MAS361/61002 Medical Statistics
Part II: Survival Analysis

Module Information

1. Course lecturer
   Kevin Walters in office I22a (k.walters@sheffield.ac.uk).

2. Course materials
   All course materials are on the MAS361/61002 Blackboard (BB) page.

3. Getting help with this module
   Ask questions on BB or book an appointment in one of my office hour slots (linked to on BB).

4. Task Sheets
   There are 5 task sheets on BB, one for each week. Solutions are available on BB.

5. Getting feedback
   You will be given feedback on a set of exercises (separate to the task sheets).

6. Datasets
   These are available on Blackboard as .Rdata files so they will open directly in R.

7. Examinable material
   The material in sections marked with an * is **not** examinable in MAS361 but **is** examinable in MAS61002.

8. Software
   We will use R to illustrate the techniques. Download the datasets on Blackboard into a folder ready to load into R.

9. R code
   The R code to run the examples in the notes is available on BB.

10. Background knowledge
    We assume familiarity with manipulating likelihoods and ideas from linear regression. Sheffield undergraduates may find their MAS223 notes helpful for revision. There is a brief recap of maximum likelihood estimation and likelihood ratio testing (with exercises) at the end of these notes.

Contents

1 Background and Basic Concepts **4**

1.1 Preliminary Discussion 4

1.1.1 Example: Survival of Angina Pectoris 4

1.2 Censoring 5

1.3 Types of Censoring 5

1.4 Informative and Non-Informative Censoring 6

1.5 Analysis Approaches 7

2 Single Sample Models **7**

2.1 Introduction 7

2.1.1 Survivor Function 8

2.1.2 Hazard Function 8

2.1.3 Integrated Hazard Function 8

2.2 Common Failure time Distributions 9

2.2.1 Exponential Distribution 9

2.2.2 Weibull Distribution 9

2.3 Lifetables 9

2.3.1 Example: No Loss to Follow-Up 10

2.3.2 Example: Some Loss to Follow-Up 11

2.4 Kaplan-Meier Product Limit Estimate of $S(t)$ 13

2.4.1 Simple Case: No Censoring 13

2.4.2 Standard Case: Censoring 15

2.4.3 Finding the Median Survival Time from KM plots 15

2.4.4 Example: Tumour Remission Times 16

2.4.5 Calculating Kaplan-Meier Estimates in R 17

2.5 Parametric Models 18

2.5.1 Exponential Failure Times: Uncensored Data 19

2.5.2 Exponential Failure Times: Censored Data 19

2.5.3 Exponential Failure Times: Interest in Other Parameters 21

2.5.4 Exponential Failure Times Example: Lung Cancer Survival 21

2.5.5 $\star$ Exponential Failure Times: Allowing Random Censoring (only examinable in MAS61002) 22

2.5.6 Other Lifetime Distributions 22

2.5.7 R Implementation 23

3 Two Sample Comparisons **27**

3.1 Introduction 27

3.2 Log Rank Test (Non-Parametric) 27

3.2.1 Example: Brain Tumour Survival Times 27

3.2.2 Implementing the Log Rank Test . . . 27
3.2.3 Log Rank Test applied to the Brain Tumour Data . . . 29
3.2.4 Performing the Log Rank Test in R . . . 29
3.3 Parametric Tests . . . 30
3.3.1 MLE Test . . . 30
3.3.2 MLE Test applied to the Brain Tumour Data . . . 31
3.3.3 Likelihood Ratio Test . . . 31
3.3.4 Likelihood Ratio Test applied to the Brain Tumour Data . . . 32
3.3.5 Performing the Likelihood Ratio Test in R . . . 32

4 Regression Models 33
4.1 Accelerated Failure Time (AFT) Models . . . 33
4.1.1 Two-Sample Example . . . 33
4.2 General AFT model Framework for including Explanatory Variables . . . 34
4.2.1 Using this Model . . . 35
4.2.2 An Example — Exponential Model . . . 35
4.3 Fitting an Exponential AFT Model to the Myelogenous Leukemia Data in R . . . 36
4.4 Fitting a Weibull AFT Model to the Myelogenous Leukemia Data in R . . . 38
4.5 Proportional Hazards . . . 39
4.5.1 Which AFT models exhibit the proportional hazards property? . . . 40
4.5.2 An Initial Attempt at Parameter Estimation . . . 40
4.5.3 Partial Likelihood . . . 41
4.5.4 Notes on Partial Likelihood Approach . . . 42
4.5.5 Interpreting an Example: Atrial Fibrillation . . . 43
4.5.6 R Implementation . . . 45
4.5.7 Model Checking . . . 46

A Maximum Likelihood Estimation 49
A.1 Estimation . . . 49
A.2 Definition . . . 49
A.3 Examples . . . 49
A.4 Asymptotic Properties of MLEs . . . 51
A.4.1 Confidence Intervals . . . 52
A.5 [Generalized] Likelihood Ratio Tests . . . 53
A.5.1 Examples . . . 53
A.6 Exercises . . . 54

1 Background and Basic Concepts

1.1 Preliminary Discussion

The objective of a survival data analysis may be to model a single sample of data and describe the lifetimes of a single population or it may be to compare the lifetimes of two or more groups of subjects; for example the two groups may have received different medical treatments and the lengths of survival time measure how effective the treatments are.

A distinctive feature of survival data is that some observations may be censored. Censoring means that the event of interest (e.g. death of patient, failure of component, recovery of patient) has not occurred by the time of recording. In this case, all we know is that the lifetime for that subject is **at least** some value. We cannot just remove censored observations since they carry important information about the effectiveness of the treatment. This complicates the data analysis.

1.1.1 Example: Survival of Angina Pectoris

Data on the survival times of patients with angina pectoris are given by Gehan (1969: J.Chronic Disease) in Table 1. These patients form part of a large group of patients examined at the Mayo Clinic during the 15 year period January 1, 1927 to December 31, 1941.

This table highlights several important factors that are typically present in survival data:

- Grouping — often when people report survival data they group the data into disjoint periods of time, in this case yearly. We do not have the exact timing of the events that have occurred but simply the yearly interval.
- Hidden change in the time-axis — clinical trials often do not recruit all participants at the same time but instead continue to recruit over the entire period of study. Working directly with the date of study entry and exit is cumbersome (and not strictly necessary) so we change the time axis to start recording only once they have entered the study. This is illustrated in Figure 1.
- Censoring — there are *lost* patients i.e. patients for which we do not observe failure (death). These observations are said to be *censored*. Censoring occurs in two ways in this study: some patients are lost while still alive during the course of the study (they may have chosen to no longer take part); some have survived until the study has ended. In both cases we only know that their survival was longer than a certain time.

When analysing these data we should be aware that:

- there are likely to be several possible causes of death;

Survival time (years)	Number of patients known to survive at beginning of interval	Number of patients lost to follow up
0 – 1	2418	0
1 – 2	1962	39
2 – 3	1697	22
3 – 4	1523	23
4 – 5	1329	24
5 – 6	1170	107
6 – 7	938	133
7 – 8	722	102
8 – 9	546	68
9 – 10	427	64
10 – 11	321	45
11 – 12	233	53
12 – 13	146	33
13 – 14	95	27
14 – 15	59	23
15 – 16	30	

Table 1: *Survival times of patients with angina pectoris are from Gehan (1969: J.Chronic Disease)*

- There may be several covariates which might influence an individual's lifetime distributions, e.g. age, sex, ...

1.2 Censoring

Suppose that individual i is observed for a time c_i and their failure time is t_i. We will only observe the failure if $t_i \leq c_i$. This is known as time censoring and is demonstrated in Figure 2

1.3 Types of Censoring

Some forms of censoring are mentioned below.

- **Right Censoring** — the failure time exceeds some value. Failure hadn't occurred when the individual was lost to the study or the study ended;
- **Left Censoring** — the failure time is less than some value. This might occur if subjects are only observed after a fixed period of time. Here, we would know that survival time is less than the period of observation. Another example is when the

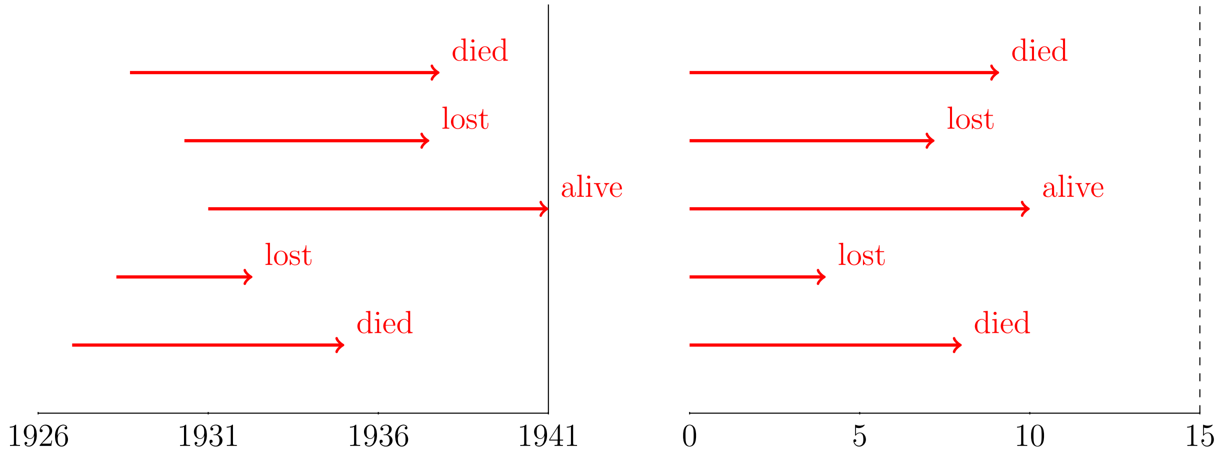


Figure 1: *In the left hand plot we show the progress of individuals through the clinical trial dependent upon true calendar date. It shows the different dates that people entered the trial and the date they exited — either through death, being lost, or being withdrawn alive at the trial end. In the right hand plot we change the x-axis to represent the number of years in the study from entry in order to allow easier comparison between times.*

failure time is the time to recurrence of a tumour, where the presence of such a tumour is only observable during surgery.

