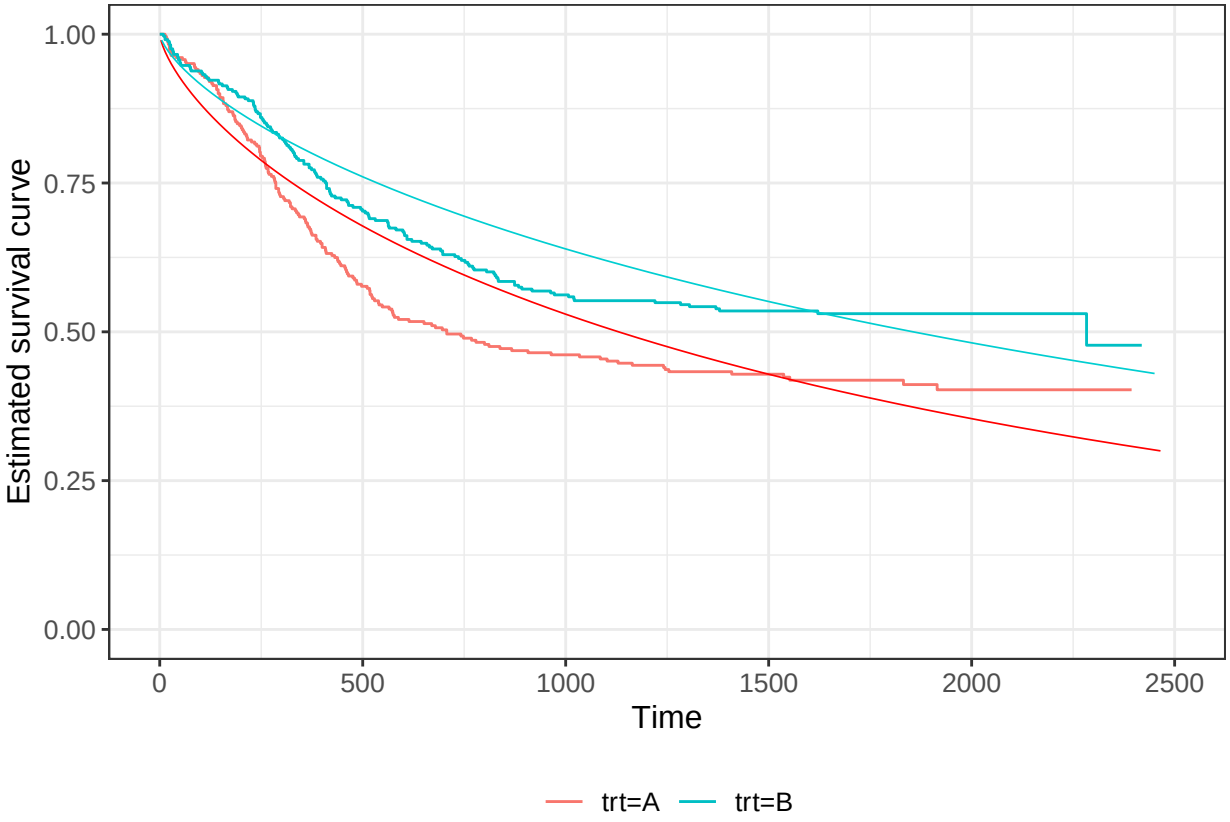


Figure 7.7: Weibull fit to survival curve of Myeloid data, dashed lines (Kaplan Meier estimate also shown in solid lines). Red for group T, black for group C.

Aside: Sample size calculations for time-to-event data

There are implications here for sample size calculations, which must take into account the duration of a trial; it is important that trials monitor patients until a sufficient proportion have experienced the event (whatever it is). Sample size calculations for time-to-event data therefore have two components:

1. The power of the trial can first be expressed in terms of m, the number of complete observations.
2. A separate calculation is needed to estimate the number of participants needing to be recruited, and length of trial, to be sufficiently likely to achieve that value of m.

Both of these calculations rely on a number of modelling assumptions, and on previous scientific/clinical data (if available).

We will think more about how this can be used in the next section, when we come to compare treatment effects.

Chapter 8

Comparing survival curves

Really what we would like to be able to do is to compare two survival curves (showing, for example, the results from different treatments), so that we can say whether one is significantly different from the other. In most cases, this boils down to constructing a hypothesis test along the lines of

$$H_0 : \text{the treatments are the same}$$
$$H_1 : \text{the treatments are different.}$$

There are various ways to do this, and we will look at some now.

8.1 Parametric: likelihood ratio test

For a parametric analysis, our null hypothesis that the two treatments are the same can be reduced to a test of whether the parameter(s) for each group are the same. We can do this using a likelihood ratio test. We've already calculated the log-likelihood for the exponential distribution in Section 7.2.2.1, and found the MLE.

$$\ell(\lambda) = m \log \lambda - \lambda \sum_{i=1}^{n} t_i$$
$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}.$$

Working with the exponential distribution, we can model the survival function as

$$S(t) = \begin{cases} e^{-\lambda_C t} & \text{for participants in group C} \\ e^{-\lambda_T t} & \text{for participants in group T} \end{cases}$$

and the null hypothesis boils down to

$$H_0 : \ \lambda_C = \lambda_T = \lambda.$$

We can adapt the log-likelihood we found in Section 7.2.2.1 in light of the separate groups, and we find

$$\ell(\lambda_C, \lambda_T) = m_C \log \lambda_C - \lambda_C \sum_{i=1}^{n_C} t_{iC} + m_T \log \lambda_T - \lambda_T \sum_{i=1}^{n_T} t_{iT} \tag{8.1}$$

and

$$\hat{\lambda}_X = \frac{m_X}{\sum_{i=1}^{n_X} t_{iX}}$$

where $X = C$ or T. In these equations m_X is the number of non-censored observations in group X, n_X is the total number of participants in group X and t_{iX} is the time for participant i in group X. To simplify notation, we will write

$$t_X^+ = \sum_{i=1}^{n_X} t_{iX},$$

and t^+ for the sum over both groups.

Substituting the MLEs into Equation (8.1) gives

$$\ell\left(\hat{\lambda}_C, \hat{\lambda}_T\right) = m_C \log\left(\frac{m_C}{t_C^+}\right) - m_C + m_T \log\left(\frac{m_T}{t_T^+}\right) - m_T$$

and

$$\ell\left(\hat{\lambda}, \hat{\lambda}\right) = m \log\left(\frac{m}{t^+}\right) - m,$$

where n, m are the corresponding totals over both groups.

We can therefore perform a maximum likelihood test by finding

$$\begin{aligned}
\lambda_{LR} &= -2\left[\ell\left(\hat{\lambda}, \hat{\lambda}\right) - \ell\left(\hat{\lambda}_C, \hat{\lambda}_T\right)\right] \\
&= 2\left[\left(m_C \log\left(\frac{m_C}{t_C^+}\right) - m_C + m_T \log\left(\frac{m_T}{t_T^+}\right) - m_T\right) - \left(m \log\left(\frac{m}{t^+}\right)\right)\right] \\
&= 2\left(m_C \log\left(\frac{m_C}{t_C^+}\right) + m_T \log\left(\frac{m_T}{t_T^+}\right) - m \log\left(\frac{m}{t^+}\right)\right)
\end{aligned}$$

and referring this value to a χ_1^2 distribution.

We can also find a confidence interval for the difference between λ_T and λ_C, by using the asymptotic variances of the MLEs, which are $\frac{\lambda_C^2}{m_C}$ and $\frac{\lambda_T^2}{m_T}$. Therefore, the limits of a $100\,(1-\alpha)\%$ CI for $\lambda_T - \lambda_C$ is given by

$$\frac{m_T}{t_T^+} - \frac{m_C}{t_C^+} \pm z_{\alpha/2} \sqrt{\frac{m_T}{\left(t_T^+\right)^2} + \frac{m_C}{\left(t_C^+\right)^2}}.$$

Example 8.1. In this example we'll conduct a likelihood ratio test for each of the datasets in Example 7.1. For each dataset, the quantities we need are:

- m_C, m_T: the number of complete observations in each group
- t_C^+, t_T^+ the sum of all observation times (including censored times) in each group

Note that $m = m_C + m_T$ and $t^+ = t_C^+ + t_T^+$.

For the `ovarian` data we have

```
mC_ov = sum((ovarian$fustat==1)&(ovarian$rx==1))
mT_ov = sum((ovarian$fustat==1)&(ovarian$rx==2))
tsum_ov_C = sum(ovarian$futime[ovarian$rx==1])
tsum_ov_T = sum(ovarian$futime[ovarian$rx==2])
m_ov = mT_ov + mC_ov
tsum_ov = tsum_ov_C + tsum_ov_T

## Can now plug these into LR test stat
LRstat_ov =  2*(mC_ov*log(mC_ov/tsum_ov_C) + mT_ov*log(mT_ov/tsum_ov_T) - m_ov*log(m_ov/tsum_ov))
LRstat_ov
```

```
## [1] 1.114895
```

We can find the p-value of this test by

```
1-pchisq(LRstat_ov, df=1)
```

```
## [1] 0.2910204
```

and we find that it isn't significant. A 95% confidence interval for the difference is given by

```
## [1] -0.0013927697  0.0004392714
```

For the Myeloid data we can do the same thing

```
mC_my = sum((myeloid$death==1)&(myeloid$trt=="A"))
mT_my = sum((myeloid$death==1)&(myeloid$trt=="B"))
tsum_my_C = sum(myeloid$futime[myeloid$trt=="A"])
tsum_my_T = sum(myeloid$futime[myeloid$trt == "B"])
m_my = mT_my + mC_my
tsum_my = tsum_my_C + tsum_my_T

## Can now plug these into LR test stat
LRstat_my =  2*(mC_my*log(mC_my/tsum_my_C) + mT_my*log(mT_my/tsum_my_T) - m_my*log(m_my/tsum_my))
LRstat_my
```

```
## [1] 11.95293
```

Again, we refer this to χ_1^2:

```
1-pchisq(LRstat_my, df=1)
```

```
## [1] 0.0005456153
```

This time we find that the difference is significant at even a very low level, and the 95% CI is given by

```
## [1] -3.028814e-04 -8.108237e-05
```

Although the confidence around $\hat{\lambda}_X$ is high (ie. small standard error of the estimate), because of the large amount of data, the fit appears to actually be rather poor (recall Figure 7.6), mainly because of the inflexibility of the exponential distribution.

One feature of the exponential model that is convenient is that the hazard function is constant. Comparisons between treatment groups in survival trials are often summarised by the **hazard ratio**: the ratio of the hazard functions for the two groups. In general this is a function of t, but for two exponential hazard functions it is simply the ratio of the λ values.

We could also perform LR tests with the fitted Weibull distributions, but instead we will continue on through some more commonly used methods.

8.2 Non-parametric: the log-rank test

The log-rank test is performed by creating a series of tables, and combining the information to find a test statistic.

We work through each time t_j at which an event is observed (by which we mean a death or equivalent, not a censoring) in either of the groups.