- **Interval Censoring** — the failure time is within some range of values. This might occur if individuals are observed at a sequence of fixed appointment times.
- **Type I Censoring** — identical starting points and subjects are observed for a fixed time c_i (typically $c_1 = c_2 = \ldots = c_n$). The number of censored observations is then random.
- **Type II Censoring** — the trial finishes after a certain number of failures. Here we do not specify the end of the trial in terms of time but in terms of the number of failures. This type of censoring is less common in medical studies but is widely used in electronic component testing and reliability studies. Here the number of censored observations is not random but is fixed in advance.

We typically consider Type I censoring.

1.4 Informative and Non-Informative Censoring

For all the techniques presented in this module we will assume that the censoring is random (or non-informative). This means that the censoring time is **statistically independent** of the failure time. This excludes situations whereby people are censored in a clinical trial if they suddenly become seriously ill and, as a consequence, are removed from the trial. Here censoring suggests that an individual is near death. It also excludes situations where they are censored/lost if they become/feel better. Here censoring would suggest that an individual is a long way from death. Examples such as these are known as informative sampling as the act of censoring gives extra information about the survival time.

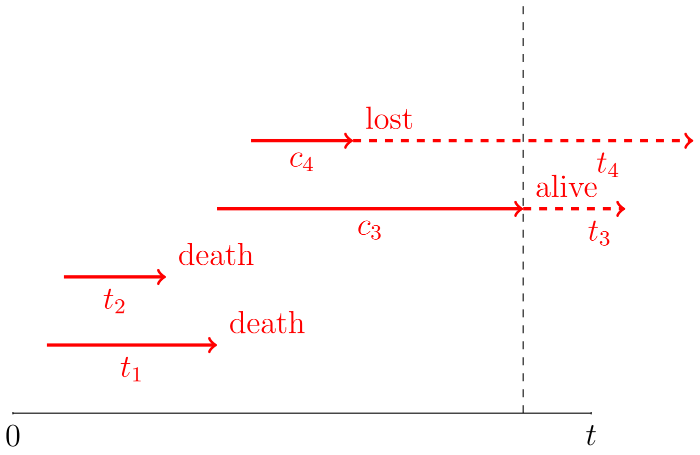


Figure 2: *Individuals enter the study at different times. We observe the failure times of individuals 1 and 2 but not of 3 and 4 who have not failed when we stop observing them at times c_3 and c_4. Individual 3 survives until the end of the study while 4 is lost midway through. All we know is that $t_i > c_i$ for $i = 3, 4$.*

Analysing informatively censored data is very difficult. If such informative censoring is present then the methods described here would often lead to biased estimates. In any survival analysis one should think carefully about whether the censoring is informative.

1.5 Analysis Approaches

Suppose that the time to failure for individual i is T_i. Generally we want to

- estimate lifetime distributions;
- predict survival times – we will consider both non-parametric approaches (lifetables and Kaplan-Meier) and parametric methods (modelling lifetimes via a particular distribution e.g. Exponential, Weibull)

When doing the above we need to allow for censoring, where some T_i are not observed exactly. In this course we will predominantly consider Type I right censoring and assume that censoring is independent of failure time.

2 Single Sample Models

2.1 Introduction

Suppose we have a homogeneous population of individuals, and the failure time of each individual is a continuous random variable $T > 0$ with probability density $f(t)$ and distribution function $F(t)$:

$$F(t) = P(T \leq t) \quad \text{and} \quad f(t) = F'(t) \qquad (\text{so } F(t) = \int_0^t f(u)du). \tag{1}$$

We wish to estimate the distribution of T given, possibly censored, survival data. We first need to define some other functions which are all inter-related.

2.1.1 Survivor Function

Typically we want to work out the probability that an individual survives longer than a certain time. For this reason we generally work with $S(t)$ defined as

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(u)du. \tag{2}$$

Hence

$$f(t) = -S'(t) = \lim_{h\to 0} \frac{P(t < T < t+h)}{h}. \tag{3}$$

2.1.2 Hazard Function

We often model the lifetime through the hazard function, $h(t)$, which measures the risk or proneness to death at time t, given survival up to time t. The hazard function represents the instantaneous death rate for an individual surviving to time t.

$$h(t) = \lim_{h\to 0} \frac{P(t \leq T < t+h | T \geq t)}{h}. \tag{4}$$

By the definition of conditional probability, we find that

$$\begin{aligned} h(t) &= \lim_{h\to 0} \frac{P(t \leq T < t+h, T \geq t)}{P(T \geq t)h} \\ &= \lim_{h\to 0} \frac{P(t < T < t+h)}{P(T \geq t)h} \\ &= \left\{ \lim_{h\to 0} \frac{P(t < T < t+h)}{h} \right\} \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{5}$$

Substituting $f(t) = -S'(t)$, we have

$$\begin{aligned} h(t) &= \frac{-S'(t)}{S(t)} \\ &= -\frac{d}{dt} \log S(t). \end{aligned} \tag{6}$$

2.1.3 Integrated Hazard Function

$$H(t) = \int_0^t h(u)du = -\log S(t). \tag{7}$$

Note furthermore that

$$S(t) = \exp\left\{ -\int_0^t h(u)du \right\} = \exp\{-H(t)\}, \tag{8}$$

and finally that

$$f(t) = h(t)\exp\{-H(t)\}. \tag{9}$$

2.2 Common Failure time Distributions

2.2.1 Exponential Distribution

Suppose that $T_i \sim Exp(\lambda)$, then for $t > 0$

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} && (10) \\ F(t) &= 1 - e^{-\lambda t} && (11) \\ S(t) = 1 - F(t) &= e^{-\lambda t} && (12) \\ h(t) &= \lambda. && (13) \end{aligned}$$

The exponential distribution is the only distribution with a constant hazard function.

2.2.2 Weibull Distribution

Suppose that $T_i \sim Weibull(\lambda, \gamma)$, then

$$\begin{aligned} f(t) &= \lambda\gamma(\lambda t)^{\gamma-1} \exp\left[-(\lambda t)^{\gamma}\right] && (14) \\ S(t) &= \exp\left[-(\lambda t)^{\gamma}\right] && (15) \\ h(t) &= \lambda\gamma(\lambda t)^{\gamma-1} && (16) \end{aligned}$$

The Weibull distribution provides a very flexible family of survival distributions where the value of γ is key:

- $\gamma > 1$ means the hazard is increasing over time;
- $\gamma = 1$ means the hazard is constant (i.e. it's the exponential distribution);
- $\gamma < 1$ means the hazard is decreasing over time.

The parameters can be difficult to estimate, particularly if the true value of γ is close to 1. There are several alternative parameterisations for the Weibull distribution so care is needed to make sure you know what parameterisation is being used in any give situation. The parameterisation above is the one used in the `survreg` function in R (we will use this key function extensively in later Chapters). Confusingly, this is not the parameterisation used in `dweibull` and related R functions.

2.3 Lifetables

Before trying to fit a formal statistical model, an initial non-parametric investigation is sensible — often it provides sufficient information for the study and it will always give useful information to help in selecting a suitable family of distributions. A lifetable is a way of expressing or tabulating the death rates experienced by some particular population during a particular period of time.

2.3.1 Example: No Loss to Follow-Up

The data in Table 2 is from a study where every patient was followed up after a particular treatment, either until death or up to the end of 1992. This is still a form of censoring since we do not know the lifetimes of patients surviving until the end of the study.

Year of treatment	Number treated	Number alive on each anniversary				
		1st	2nd	3rd	4th	5th
1987	62	58	51	46	45	42
1988	39	36	33	31	28	42
1989	47	45	41	38	73	
1990	58	53	48	115		
1991	42	40	173			
	248	232				

Table 2: *Table showing follow up study of patients after a medical treatment*

To find the lifetable we need to estimate

$$P(\text{Survive till end year } i|\text{Alive at start of year } i) = \frac{\text{Number survive in year } i}{\text{Number studied for whole of year } i}.$$

The lifetable is shown in Table 3. Note that ”year i” indicates the number of years after treatment. For example in year 2 (i.e. two years after treatment) the effective number of individuals must not include those who had treatment in 1991 as they exit the study one year after treatment and we ignore what happened to them. Hence

					Lifetable (per 1000)	
Year after treatment	Number obs. year i	Number alive end year i	Prob. survive year i	Prob. die year i	Number alive on each anniversary	Number dying during year i
i			p_i	q_i	l_i	d_i
0					1000	
1	248	232	0.936	0.064	936	64
2	192	173	0.901	0.099	843	93
3	125	115	0.92	0.08	776	67
4	77	73	0.948	0.052	736	40
5	45	42	0.933	0.067	687	49

Table 3: *Lifetable constructed from the data in Table 2*

$$P(\text{Survive until end year 2}|\text{Alive at start of year 2}) = \frac{173}{232-40} = 0.901.$$

The p_i are calculated from the numbers surviving from one year to the next. They are estimates of the conditional probabilities of surviving year i, conditional on surviving up until the start of the year.

2.3.2 Example: Some Loss to Follow-Up

Complications begin to arise when patients are lost to follow-up during the study. With these patients, we do not know if they have died or not and so they are instead considered as *withdrawn.* An example can be seen in Table 4 (taken from Armitage, 1971).

In this example the columns report:

- d_x — the number who died during $(x, x+1)$.
- w_x — the number who disappeared during the period $(x, x+1)$, i.e. withdrawn alive during that year.
- n_x — number living at start of interval $(x, x+1)$. You can accumulate $d_x + w_x$ from the bottom of the table.
- n'_x — we assume the withdrawals are uniformly spread over interval (i.e. the individuals who withdrew did so half way through the year on average) so that

$$n'_x = n_x - \frac{1}{2}w_x. \tag{17}$$

- q_x — conditional probability of dying in $(x, x+1)$,

$$q_x = d_x/n'_x. \tag{18}$$

- $p_x = 1 - q_x$ — the conditional probability of surviving in $(x, x+1)$.
- l_x — life survival rates, assuming start of with 100 individuals. Then $l_0 = 100$ and $l_x = l_0 p_0 p_1 \dots p_{x-1}$. This also allows us to estimate the survivor function since

$$l_x = 100\hat{S}_x \quad x = 0, 1, 2, \dots. \tag{19}$$

- $\hat{f}_{x+1/2}$ — the estimate of the pdf of the failure distribution T,

$$\hat{f}_{x+1/2} = \hat{S}_x - \hat{S}_{x+1} = \hat{S}_x - \hat{S}_x p_x = \hat{S}_x q_x \quad x = 0, 1, 2, \dots. \tag{20}$$

- $\hat{h}_{x+1/2}$ — the estimate of the hazard function of T,

$$\hat{h}_{x+1/2} = \frac{\hat{f}_{x+1/2}}{\hat{S}_{x+1/2}} = \frac{\hat{f}_{x+1/2}}{\frac{\hat{S}_x + \hat{S}_{x+1}}{2}} = \frac{2q_x}{1+p_x}. \tag{21}$$

Notes

1. We have assumed that withdrawals are subjected to the same probability of death as non-withdrawals. This is reasonable if they are *withdrawn alive* but possibly it is not for *lost to follow up* (since the reason they may be *lost* could also affect their chance of dying).

Interval since operation (years)	Last reported during this interval		Living at start of interval	Adjusted number at risk	Estimated probability of death	Estimated probability of survival	% of survivors after x years	Estimate of p.d.f.	Estimate of hazard function
	Died	Withdrawn							
x to $x+1$	d_x	w_x	n_x	n'_x	q_x	p_x	l_x	$\hat{f}_{x+1/2}$	$\hat{h}_{x+1/2}$
0 – 1	90	0	374	374	0.2406	0.7594	100	0.24	0.27
1 – 2	76	0	284	284	0.2676	0.7324	75.9	0.20	0.31
2 – 3	51	0	208	208	0.2452	0.7548	55.6	0.14	0.28
3 – 4	25	12	157	151	0.1656	0.8344	42	0.07	0.18
4 – 5	20	5	120	117.5	0.1702	0.8298	35	0.06	0.19
5 – 6	7	9	95	90.5	0.0773	0.9227	29.1	0.02	0.08
6 – 7	4	9	79	74.5	0.0537	0.9463	26.8	0.01	0.06
7 – 8	1	3	66	64.5	0.0155	0.9845	25.4	0.004	0.02
8 – 9	3	5	62	59.5	0.0504	0.9496	25	0.01	0.05
9 – 10	2	5	54	51.5	0.0388	0.9612	23.7	0.01	0.04
10 –	21	26	47	—	—	—	22.8		

Table 4: *Example of a lifetable where patients are lost to follow-up during the study (taken from Armitage, 1971)*

2. If the w_x withdrawals are evenly spread through the year then this is equivalent to half of these being withdrawn at the beginning of the year and then no further withdrawals. Thus it is appropriate to adjust the number at risk y subtracting $0.5w_x$, noting also that in life tables the intervals are usually quite long (typically one year or more) and so there are usually several withdrawals in that interval.

3. We have assumed that p_x and q_x remain constant over a single year within the study. This is a relatively short period so could be considered a reasonable approximation.

However, this example starts to indicate the difficulty of calculating expectations of life with censored data and so we might wish to start thinking in terms of a parametric model.

2.4 Kaplan-Meier Product Limit Estimate of $S(t)$

The lifetable methods in the previous section all consider the data in groups. If the actual lifetimes (perhaps censored) are available, performing such grouping will lose information. Instead we can form a Kaplan-Meier estimate.

2.4.1 Simple Case: No Censoring

Suppose that we have n observations of lifetimes at $t_1, t_2, \ldots, t_n$. We can order these observed lifetimes so that $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ (if there are k distinct lifetimes). Let d_i be the number of deaths at time $t_{(i)}$ so that

$$\sum_{i=1}^{k} d_i = n. \tag{22}$$

This is illustrated in Figure 3 below,

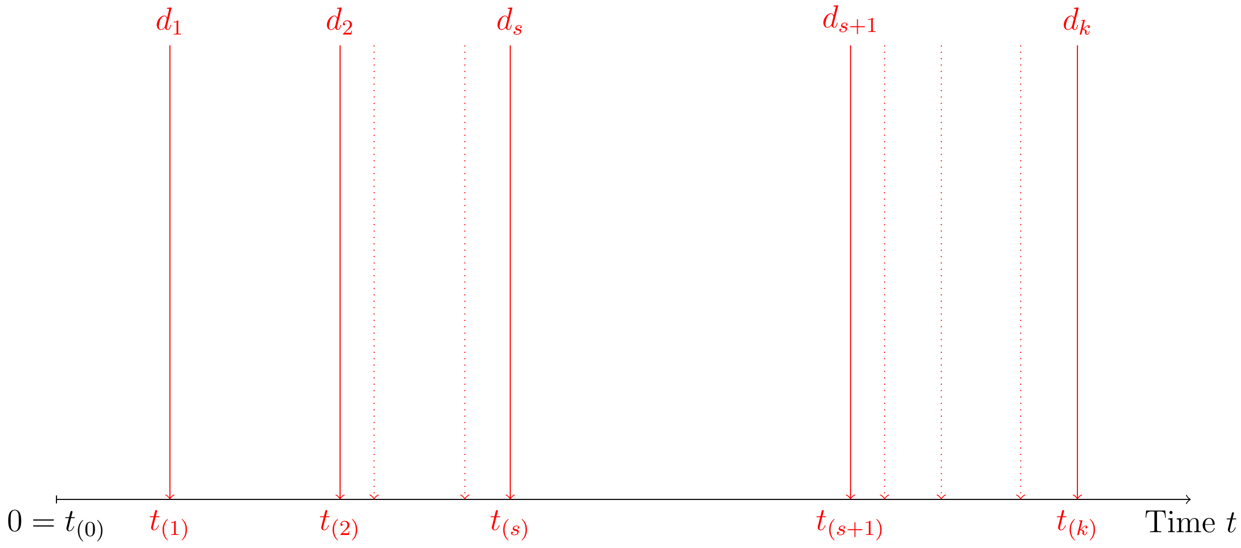


Figure 3: *Illustration of possible failure times with no censoring*

Now we can estimate

$$\hat{F}(t) = P(T \leq t) = \{\text{proportion of lifetimes} \leq t\} \tag{23}$$

$$= \frac{1}{n}\sum_{i=1}^{s} d_i \qquad \text{for } t_{(s)} \leq t < t_{(s+1)} \tag{24}$$

and

$$\hat{S}(t) = 1 - \hat{F}(t) = P(T > t) \tag{25}$$

$$= \frac{n - \sum_{i=1}^{s} d_i}{n} \qquad \text{for } t_{(s)} \leq t < t_{(s+1)}. \tag{26}$$

We can plot $\hat{S}(t)$ as shown in Figure 4. If we let r_j be the number at risk (the number alive) just before time $t_{(j)}$, then $r_{j+1} = r_j - d_j$. Hence we can re-write

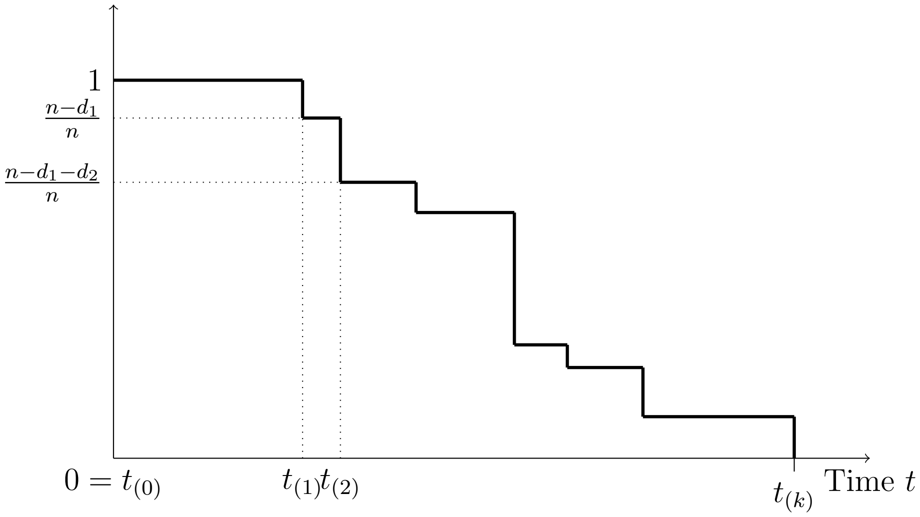


Figure 4: *Showing the step function nature of a typical Kaplan-Meier estimate of the survivor function*

$$\hat{S}(t) = \frac{n - \sum_{i=1}^{s} d_i}{n} \tag{27}$$

as

$$\hat{S}(t) = \left(\frac{n-d_1}{n}\right)\left(\frac{n-d_1-d_2}{n-d_1}\right)\left(\frac{n-d_1-d_2-d_3}{n-d_1-d_2}\right)\ldots\left(\frac{n-d_1-\ldots-d_s}{n-d_1-\ldots-d_{s-1}}\right) \tag{28}$$

$$= \left(1-\frac{d_1}{r_1}\right)\left(1-\frac{d_2}{r_2}\right)\ldots\left(1-\frac{d_s}{r_s}\right) \tag{29}$$

$$= \prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \qquad \text{for } t_{(s)} \leq t < t_{(s+1)} \text{ and } s \geq 1. \tag{30}$$

Note that $\hat{S}(t) = 1$ for $t < t_{(1)}$. For convenience we drop explicitly stating the $s \geq 1$ condition and assume that this is implied (the definition does not make sense for $s = 0$ since in this case the upper limit is smaller than the lower limit).