For notation, we will say that at time t_j,

- n_j patients are 'at risk' of the event
- d_j events are observed (often the 'event' is death, so we will sometimes say this)

For groups C and T we would therefore have a table representing the state of things at time t_j, with this general form:

Group	No. surviving	No. events	No. at risk
Treatment	$n_{Tj} - d_{Tj}$	d_{Cj}	n_{Cj}
Control	$n_{Cj} - d_{Cj}$	d_{Tj}	n_{Tj}
Total	$n_j - d_j$	d_j	n_j

Under H_0, we expect the deaths (or events) to be distributed proportionally between groups C and T, and so the expected number of events in group X (C or T) at time t_j is

$$e_{Xj} = n_{Xj} \times \frac{d_j}{n_j}.$$

This means that $e_{Cj} + e_{Tj} = d_{Cj} + d_{Tj} = d_j$.

If we take the margins of the table (by which we mean n_j, d_j, n_{Cj} and n_{Tj}) as fixed, then d_{Cj} has a **hypergeometric distribution**.

Definition 8.1. The **hypergeometric distribution** is a discrete probability distribution describing the probability of k successes in n draws (without replacement), taken from a finite population of size N that has exactly K objects with the desired feature. The probability mass function for a variable X following a hypergeometric function is

$$p\left(X = k \mid K, N, n\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

An example would be an urn containing 50 (N) balls, of which 16 (K) are green and the rest (34, $N - K$) are red. If we draw 10 (n) balls **without replacement**, X is the random variable whose outcome is k, the number of green balls drawn.

In the notation of the definition, the mean is

$$\mathrm{E}(X) = n\frac{K}{N}$$

and the variance is

$$\mathrm{var}(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}.$$

In the notation of our table at time t_j, we have

$$\mathrm{E}(d_{Cj}) = e_{Cj} = n_{Cj} \times \frac{d_j}{n_j}$$
$$\mathrm{var}(d_{Cj}) = v_{Cj} = \frac{d_j n_{Cj} n_{Tj} (n_j - d_j)}{n_j^2 (n_j - 1)}$$

With the marginal totals fixed, the value of d_{Cj} fixes the other three elements of the table, so considering this one variable is enough.

Under H_0, the numbers dying at successive times are independent, so

$$U = \sum_j (d_{Cj} - e_{Cj})$$

will (asymptotically) have a normal distribution, with

$$U \sim N\left(0, \sum_j v_{Cj}\right).$$

We label $V = \sum_j v_{Cj}$, and in the log-rank test we refer $\frac{U^2}{V}$ to χ^2_1.

A somewhat simpler, and more commonly used, version of the log-rank test uses the fact that under H_0, the expected number of events (eg. deaths) in group X is $E_X = \sum_j e_{Xj}$, and the observed number is $O_X = \sum_j d_{Xj}$. The standard χ^2 test formula can then be applied, and the test-statistic is

$$\frac{(O_C - E_C)^2}{E_C} + \frac{(O_T - E_T)^2}{E_T}.$$

It turns out that this test statistic is always smaller than $\frac{U^2}{V}$, so this test is slightly more conservative (ie. it has a larger p-value).

Notice that for both of these test statistics, the actual difference between observed and expected is used, not the absolute difference. Therefore if the differences change in sign over time, the values are likely to cancel out (at least to some extent) and the log-rank test is not appropriate.

Example 8.2. Let's now perform a log-rank test on our data from Example 8.1.

First, the ovarian cancer dataset. To do this, we can tabulate the key values at each time step.

Time	n_Cj	d_Cj	e_Cj	n_Tj	d_Tj	e_Tj	n_j	d_j
59	13	1	0.5000000	13	0	0.5000000	26	1
115	12	1	0.4800000	13	0	0.5200000	25	1
156	11	1	0.4583333	13	0	0.5416667	24	1
268	10	1	0.4347826	13	0	0.5652174	23	1
329	9	1	0.4090909	13	0	0.5909091	22	1
353	8	0	0.3809524	13	1	0.6190476	21	1
365	8	0	0.4000000	12	1	0.6000000	20	1
431	8	1	0.4705882	9	0	0.5294118	17	1
464	6	0	0.4000000	9	1	0.6000000	15	1
475	6	0	0.4285714	8	1	0.5714286	14	1
563	5	0	0.4166667	7	1	0.5833333	12	1
638	5	1	0.4545455	6	0	0.5454545	11	1

From this, we can find the v_j and the test statistic $\frac{U^2}{V}$:

```
# Add up the differences
UC = sum(logrank_df$d_Cj - logrank_df$e_Cj)
vCj_vec = sapply(
  1:n_event,
  function(j){
    nCj = logrank_df$n_Cj[j]
    nTj = logrank_df$n_Tj[j]
    dj = logrank_df$d_j[j]
    nj = logrank_df$n_j[j]

    (nCj*nTj*dj*(nj-1))/((nj^2)*(nj-1))
})
VC = sum(vCj_vec)
cs_ov_stat = (UC^2)/VC
1-pchisq(cs_ov_stat, df=1)
```

```
## [1] 0.3025911
```

For the simpler, more conservative, version of the log-rank test, we have

```
EC = sum(logrank_df$e_Cj)
ET = sum(logrank_df$e_Tj)
OC = sum(logrank_df$d_Cj)
OT = sum(logrank_df$d_Tj)

test_stat = ((EC-OC)^2)/EC + ((ET-OT)^2)/ET

test_stat
```

```
## [1] 1.057393
```

and we can find the p-value by

```
1-pchisq(test_stat, df=1)
```

```
## [1] 0.3038106
```

As we expected, slightly larger, but not much different from the first version. These values are also pretty close to the results of our LR test in Example 8.1, where we had $p = 0.291$.