2.4.2 Standard Case: Censoring

We can use the same type of estimate based upon a product when there is censoring too. As before, let $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ be the k distinct lifetimes with d_j individuals failing at time $t_{(j)}$. However suppose also that I_j denotes the number of individuals who were censored in the interval $t_{(j-1)} \leq t < t_{(j)}$ as shown in Figure 5.

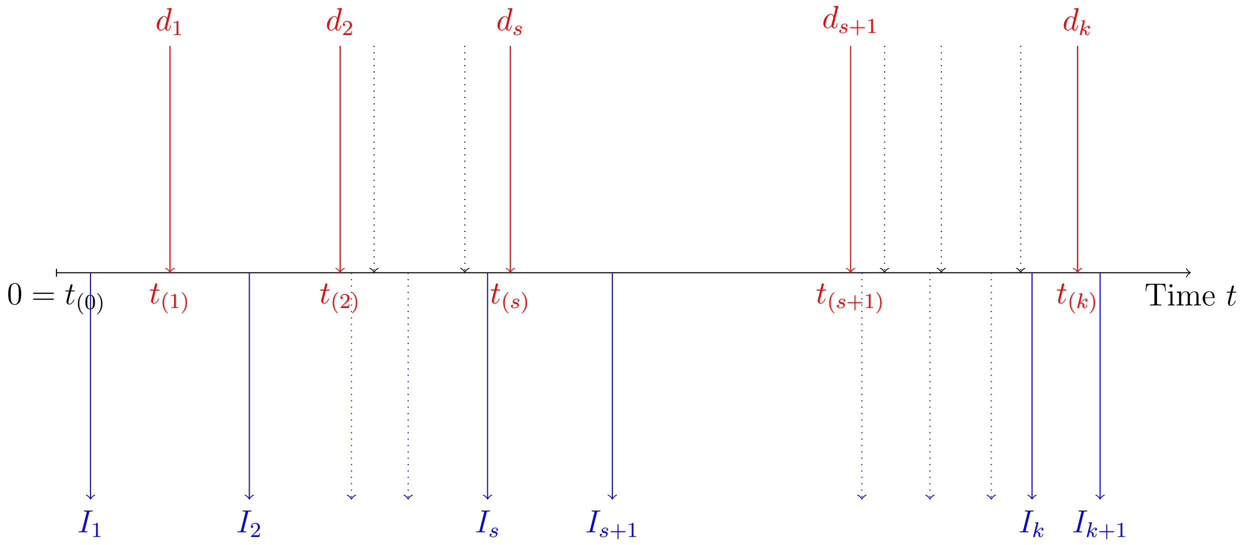


Figure 5: *Illustration of possible failure times with censoring*

In this case, when working out the number at risk just before time t we need to be aware of those individuals who have been censored and so

$$r_1 = n - I_1 \tag{31}$$

$$r_j = r_{j-1} - d_{j-1} - I_j \tag{32}$$

$$= n - (d_1 + d_2 + \ldots + d_{j-1}) - (I_1 + I_2 + \ldots + I_j) \quad \text{for } j \geq 2 \tag{33}$$

and so we can find the same Kaplan-Meier product limit

$$\hat{S}(t) = \prod_{j=1}^{s} \left(1 - \frac{d_j}{r_j}\right) \quad \text{for } t_{(s)} \leq t < t_{(s+1)} \tag{34}$$

2.4.3 Finding the Median Survival Time from KM plots

The median, M, is defined as

$$M = \underset{t}{\operatorname{argmin}}\{S(t) \leq 0.5\}$$

I.e. it is the smallest value of t where the survivor function takes a value of 0.5 or less.

Notes

- Assumes that the I_j censorings survive up to $t_{(j-1)}$ and then are removed immediately. This is slightly different from the adjustment usually used in life tables. Kaplan-Meier estimates are usually used when the intervals between events is quite short and the number of withdrawals in any interval is thus quite small.

- The uncensored case is just a special case with $I_j = 0 \quad \forall j$.
- If $I_{k+1} > 0$ then
$$\hat{S}(t) = \prod_{j=1}^{s}\left(1 - \frac{d_j}{r_j}\right) > 0 \tag{35}$$
since $r_k > d_k$. Hence if there are still individuals left in the trial after the last observed death (possibly as the observation period has ended) then $\hat{S}(t)$ will not tend to 0. However we know, by definition that as $t \to \infty$ then $S(t) \to 0$ since everyone will fail eventually, so the Kaplan-Meier estimate is biased if the maximum observation is censored.
- $\hat{S}(t)$ is subject to sampling error. Greenwood gives
$$var\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{j=1}^{s} \frac{d_j}{r_j(r_j - d_j)} \quad \text{for } t_{(s)} \leq t < t_{(s+1)} \tag{36}$$
- We can also estimate $H(t)$ by $\hat{H}(t) = -\log \hat{S}(t)$ or we can use the simpler approximation
$$\tilde{H}(t) = \sum_{j=1}^{s} \frac{d_j}{r_j} \quad \text{for } t_{(s)} \leq t < t_{(s+1)} \tag{37}$$

2.4.4 Example: Tumour Remission Times

A study (data available on Blackboard as `tumour.Rdata`) investigates the remission times for 10 patients with tumours. During the study, 6 relapse after 3.0, 6.5, 6.5, 10, 12, and 15 months and one was lost to follow-up at 8.4 months. The other three were still in remission at end of study after 4.0, 5.7, 10.1 months respectively. We show this data in Table 5 along with the Kaplan-Meier estimate of the survivor function. A plot of this Kaplan-Meier estimate is given in Figure 6.

j	$t_{(j)}$	I_j	r_j	d_j	$\left(1 - \frac{d_j}{r_j}\right)$	$\hat{S}(t)$	t	Calculation of $\hat{S}(t)$
0	0	0				1.00	$0 \leq t < 3.0$	
1	3.00	0	10	1	9/10	0.90	$3.0 \leq t < 6.5$	9/10
2	6.50	2	7	2	5/7	0.64	$6.5 \leq t < 10.0$	9/10 × 5/7
3	10.00	1	4	1	3/4	0.48	$10.0 \leq t < 12.0$	9/10 × 5/7 × 3/4
4	12.00	1	2	1	1/2	0.24	$12.0 \leq t < 15.0$	9/10 × 5/7 × 3/4 × 1/2
5	15.00	0	1	1	0	0.00	$15 \leq t$	

Table 5: *Kaplan-Meier Estimate of the Survivor Function for the Tumour Remission Times Data*

We can see from Table 5 that $\hat{S}(t) > 0.5$ for $0 \leq t < 10$ and $\hat{S}(10) \leq 0.5$ so that the median tumour remission time is 10 months.

2.4.5 Calculating Kaplan-Meier Estimates in R

- R functions for analysing survival data are in the survival package (load it by ticking the box next to the package in RStudio). You shouldn't need to install it.
- The first step is to create a *survival object* with the function Surv() (note the capital S). It contains information on which observations are censored (coded with a 0) and which are observed events (coded with a 1).
- The next step is to estimate the survivor curve with the function survfit().
- To produce flexible Kaplan-Meier plots use the ggsurvplot function in the survminer package (install and load it as above). The function summary() will give the estimate of the survivor function and other details of the model fitting.