Since the Myeloid dataset is much bigger, we won't go through the rigmarole of making the table, but will instead use an inbuilt R function from the `survival` package (more on this in practicals).

```
myeloid$trt = as.factor(myeloid$trt)
survdiff(Surv(futime, death) ~ trt, data = myeloid, rho=0)
```

```
## Call:
## survdiff(formula = Surv(futime, death) ~ trt, data = myeloid,
##     rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=A 317      171      143      5.28      9.59
## trt=B 329      149      177      4.29      9.59
##
##  Chisq= 9.6  on 1 degrees of freedom, p= 0.002
```

This time the p-value is quite far from the one we found using the likelihood ratio test (p=0.00055), further supporting the view that the likelihood ratio test was not appropriate because of the poor fit of the exponential distribution.

8.3 Semi-parametric: the proportional hazards model

As with continuous and binary outcome variables, what we would really like to be able to do is to adjust our model for baseline covariates. It seems intuitively reasonable to suppose that factors like age, sex, disease status etc. might affect someone's chances of survival (or whatever event we're concerned with).

The conventional way to do this is using a **proportional hazards model**, where we assume that

$$h_T(t) = \psi h_C(t)$$

for any $t > 0$ and for some constant $\psi > 0$. We call ψ the **relative hazard** or **hazard ratio**. If $\psi < 1$ then the hazard at time t under treatment T is smaller than under control C. If $\psi > 1$ then the hazard at time t is greater in greater in group T than in group C. The important point is that ψ doesn't depend on t. The hazard for a particular patient might be greater than for another, due to things like their age, disease history, treatment group and so on, but the extent of this difference doesn't change over time.

We can adopt the concept of a **baseline hazard function** $h_0(t)$, where for someone in group C (for now), their hazard at time t is $h_0(t)$, and for someone in group T it is $\psi h_0(t)$. Since we must have $\psi > 0$, it makes sense to set

$$\psi = e^{\beta},$$

so that $\beta = \log \psi$ and $\psi > 0 \; \forall \beta \in \mathbb{R}$. Note that $\beta > 0 \iff \psi > 1$.

We can now (re)-introduce our usual indicator variable G_i, where

$$G_i = \begin{cases} 0 \text{ if participant } i \text{ is in group } C \\ 1 \text{ if participant } i \text{ is in group } T \end{cases}$$

and model the hazard function for participant i as

$$h_i(t) = \exp[\tau G_i] h_0(t).$$

This is the proportional hazards model for the comparison of two groups. Now, the relative hazard is a function of the participant's characteristics. Naturally, we can extend it to include other baseline covariates, as we have with linear models in ANCOVA, and with logistic regression.

8.3.1 General proportional hazards model

Extending the model to include baseline covariates $X_1, \ldots, X_p$, we have

$$\psi(\mathbf{x}_i) = \exp(\tau G_i + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right),$$

where we collect τ into $\boldsymbol{\beta}$ and G into $\mathbf{x}$, and the hazard function for participant i is

$$h_i(t) = \psi(\mathbf{x}_i) h_0(t).$$

Now, our baseline hazard function $h_0(t)$ is the hazard function for a participant in group C for whom all baseline coviariates are either zero (if continuous) or the reference level (if a factor variable). For factor covariates this makes sense, since all levels are realistic values, but for continuous variables zero is likely to be unrealistic (for example you'd never expect zero for age, weight, height, blood pressure etc.). So, if any continuous variables are present, the baseline will always need to be adjusted, but if all covariates are factors, it is likely that the baseline hazard function will be applicable for some set of participants.

The linear component $\mathbf{x}_i^T \boldsymbol{\beta}$ is often called the **risk score** or **prognostic index** for participant i.

The general form of the model is therefore

$$h_i(t) = \exp\left[\mathbf{x}_i^T \boldsymbol{\beta}\right] h_0(t), \tag{8.2}$$

and we can rewrite it as

$$\log\left(\frac{h_1(t)}{h_0(t)}\right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Notice that there is no constant in the linear term - if there was, it could just be absorbed into the baseline hazard function.

There are ways of fitting this model that rely on specifying the hazard function using parametric methods, but the method we will study (and the most widely used) is one developed by Cox (1972).

8.3.1.1 Interpreting the parameters in a proportional hazards model

Since our primary interest is in comparing the effect of some new treatment with that of the control, it is important that we understand what the coefficients mean, and in particular how they relate to the treatment effect. Let's do that (as usual) by considering two participants who are identical in all baseline covariates, one in group C and one in group T. We have

$$\begin{aligned} h_i^C(t) &= \exp(\beta_1 x_{1i} + \ldots + \beta_p x_{pi}) h_0(t) && \text{in group C} \\ h_i^T(t) &= \exp(\tau + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}) h_0(t) && \text{in group T.} \end{aligned}$$

From this we can find the **hazard ratio** at time t for the two treatments to be

$$\frac{h_i^T(t)}{h_i^C(t)} = \exp(\tau)$$

and τ is the log of the hazard ratio for the two treatments, adjusting for the other covariates. A value of $\tau = 0$ implies a hazard ratio of one, and of no evidence of difference between the treatments.

8.3.2 Cox regression

The beauty of Cox regression is that it avoids specifying a form for $h_0(t)$ altogether.

To fit the model in Equation (8.2) we must estimate the coefficients $\boldsymbol{\beta} = (\tau, \beta_1, \ldots, \beta_p)^T$. It also appears as though we should estimate the baseline hazard $h_0(t)$ somehow too, but the great advance made by Cox was to develop a method where this isn't necessary. We don't need to estimate $h_0(t)$ to make inferences about the hazard ratio

$$\frac{h_i(t)}{h_0(t)}.$$

We will estimate the coefficients $\boldsymbol{\beta}$ using maximum likelihood, and so we'll need to specify a likelihood function for the $\boldsymbol{\beta}$, which will be a function of $\mathbf{x}^T\boldsymbol{\beta}$ and our observed data, the survival times t_i.