After downloading the dataset tumour.Rdata and saving it in your working directory, analysis can be performed as described below. This will produce the plot shown in Figure 6.

```
> load("tumour.Rdata")
> tumour # 0 means censored, 1 means observed
   time censor
1   3.0      1
2   4.0      0
3   5.7      0
4   6.5      1
5   6.5      1
6   8.4      0
7  10.0      1
8  10.1      0
9  12.0      1
10 15.0      1
> tumour_sv <- Surv(tumour$time, tumour$censor, type = "right")
> tumourSurv <-survfit(tumour_sv ~ 1, data=tumour)
> summary(tumourSurv)
Call: survfit(formula = tumour_sv ~ 1, data = tumour)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3.0     10       1    0.900  0.0949       0.7320            1
  6.5      7       2    0.643  0.1679       0.3852            1
 10.0      4       1    0.482  0.1877       0.2248            1
 12.0      2       1    0.241  0.1946       0.0496            1
 15.0      1       1    0.000     NaN           NA           NA
> ggsurvplot(tumourSurv, data = tumour_sv, surv.median.line = 'hv',
+             conf.int = TRUE, conf.int.style = "ribbon",
+             conf.int.alpha = 0.4, xlab = "Time (months)",
+             risk.table = T, break.time.by = 1,
+             ncensor.plot = T, color = "black", censor = T,
+             censor.shape = 3, censor.size = 5, tables.height = 0.25)
```

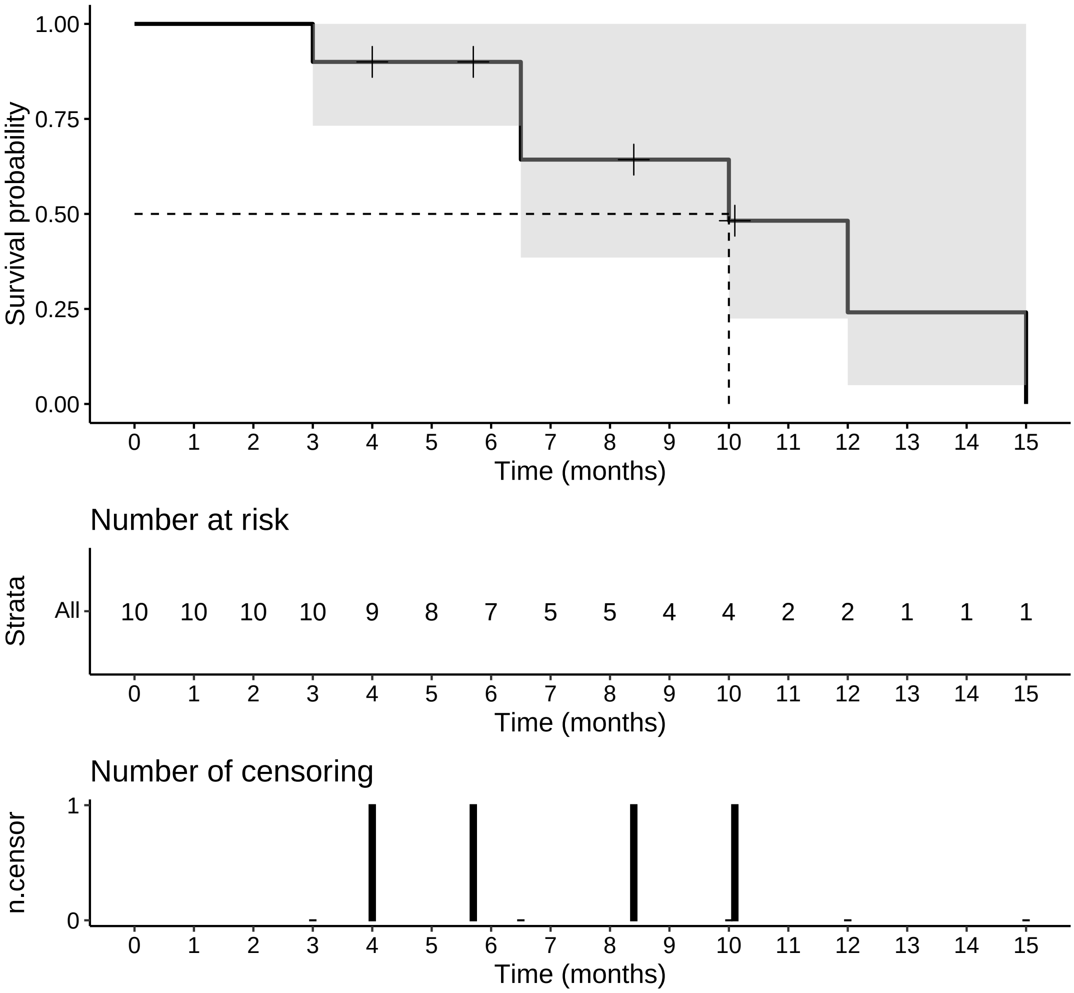


Figure 6: *Kaplan-Meier Estimate (plus 95% confidence interval) of the survivor function for the Tumour Remission Times data*

2.5 Parametric Models

Suppose we have a non-negative failure time T with p.d.f. $f(t)$; distribution function $F(t)$; survival function $S(t) = 1 - F(t)$; and hazard function $h(t)$. Typically the pdf depends on an unknown parameter θ that needs to be estimated from the data. There are many methods of estimation but we concentrate on maximum likelihood estimation (m.l.e.) and then perform tests/inference relying on its nice asymptotic (large sample) properties. Some details of likelihoods, maximum likelihood estimation and likelihood ratio tests are given in Appendix A.

To find the mle, we first find the likelihood $L(\theta)$ of a parameter θ for data $x_1, \ldots, x_n$. We can now maximize $L(\theta)$ wrt θ to find the maximum likelihood estimate of θ. To do this we usually maximise the log-likelihood as this is generally simpler to work with.

[For small samples we might need to look at alternative methods, e.g. Bayesian methods such as in Smith & Naylor (1987) Applied Statistics]

2.5.1 Exponential Failure Times: Uncensored Data

Refer to Equations (10) to (13) on page 9 in this section. Suppose we observe (non-censored) failure times $t_1, t_2, \ldots, t_n$, Then

$$L(\lambda; t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i) = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i} \tag{38}$$

$$\log(L) = l(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} t_i \tag{39}$$

$$\Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} \tag{40}$$

We can also find a confidence interval for λ since $Y = \sum_{i=1}^{n} T_i \sim \Gamma(n, \lambda)$ and

$$Z = 2\lambda Y \sim \chi^2_{2n}. \tag{41}$$

Hence $P(\chi^2_{2n;\alpha/2} < 2\lambda \sum_{i=1}^{n} T_i < \chi^2_{2n;1-\alpha/2}) = 1 - \alpha$ and so a $100(1-\alpha)\%$ confidence interval for λ is given by

$$\left(\frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^{n} t_i}, \frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^{n} t_i} \right) \tag{42}$$

Finally we can find the mle of $S(t)$ as

$$\hat{S}(t) = e^{-\hat{\lambda} t} \tag{43}$$

2.5.2 Exponential Failure Times: Censored Data

Now suppose that we have right censored data. Specifically, let us consider n patients, with potential i.i.d. lifetimes $T_i \sim Exp(\lambda)$. We observe either the lifetime t_i or the fact that $t_i > c_i$ for individual i. The simplest case is to assume that the c_i are fixed and known for all individuals, i.e. non-random. For individuals where we observe the failure time (where $t_i \leq c_i$), the contribution of each individual to the likelihood is

$$f(t_i) = \lambda e^{-\lambda t_i}. \tag{44}$$

For individuals where we observe the right-censored value (where $t_i > c_i$), using the fact that $P(T_i > c_i) = S(c_i)$, the contribution to the likelihood is

$$S(c_i) = e^{-\lambda c_i}. \tag{45}$$

If we define

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq c_i \text{ (i.e. uncensored)} \\ 0 & \text{if } t_i > c_i \text{ (i.e. censored)} \end{cases} \tag{46}$$

then

$$L(\lambda) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(c_i)]^{1-\delta_i} = \prod_{i=1}^{n} \left[\lambda e^{-\lambda t_i}\right]^{\delta_i} \left[e^{-\lambda c_i}\right]^{1-\delta_i}. \tag{47}$$

It follows that

$$l(\lambda) = \log \lambda \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} t_i \delta_i - \lambda \sum_{i=1}^{n} (1-\delta_i) c_i \tag{48}$$

$$\frac{\partial l}{\partial \lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} (t_i \delta_i + (1-\delta_i) c_i) \tag{49}$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} (t_i \delta_i + (1-\delta_i) c_i)}. \tag{50}$$

We can use the asymptotic properties of maximum likelihood estimators given in Appendix A.4 to find confidence intervals for $\hat{\lambda}$. We have

$$\hat{\lambda} \xrightarrow{d} N\left(\lambda, I^{-1}\right), \tag{51}$$

where $I^{-1} = var(\hat{\lambda})$ is the variance-covariance matrix of $\hat{\lambda}$. If we use the Fisher information then I is given by

$$I = E\left[-\frac{\partial^2 l}{\partial \lambda^2}\right]. \tag{52}$$

Now, in our case

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum \delta_i}{\lambda^2} \tag{53}$$

and (implicitly assuming that the c_i are considered non-random)

$$E[\delta_i] = 1P(T_i \le c_i) + 0P(T_i > c_i) \tag{54}$$

$$= (1 - e^{-\lambda c_i}). \tag{55}$$

Replacing λ with $\hat{\lambda}$ we have that

$$var(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_{i=1}^{n} (1 - e^{-\hat{\lambda} c_i})}. \tag{56}$$

Alternatively, we can use the observed information to specify I. In this case we have

$$I = -\left.\frac{\partial^2 l}{\partial \lambda^2}\right|_{\lambda = \hat{\lambda}} \tag{57}$$

giving

$$var(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_{i=1}^{n} \delta_i}. \tag{58}$$

To find a $100(1-\alpha)\%$ confidence interval for λ we use the fact that the distribution of $\hat{\lambda}$ is normal giving us the interval

$$\hat{\lambda} \pm 1.96 \times \sqrt{var(\hat{\lambda})} \tag{59}$$

2.5.3 Exponential Failure Times: Interest in Other Parameters

Our interest may be in other aspects, e.g. the mean lifetime $\mu = E[T] = \lambda^{-1}$; or the age S_α beyond which $100\alpha\%$ survive. To estimate these we use the result that for any differentiable, monotonic function $g(\lambda)$, we have

$$var(g(\hat{\lambda})) \approx \{g'(\lambda)\}^2 \mid_{\lambda=\hat{\lambda}} var(\hat{\lambda}). \tag{60}$$

Suppose we want to find the variance of the m.l.e of the mean lifetime $var(\hat{\mu})$. Since $\hat{\mu} = 1/\hat{\lambda}$ we have $g(\lambda) = 1/\lambda$. Using the fisher information we get that

$$var(\hat{\mu}) \approx \frac{\hat{\mu}^2}{\sum_{i=1}^{n}(1 - e^{-\hat{\lambda}c_i})}. \tag{61}$$