Suppose we have data for n participants, and that these include m complete observations (often referred to as deaths) and $n-m$ right-censored survival times. Suppose also that all the complete observation times are distinct. Since time itself is continuous, this is always technically true, but in data the time will be rounded and so there may be multiple observations at one time.

We can order these m event times

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)},$$

such that $t_{(j)}$ is the time of the j^{th} event to be observed.

At time $t_{(j)}$, there will be some number of individuals who are 'at risk' of the event, because either their observation time or their censored survival time is greater than $t_{(j)}$. The set of these individuals is the **risk set**, denoted $R\left(t_{(j)}\right)$.

Cox (1972) shows that the relevant likelihood function for the proportional hazards model in Equation (8.2) is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp\left[\mathbf{x}_{(j)}^T\boldsymbol{\beta}\right]}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left[\mathbf{x}_l^T\boldsymbol{\beta}\right]} \tag{8.3}$$

where $\mathbf{x}_{(j)}$ is the vector of covariates for the individual who dies (or equivalent) at time $t_{(j)}$. Notice that the product is over only those individuals with complete observations, but individuals with censored data do contribute to the sum in the denominator.

The numerator of the fraction inside the product in Equation (8.3) is the relative hazard for the person who actually did die at time $t_{(j)}$. The denominator is the sum of the relative hazards for all those who possibly could have died at time $t_{(j)}$ (the risk set $R\left(t_{(j)}\right)$). Thus, in very loose terms, maximizing the likelihood means finding values for $\boldsymbol{\beta}$ that mean the people who did die were 'the most likely' to die at the time they did.

Notice that this is not a true likelihood, since it depends only on the ordering of the data (the observation and censoring times) and not the data itself. This makes it a **partial likelihood**. The argument given to justify this is that because the baseline hazard $h_0(t)$ has an arbitrary form, it's possible that except for at these observed times, $h_0(t) = 0$, and therefore $h_i(t) = 0$. This means the intervals between successive observations convey no information about the effect of the covariates on hazard, and therefore about the $\boldsymbol{\beta}$ parameters.

If you want to know more detail about how this likelihood was derived, you can find in in Section 3.3 of Collett (2003b), or in Cox's original paper (Cox (1972)).

Moving on, if we set

$$\delta_i = \begin{cases} 0 & \text{if individual } i \text{ is censored} \\ 1 & \text{if individual } i \text{ is observed} \end{cases}$$

then we can write Equation (8.3) as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} \left(\frac{\exp\left[\mathbf{x}_i^T \boldsymbol{\beta}\right]}{\sum_{l \in R(t_i)} \exp\left[\mathbf{x}_l^T \boldsymbol{\beta}\right]} \right)^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i.

From this we can find the log-likelihood

$$\ell(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \log \sum_{l \in R(t_i)} \exp\left(\mathbf{x}_l^T \boldsymbol{\beta}\right) \right].$$

The MLE $\hat{\boldsymbol{\beta}}$ is found using numerical methods (often Newton-Raphson, which you'll have seen if you did Numerical Analysis II).

How can we tell if a proportional hazards model is appropriate?

We can't easily visualise the hazard function for a dataset, and instead would plot the survival curve. So can we tell if the proportional hazards assumption is met by looking at the survival curve?

It turns out that if two hazard functions are proportional, their survival functions won't cross one another, as we will show now.

Suppose $h_C(t)$ is the hazard at time t for an individual in group C, and $h_T(t)$ is the hazard for that same individual in group T. If the two hazards are proportional then we have

$$h_C(t) = \psi h_T(t)$$

for some constant ψ.

Recall from Section 7.2 that

$$h(t) = \frac{f(t)}{S(t)},$$

where $S(t)$ is the survival function and $f(t)$ is the probability density of T. We can therefore write

$$h(t) = -\frac{d}{dt}\left[\log\left(S(t)\right)\right]$$

and rearrange this to

$$S(t) = \exp(-H(t)) \tag{8.4}$$

where

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

Therefore for our two hazard functions, we have

$$\exp\left\{-\int_0^t h_C(u)\, du\right\} = \exp\left\{-\int_0^t \psi h_T(u)\, du\right\}$$

From Equation (8.4) we see that therefore

$$S_C(t) = [S_T(t)]^{\psi}.$$

Since the survival function is always between 0 and 1, we can see that the value of ψ determines whether $S_C(t) < S_T(t)$ (if $\psi > 1$) or $S_C(t) > S_T(t)$ (if $0 < \psi < 1$). The important thing is that **the survival curves will not cross**. This is an informal conclusion, and lines not crossing is a necessary condition but not a sufficient one. It may also be that the survival curves cross when a particular [influential] covariate is factored out, but not when it isn't.

Example 8.3. First of all, we can use Cox regression adjusted only for the Group (or treatment arm) of the participants.

For the `ovarian` dataset

```
coxph(formula = Surv(futime, fustat)~rx, data=ovarian)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ rx, data = ovarian)
##
##       coef exp(coef) se(coef)      z    p
## rx -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
```

and for the `myeloid1` dataset

```
coxph(formula = Surv(futime, death)~trt, data=myeloid)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ trt, data = myeloid)
##
##         coef exp(coef) se(coef)      z       p
## trtB -0.3457    0.7077   0.1122 -3.081 0.00206
##
## Likelihood ratio test=9.52  on 1 df, p=0.002029
## n= 646, number of events= 320
```

We see that for both results, our p-values are close to what we have found with the log rank test.

For the `ovarian` dataset there is no evidence of a significant difference (likely due to the small sample size).

For the `myeloid` data we find that there is evidence of a difference - we can use the coefficient estimate and standard error to construct a 95% confidence interval for the log hazard ratio of

$$-0.346 \pm 1.96 \times 0.112 = (-0.566,\ -0.126)$$

and therefore for the hazard ratio itself of

$$(0.568,\ 0.881)\,.$$

We see that there is strong evidence that the intervention reduces the hazard.

We can also account for more baseline covariates. For the `ovarian` data we can include `age` and `resid.ds` (whether residual disease is present):

```
coxph(formula = Surv(futime, fustat)~rx+age+resid.ds, data=ovarian)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ rx + age + resid.ds, data = ovarian)
##
##              coef exp(coef) se(coef)      z       p
## rx       -0.8489    0.4279   0.6392 -1.328 0.18416
## age       0.1285    1.1372   0.0473  2.718 0.00657
## resid.ds  0.6964    2.0065   0.7585  0.918 0.35858
##
## Likelihood ratio test=16.77  on 3 df, p=0.0007889
## n= 26, number of events= 12
```

What this shows is that the most significant factor by far is the particpant's age, with the hazard function increasing as age increases. The coefficient for treatment group (`rx`) has increased in magnitude and the p-value has decreased now that age is being adjusted for (although it is still not significant).

We can do the same for the `myeloid` data:

```
coxph(formula = Surv(futime, death)~trt+sex, data=myeloid)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ trt + sex, data = myeloid)
##
##         coef exp(coef) se(coef)      z       p
## trtB -0.3582    0.6989   0.1129 -3.174 0.00151
## sexm  0.1150    1.1219   0.1128  1.020 0.30782
##
## Likelihood ratio test=10.56  on 2 df, p=0.005093
## n= 646, number of events= 320
```

We see that the only covariate we have, `sex` has very little effect, and that our confidence interval for the treatment effect will not have changed much at all.

8.3.3 Diagnostics for Cox regression

Having fit a Cox proportional hazards model, it's important to check that it is an appropriate fit to the data. We've seen already that the survival curves mustn't cross, but there are other more sophisticated methods we can use to assess the model.

It is important to examine the proportional hazards assumption for every covariate we include in the model (including the group / arm variable), and how we do this depends on whether the covariate is continuous or categorical.

8.3.3.1 Continuous variables

Schoenfeld (1982) derived partial residuals, known as **Schoenfeld residuals**, that can be used to assess whether the proportional hazards assumption is appropriate for a continuous variable.

We can think of $X_i = (X_{i1}, \ldots, X_{ip})'$, the set of covariates for a participant who experiences the event at time t_i, as a random variable. Schoenfeld (1982) showed that

$$E\left(X_{ij} \mid R_i\right) = \frac{\sum\limits_{k \in R_i} X_{kj} \exp\left(\beta' X_k\right)}{\sum\limits_{k \in R_i} \exp\left(\beta' X_k\right)},$$

where R_i are the indices of those at risk at time t_i. You can think of this as the average of the $X_{\cdot j}$ values of those at risk as time t_i, weighted by their relative hazard. We can write

$$\hat{E}\left(X_{ij} \mid R_i\right)$$

to denote this quantity with the MLE $\hat{\beta}$ substituted for β.

The partial residual at time t_i is therefore the vector

$$\hat{r} = \left(\hat{r}_{i1}, \ldots, \hat{r}_{ip}\right),$$

where

$$\hat{r}_{ik} = X_{ik} - \hat{E}\left(X_{ik} \mid R_i\right).$$

If we plot the Schoenfeld residuals against time, we should see a random scatter around zero (the kind of plot we look for when assessing residuals against fitted values of a linear regression model).

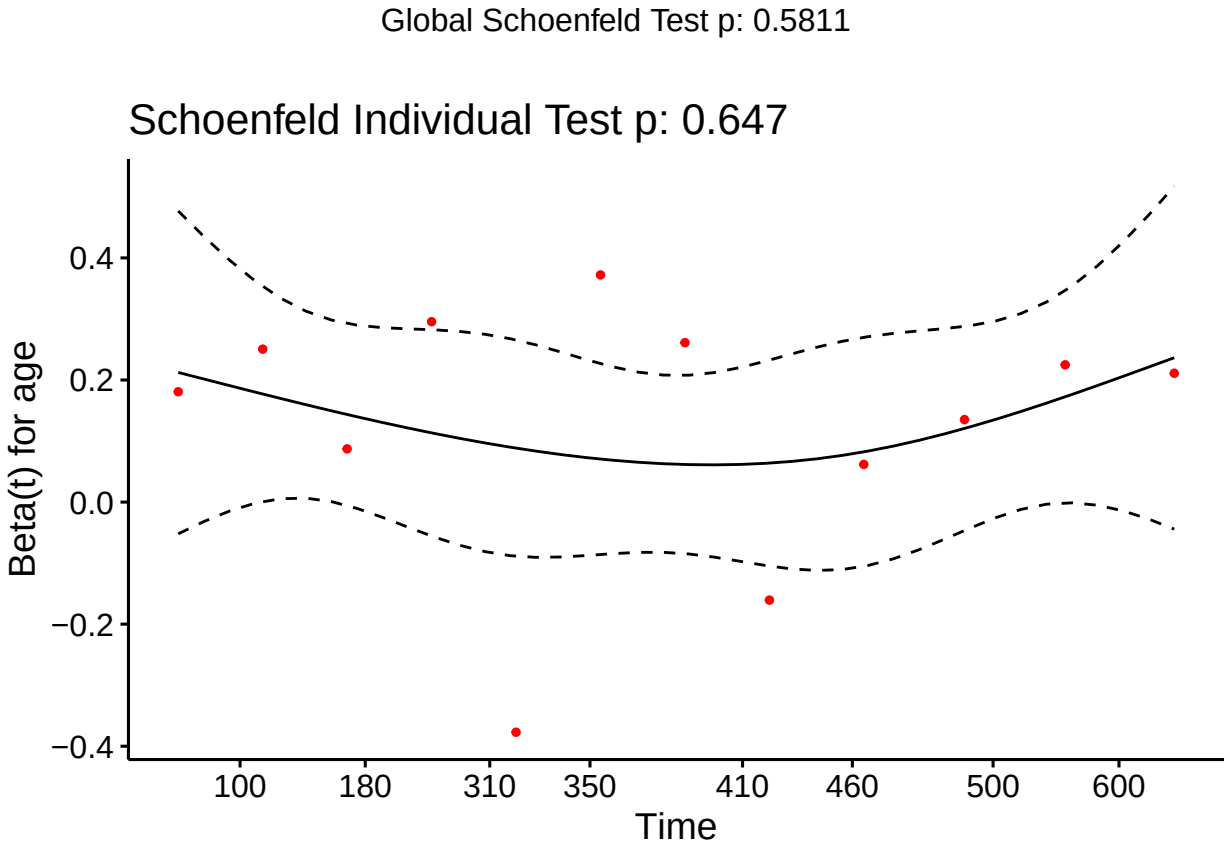
Grambsch and Therneau (1994) proposed a statistical test using the Schoenfeld residuals, in which the null hypothesis is that the proportional hazards assumption holds. This can be implemented by the function `cox.zph` in the `survival` package.