If using the observed information we replace $\sum_{i=1}^{n}(1 - e^{-\hat{\lambda}c_i})$ with $\sum_{i=1}^{n}\delta_i$. Suppose we want to find the variance of the estimator of the time beyond which $100\alpha\%$ survive, $var(\hat{S}_\alpha)$. We have $\alpha = P(T \geq S_\alpha) = S(S_\alpha) = e^{-\lambda S_\alpha}$ and so

$$S_\alpha = -\lambda^{-1}\log\alpha \tag{62}$$

$$\Rightarrow \hat{S}_\alpha = -\hat{\lambda}^{-1}\log\alpha \tag{63}$$

and to find the variance,

$$var(\hat{S}_\alpha) = var(-\hat{\lambda}^{-1}\log\alpha) \tag{64}$$

$$= [-\log\alpha]^2\, var(\hat{\lambda}^{-1}). \tag{65}$$

2.5.4 Exponential Failure Times Example: Lung Cancer Survival

The survival times (in days) of 10 patients with advanced lung cancer are given in data set `lcancer.Rdata` and shown in Table 6. The study terminated after 90 days and any patient still alive at that point was no longer observed. Patients joined the study at different times. There were $\sum\delta_i = 7$ deaths during the study with

Patient.no.	1	2	3	4	5	6	7	8	9	10
Entry time	9	18	20	30	49	59	59	60	61	69
Max. possible c_i	81	72	70	60	41	31	31	30	29	21
Survival time t_i	2	> 72	51	> 60	33	27	14	24	4	> 21
δ_i	1	0	1	0	1	1	1	1	1	0

Table 6: *90 day study of 10 patients with lung cancer*

$$\sum\delta_i t_i = 155 \quad \text{and} \quad \sum(1-\delta_i)c_i = 153. \tag{66}$$

Thus our estimate of the expected number of deaths per day $(\hat{\lambda})$ is 0.0227 and the m.l.e. of the mean lifetime is $\hat{\mu} = 1/\hat{\lambda} = 44.0$ days. We can find a 95% confidence interval for $\hat{\lambda}$ using Equation (59) giving (0.00586, 0.0395).

2.5.5 ⋆ Exponential Failure Times: Allowing Random Censoring (only examinable in MAS61002)

A full treatment of this topic is beyond the scope of this module but we'll consider the simplest approaches and state some results. The simplest case is the where observations are collected over a fixed time interval of length T_{max}. To introduce the randomness of the censoring in a simple way suppose subjects arrive uniformly over the interval $(0, T_{max})$ and that there is no other loss to follow up, so the censoring is caused only because the study ends before the subject experiences the event.

Let δ_i be the censoring indicator. It can be shown (you should verify it) that

$$E[\delta_i] = 1 - \frac{1 - e^{-\lambda T_{max}}}{\lambda T_{max}} \tag{67}$$

and so analogously to the previous section

$$s.e.(\hat{\lambda}) \approx \frac{\hat{\lambda}}{\sqrt{n\left(1 - \frac{1-e^{-\lambda T_{max}}}{\lambda T_{max}}\right)}}. \tag{68}$$

If recruitment is over a shorter period R but the study lasts for a time T_{max} so the arrival times are uniformly spread over $(0, R)$ then we have

$$E[\delta_i] = 1 - \frac{1 - e^{-\lambda(T_{max}-R)} - e^{-\lambda T_{max}}}{\lambda R} \tag{69}$$

with obvious modification to the standard error.

2.5.6 Other Lifetime Distributions

Weibull

We may wish to model the lifetime distribution as a Weibull distribution, see Section 2.2.2 on page 9 for more details. This allows a flexible range of hazard functions.

Lognormal

In such instances we model $log(T) \sim N(\mu, \sigma^2)$ so

$$f(t) = (2\pi)^{-\frac{1}{2}}(\sigma t)^{-1} \exp\left\{\frac{-\frac{1}{2}(\log t - \mu)^2}{\sigma^2}\right\}, \tag{70}$$

$$F(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{71}$$

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \tag{72}$$

and $h(t) = f(t)/S(t)$ requires numerical evaluation.

If modelling the lifetime distribution as lognormal, one should be aware that estimation can be unstable if there are short lifetimes.

Others

Sometimes we use other distributions to model the lifetime, for example the Gamma or Gumbel distribution or a mixture distribution. Most of these require numerical evaluation of the hazard function.

2.5.7 R Implementation

The basic function to perform estimation using these models is `survreg()` and one of the arguments specifies which distribution to use from the options `weibull`, `exponential`, `gaussian`, `logistic`, `lognormal` and `loglogistic`. The default is `weibull`. It is also possible to use other distributions if the distribution function and density function are specified (see the help system if needed).

Care needs to be taken with the parameterization in `survreg()` since it models the parameters in what may not initially seem the most intuitive way. It generally presents value for log of the parameter and these will need interpreting/transforming back into the actual survival times themselves. The reason for this is that `survreg` can be adapted to more complicated models (as we will see in Chapter 4.1) where this parameterisation is more natural.

- For the exponential model with rate parameter λ (or mean λ^{-1}), the MLE is $\hat{\lambda} = \exp(-\beta_0)$, where β_0 is the `survreg()` intercept. It follows that β_0 represents the log(mean survival time). It is easiest to calculate a CI for $-\beta_0$, and then exponentiate the CI limits (base e) to get a confidence interval for the true value of λ.

- For the Weibull model, the MLE of λ is $\hat{\lambda} = \exp(-\beta_0)$ as for the exponential case. $\hat{\gamma}$ is the reciprocal of the 'scale' parameter provided directly in the `survreg()` summary.

Illustration using lung cancer survival times — Exponential distribution

We illustrate fitting a distribution to the lung cancer survival times given in `lcancer.Rdata`. As with calculation of the non-parametric Kaplan-Meier estimate, we need to use the `Surv()` function to create a survival object which contains the information on censoring. This object can then be used in fitting any of the available survival regression models. The first step is to load the data frame. We can plot the KM estimate of the survivor function to help us decide upon a suitable parametric family

```
> load("lcancer.Rdata")
> sum(lcancer$censor) # number non-censored observations
[1] 7
> lcancer_sv <- Surv(lcancer$time, lcancer$censor, type = "right")
> lcancerSurv <- survfit(lcancer_sv ~ 1)
```

```
ggsurvplot(lcancerSurv, data = lcancer_sv, surv.median.line='hv',
           conf.int = TRUE, conf.int.style = "ribbon",
           conf.int.alpha = 0.4, xlab = "Time (months)",
           risk.table = T, break.time.by = 10,
           ncensor.plot = TRUE, color = "black", censor = T,
           censor.shape = 3, censor.size = 5, tables.height = 0.25)
```

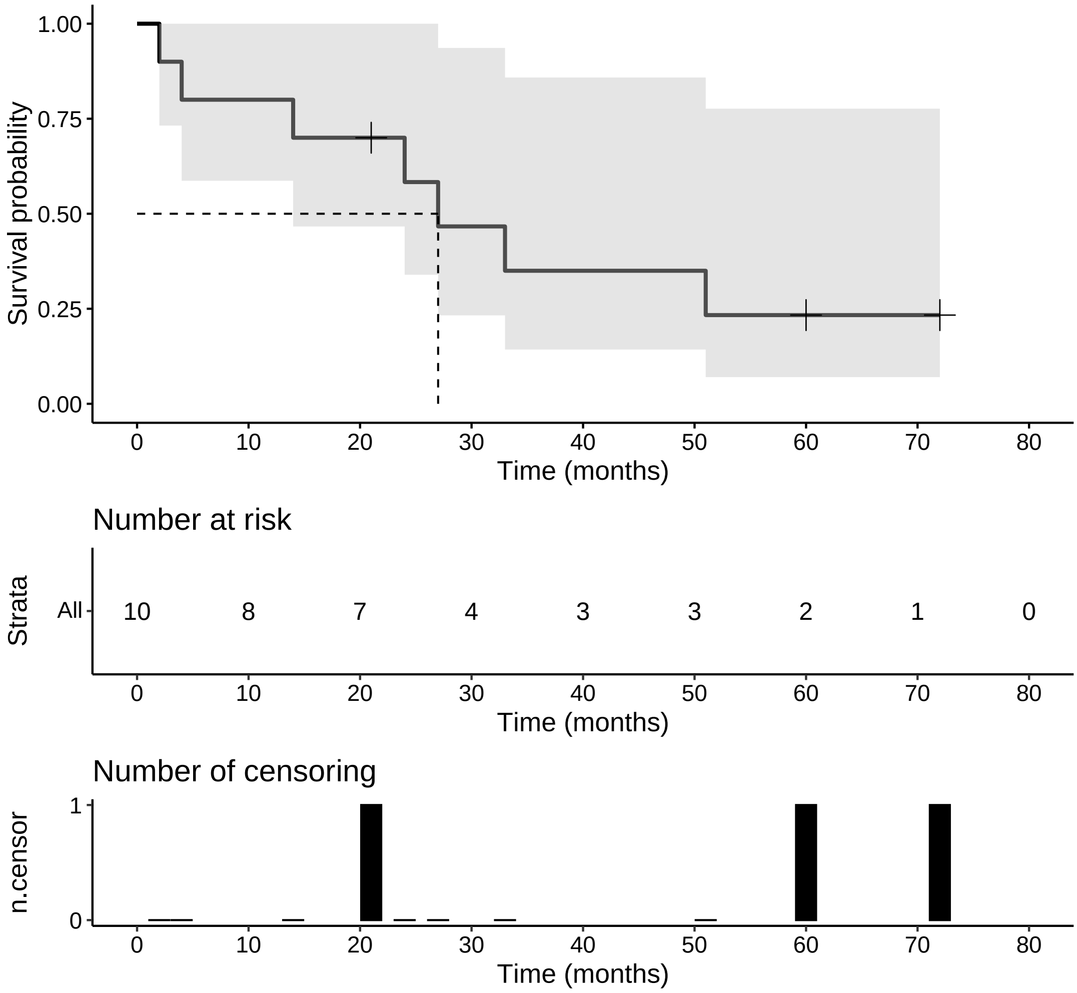


Figure 7: *Kaplan-Meier Estimate (plus 95% confidence interval) of the survivor function for the Lung Cancer data*

This gives us the KM estimate in Figure 7 which could plausibly represent exponential decay (the survivor function if $T \sim Exp(\lambda)$. We therefore proceed by fitting this model:

```
> # Lung cancer exponential MLE estimates
> lcancer_regexp <- survreg(lcancer_sv ~ 1, dist = "exponential")
> summary(lcancer_regexp)

Call:
survreg(formula = lcancer_sv ~ 1, dist = "exponential")
            Value Std. Error  z      p
```

```
(Intercept) 3.784      0.378 10 <2e-16

Scale fixed at 1

Exponential distribution
Loglik(model)= -33.5   Loglik(intercept only)= -33.5
Number of Newton-Raphson Iterations: 4
n= 10
```

Thus

$$\hat{\lambda} = \frac{1}{\exp(\text{intercept})} = \frac{1}{\exp(3.78)} = 0.0228 \tag{73}$$

and a 95% confidence interval for λ is

$$\frac{1}{\exp(3.78 \pm 1.96 \times 0.378)} = (0.0108, 0.0479) \tag{74}$$

The standard errors in the R output are not the same as the standard errors calculated from Equation (58). That would give an approximate standard error of $\hat{\lambda}$ as

$$\frac{1}{\sqrt{7}\exp(3.78)} = 0.008626 \qquad \left[s.e.(\hat{\lambda}) \approx \frac{\hat{\lambda}}{\sqrt{\sum \delta_i}} \right] \tag{75}$$

and a confidence interval of (0.00589, 0.0397). These illustrate that calculations of standard errors and confidence intervals depend on the approach being used although the differences may often not be very large.

Illustration using tumour remission times — Weibull and Exponential distribution

The default distribution for survreg is the Weibull distribution so you don't need to specify a distribution argument to fit a Weibull distribution.

```
> load("tumour.Rdata")
> tumour_sv <- Surv(tumour$time, tumour$censor, type = "right")
> tumourSurvWeib <- survreg(tumour_sv ~ 1, data = tumour)
> summary(tumourSurvWeib)

Call:
survreg(formula = tumour.svweib, data = tumour)
            Value Std. Error     z        p
(Intercept)  2.42      0.147 16.41 1.63e-60
Log(scale)  -1.02      0.312 -3.26 1.12e-03

Scale= 0.361
```

```
Weibull distribution
Loglik(model)= -18.3   Loglik(intercept only)= -18.3
Number of Newton-Raphson Iterations: 6
n= 10
```

From the R output we see that $\hat{\lambda} = \exp(-2.42) = 0.089$ and $\hat{\gamma} = 1/0.361 = 2.8$ and so our estimated distribution for the remission time is $T \sim Weibull(0.089, 2.8)$ The survival object tumour.sv can be used for fitting another model, for example the exponential:

```
> tumourSurvExp <- survreg(tumour_sv ~ 1, data = tumour, dist = "exponential")
> summary(tumourSurvExp)



Call:
survreg(formula = tumour.sv, data = tumour, dist = "exponential")
            Value Std. Error    z        p
(Intercept)  2.61      0.408 6.38 1.76e-10

Scale fixed at 1

Exponential distribution
Loglik(model)= -21.6   Loglik(intercept only)= -21.6
Number of Newton-Raphson Iterations: 4
n= 10
```

The γ parameter of the Weibull distribution is usually the most difficult to estimate, especially if the true value is actually close to 1. Fixing it to 1 reduces the Weibull to an exponential model.

3 Two Sample Comparisons

3.1 Introduction

In our work so far, we have only considered data from a single sample. However, a common problem is the comparison of two (or more) survival distributions, e.g. we might wish to find out which of two treatments is the better or whether the pattern of survivals/deaths for Males is different from that for Females.

One simple comparison would be to plot the Kaplan-Meier estimates for each group. The question is then whether there is a statistically significant difference between the two curves.

3.2 Log Rank Test (Non-Parametric)

3.2.1 Example: Brain Tumour Survival Times

Twelve brain tumour patients were randomized and received either radiation, or radiation and chemotherapy. One year after the start of the study the survival times in weeks are shown in Table 7:

Group 1 RT:	10	26	28	30	41	12*
Group 2 RT+CT:	24	30	42	15*	40*	42*

Table 7: *Survival times of 12 patients undergoing either radiation (RT) or radiation and chemotherapy (RT+CT). A * denotes a censored observation*

We can plot the Kaplan-Meier survivor function estimates (see Figure 8) but we want to **formally** test if there is a difference in the survival functions for the two types of treatment.

3.2.2 Implementing the Log Rank Test

If $S_1(t)$ and $S_2(t)$ are the survival functions under the two conditions, we wish to test

$$\begin{aligned} H_0 :& \quad S_1(t) = S_2(t) \\ H_1 :& \quad S_1(t) \neq S_2(t) \quad \text{for some } t \end{aligned} \tag{76}$$

To perform the test, we order the times of death for the two groups **combined** i.e.

$$t_{(1)} < t_{(2)} < \dots$$

We then find the expected number of deaths in each group. Under H_0 where there is no difference in groups, at the occurrence of the first death at time $t = 10$, there are 12 individuals at risk and one death. Six of those individuals are in group 1 and six are in

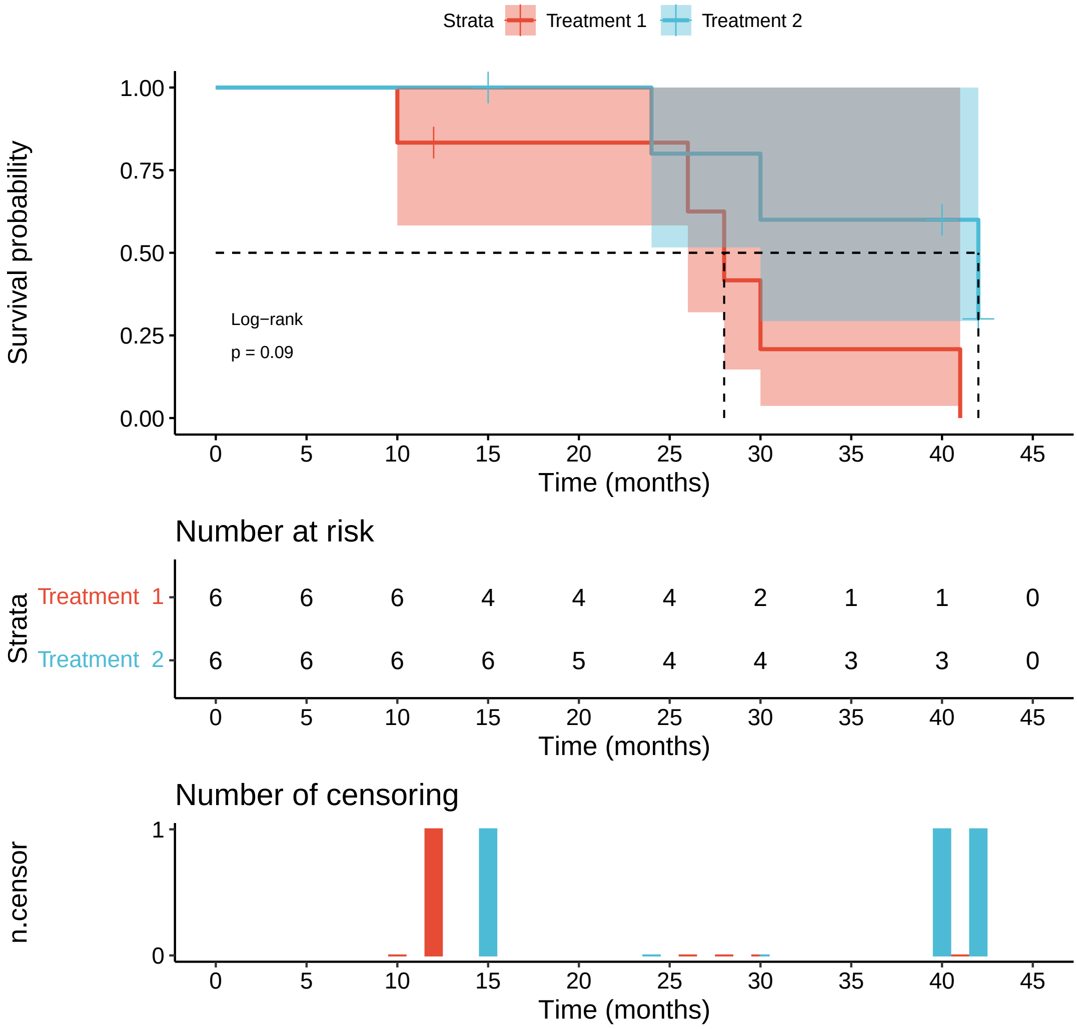


Figure 8: *Kaplan-Meier Estimates of the survivor functions for two brain tumour treatments*

group 2. Hence, under H_0 we would expect $1 \times \frac{6}{12}$ of the deaths at time $t_{(1)}$ to be in group 1 and $1 \times \frac{6}{12}$ of them to be in group 2.