Example 8.4. The `ovarian` data contains the continuous variable `age`, and so we can test the assumption of proportional hazards in relation to age using Schoenfeld residuals.

```
##          chisq df    p
## rx       0.631  1 0.43
## age      0.210  1 0.65
## resid.ds 1.120  1 0.29
## GLOBAL   1.958  3 0.58
```

We see from the `age` line that the data are consistent with the proportional hazards assumption.

The object created by `cox.zph` also contains the Schoenfeld residuals themselves, and so we can plot them:



Because there is so little data it's hard to conclude anything, and indeed our lack of significance in the test may be due to small sample size rather than excellent model fit.

8.3.3.2 Categorical variables

Recall from Equation (8.4) that

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

where $H(t)$ is the cumulative hazard function

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

From this we find that

$$\log(S(t)) = -H(t).$$

If we have two groups A and B for which the proportional hazards assumption is satisfied, then for some constant ψ

$$H_A(t) = \psi H_B(t).$$

We can combine these two equations to find

$$\begin{aligned}\log[-\log(S_A(t))] = \log(H_A(t)) &= \log\psi + \log(H_B(t)) \\ &= \log\psi + \log[-\log(S_B(t))].\end{aligned}$$

Under the Cox Regression model, $\log(H(t))$ is linear in the covariates, and so we will have [roughly] parallel lines. We can plot this in R using `ggsurvplot` and setting `fun="cloglog"`.

Example 8.5. First we'll check the `ovarian` dataset, split by treatment group (Figure 8.1).

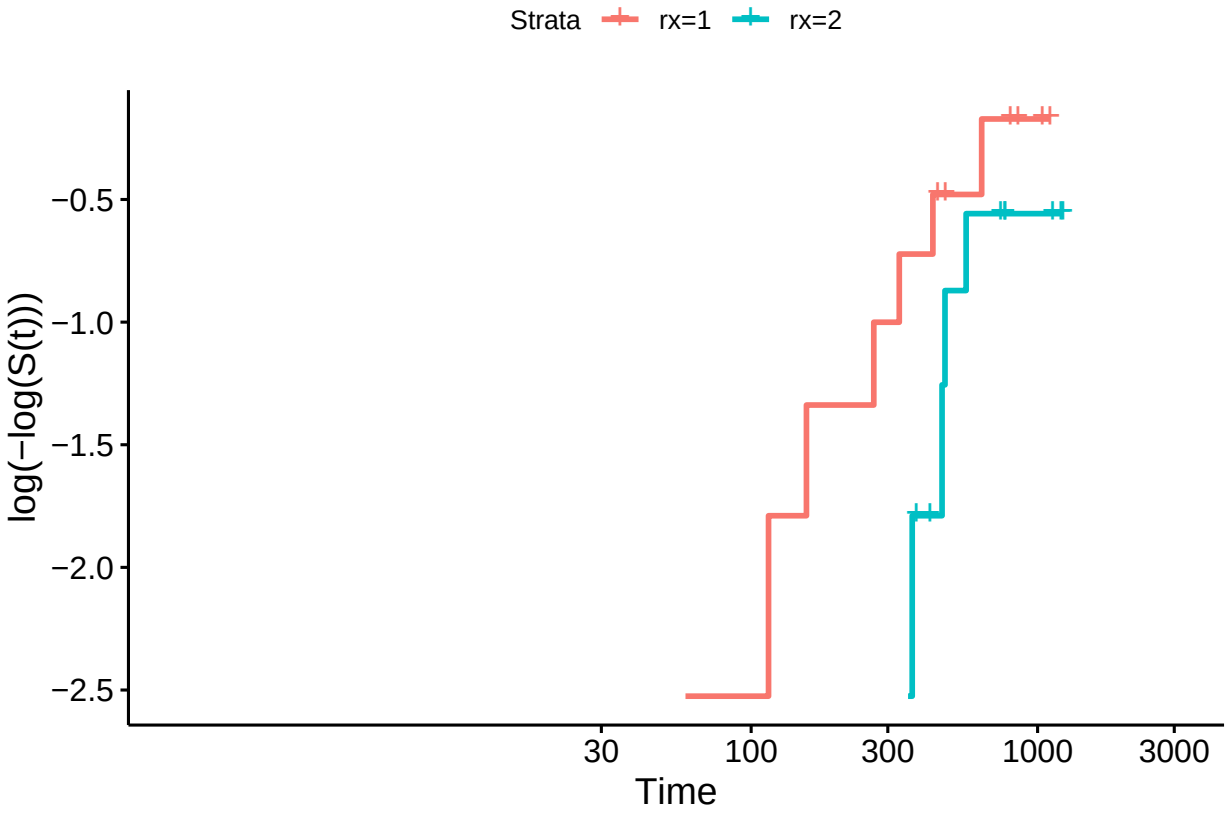


Figure 8.1: Log-log plot for ovarian data by treatment group.

As before, there isn't really enough data to tell whether the assumption is violated.

We can do the same for the `myeloid` data, as in Figures 8.2 and 8.3.

To explore these diagnostic checks further we will introduce the `veteran` dataset, which focuses on a trial for lung cancer patients. The treatment variable is `trt` and there are a mixture of continuous and categorical covariates. We will include `celltype`, a categorical variable with four levels and `karno`, a score from 0 to 100 (which we will treat as continuous).

Firstly we can check the proportional hazards assumption for the categorical covariates, in Figure 8.4.

```
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## trt                  1.2992     0.7697    0.8763    1.9262
## celltypesmallcell    2.2818     0.4382    1.3471    3.8653
## celltypeadeno        3.1708     0.3154    1.7784    5.6534
## celltypelarge        1.4838     0.6739    0.8534    2.5801
## karno                0.9692     1.0318    0.9595    0.9791
##
## Concordance= 0.737  (se = 0.022 )
## Likelihood ratio test= 61.07  on 5 df,   p=7e-12
## Wald test            = 63.41  on 5 df,   p=2e-12
## Score (logrank) test = 66.55  on 5 df,   p=5e-13
```

and check the Schoenfeld residuals for `karno`, as in Figure 8.5.

```
ggcoxzph(cox.zph(cox_vet), var = "karno")
```

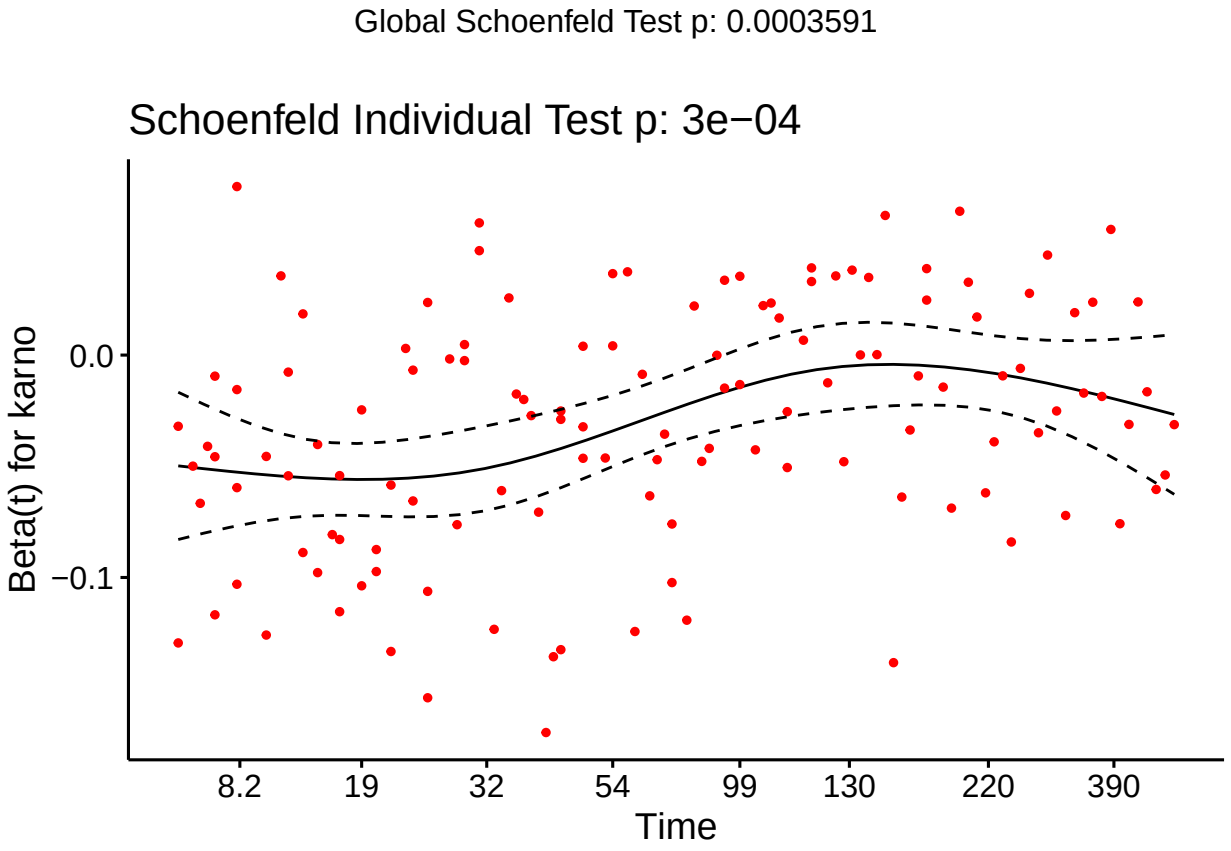


Figure 8.5: Schoenfeld residuals for Cox regression model fit to veteran data, for covariate 'karno'

Although these look quite evenly spread, they are mostly negative, and it appears there is a slight trend with time. Indeed the p-value shows a significant deviation from proportional hazards.