The next death occurs at $t_{(2)} = 24$ when there are four individuals in group 1 and five in group 2. Hence we would expect $1 \times \frac{4}{9}$ deaths in group 1 and $1 \times \frac{5}{9}$ in group 2. We can continue to do this until all the deaths have occurred.

Having done this for every time $t_{(i)}$ we can sum the number of expected deaths in each group to find E_i, the expected total number of deaths in group i. We can compare these with the total observed number of deaths O_i in the groups and find the log-rank statistic

$$LR = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi^2_1 \tag{77}$$

under H_0.

3.2.3 Log Rank Test applied to the Brain Tumour Data

We perform this test on our Brain Tumour data as shown in Table 8. Here

$$LR = 2.46 < \chi^2_{1,0.95}$$

i.e. there is no significant difference in survivor functions at the 5% level.

		Number at risk			Number of deaths			Expected no. of deaths	
i	t_i	r_{1i}	r_{2i}	r_i	d_{1i}	d_{2i}	d_i	e_{1i}	e_{2i}
1	10	6	6	12	1	0	1	1/2	1/2
2	24	4	5	9	0	1	1	4/9	5/9
3	26	4	4	8	1	0	1	1/2	1/2
4	28	3	4	7	1	0	1	3/7	4/7
5	30	2	4	6	1	1	2	2/3	4/3
6	41	1	2	3	1	0	1	1/3	2/3
7	42	0	2	2	0	1	1	0	1
Total					$O_1 = 5$	$O_2 = 3$		$E_1 = 2.87$	$E_2 = 5.13$

Table 8: *Calculations for the log rank test applied to the brain tumour data*

The log rank test can be generalised to more than 2 groups. If we have k groups then the log rank statistic has a χ^2_{k-1} distribution under H_0.

3.2.4 Performing the Log Rank Test in R

The function for performing log rank tests is `survdiff()`. The procedure is similar to calculating a non-parametric (i.e. Kaplan-Meier) survival model. We start by creating a survival object using the `Surv()` function. We then regress the `Surv` object on the group variable. We do this for the brain tumour survival times:

```
> load("braintu.Rdata")
> brain_sv <- Surv(braintu$time, braintu$censor, type = "right")
> names(braintu) # look for name of grouping variable
[1] "time"   "censor" "group" # ’group’ is the grouping variable


> survdiff(brain_sv ~ braintu$group)
Call:
survdiff(formula = brain_sv ~ braintu$group)


                N Observed Expected (O-E)^2/E (O-E)^2/V
braintu$group=1 6        5     2.87     1.575      2.88
braintu$group=2 6        3     5.13     0.882      2.88


 Chisq= 2.9  on 1 degrees of freedom, p= 0.09
```

Note that the numerical value of the test statistic is slightly different from our log rank statistic value of 2.46 since R handles ties in a more sophisticated way.

We can produce separate Kaplan-Meier plots for each group (see Figure 8) using

```
> brainSurv <- survfit(brain_sv ~ braintu$group)
> ggsurvplot(brainSurv, data = braintu, surv.median.line = 'hv',
             conf.int = TRUE, conf.int.style = "ribbon",
             conf.int.alpha = 0.4, palette = "npg", pval = TRUE,
             pval.method = TRUE, pval.size = 3, xlim = c(0,45),
             xlab = "Time (months)", risk.table = T,
             break.time.by = 5, ncensor.plot = TRUE, censor = TRUE,
             censor.shape = 3, censor.size = 5, xscale = 1,
             ncensor.plot.height = 0.25, tables.height = 0.25,
             legend.labs = c("Treatment 1", "Treatment 2"))
```

3.3 Parametric Tests

To perform parametric tests, we generally need to use asymptotic properties of maximum likelihood estimates of parameters or likelihood ratios. The exponential lifetime model is given here as an illustration, but all other models could be handled similarly with numerical estimation of parameters and the asymptotic variances.

3.3.1 MLE Test

Suppose we have n_1 observations of $T_1 \sim Exp(\lambda_1)$ and n_2 observations of $T_2 \sim Exp(\lambda_2)$. Now the mles are given by

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n_j} \delta_{ji}}{\sum_{i=1}^{n_j} t_{ji}} = \frac{\Delta_j}{\mathcal{T}_j} \quad \text{for } j = 1, 2; \tag{78}$$

where Δ_j is the number of deaths in group j and $\mathcal{T}_j$ is the total time on test in group j (i.e the sum of all observed failure times and censored times). From Equations (51) and (58) we know that

$$\hat{\lambda}_j \approx N\left(\lambda_j, \frac{\lambda_j^2}{\Delta_j}\right) \quad \text{for } j = 1, 2. \tag{79}$$

Replacing λ_j in the variance term with its estimate $\hat{\lambda}_j$ we have

$$\hat{\lambda}_1 - \hat{\lambda}_2 \approx N\left(\lambda_1 - \lambda_2, \frac{\hat{\lambda}_1^2}{\Delta_1} + \frac{\hat{\lambda}_2^2}{\Delta_2}\right). \tag{80}$$

Hence, to test the hypothesis

$$H_0: \quad S_1(t) = S_2(t) \tag{81}$$

$$H_1: \quad S_1(t) \neq S_2(t) \quad \text{for some } t \tag{82}$$

we first note that (in our parametric setting) it is equivalent to testing

$$H_0: \quad \lambda_1 = \lambda_2 \tag{83}$$

$$H_1: \quad \lambda_1 \neq \lambda_2 \tag{84}$$

and under H_0

$$W = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\frac{\hat{\lambda}_1^2}{\Delta_1} + \frac{\hat{\lambda}_2^2}{\Delta_2}}} \approx N(0,1). \tag{85}$$

3.3.2 MLE Test applied to the Brain Tumour Data

We have

- $n_1 = 6,\ \Delta_1 = 5\ \mathcal{T}_1 = 147,\ \hat{\lambda}_1 = 5/147 = 0.034$
- $n_2 = 6,\ \Delta_2 = 3,\ \mathcal{T}_2 = 193,\ \hat{\lambda}_2 = 3/193 = 0.0155$

Theses give $W = 1.05$ and comparing this with 1.96 (0.975 quantile of the $N(0,1)$ distribution) we have no evidence at the 5% level that the survivor functions differ.

3.3.3 Likelihood Ratio Test

An alternative parametric approach is to use the likelihood ratio test. We have the likelihood

$$L(\lambda_1, \lambda_2) = L(\lambda_1)L(\lambda_2) \tag{86}$$

and so to maximize $L(\lambda_1, \lambda_2)$ we can maximise with respect to λ_1 and λ_2 separately. Equation (47) gives

$$L_{max}(\lambda_1, \lambda_2) = L(\hat{\lambda}_1, \hat{\lambda}_2) = \hat{\lambda}_1^{\Delta_1} e^{-\hat{\lambda}_1 \mathcal{T}_1} \hat{\lambda}_2^{\Delta_2} e^{-\hat{\lambda}_2 \mathcal{T}_2} \tag{87}$$

where $\hat{\lambda}_i = \frac{\Delta_i}{\mathcal{T}_i}$ for $i = 1, 2$. Whereas, if H_0 is true, then the likelihood is

$$L(\lambda, \lambda) = \lambda^{\Delta_1 + \Delta_2} e^{-\lambda(\mathcal{T}_1 + \mathcal{T}_2)} \tag{88}$$

$$\Rightarrow l(\lambda, \lambda) = (\Delta_1 + \Delta_2)\log \lambda - \lambda(\mathcal{T}_1 + \mathcal{T}_2) \tag{89}$$

$$\Rightarrow \tilde{\lambda} = \frac{\Delta_1 + \Delta_2}{\mathcal{T}_1 + \mathcal{T}_2} \tag{90}$$

$$\Rightarrow L(\tilde{\lambda}, \tilde{\lambda}) = \tilde{\lambda}^{\Delta_1 + \Delta_2} e^{-\tilde{\lambda}(\mathcal{T}_1 + \mathcal{T}_2)}. \tag{91}$$

where $\tilde{\lambda} = \frac{\Delta_1 + \Delta_2}{\mathcal{T}_1 + \mathcal{T}_2}$ is the MLE under H_0. So using the generalised likelihood ratio test we have, under H_0,

$$2\left\{ l(\hat{\lambda}_1, \hat{\lambda}_2) - l(\tilde{\lambda}, \tilde{\lambda}) \right\} \approx \chi_1^2 \tag{92}$$

i.e.

$$2\left\{ \Delta_1 \log \frac{\Delta_1}{\mathcal{T}_1} + \Delta_2 \log \frac{\Delta_2}{\mathcal{T}_2} - (\Delta_1 + \Delta_2)\log \frac{\Delta_1 + \Delta_2}{\mathcal{T}_1 + \mathcal{T}_2} \right\} \approx \chi_1^2 \tag{93}$$

3.3.4 Likelihood Ratio Test applied to the Brain Tumour Data

Calculation gives the test statistic as 1.20 and since $1.20 < 3.84$ ($\chi^2_{1,0.95} = 3.84$) we again have no evidence at the 5% level that the survivor functions differ.

3.3.5 Performing the Likelihood Ratio Test in R

We can do the MLE test using the `survreg()` function by specifying the `dist` argument to be `exponential`.

```
> brain_sv <- Surv(braintu$time, braintu$censor)
> br_regexp <- survreg(brain_sv ~ as.factor(braintu$group),
                       dist = "exponential")
> summary(br_regexp)
                          Value Std. Error    z     p
(Intercept)               3.381      0.447 7.56 4e-14
as.factor(braintu$group)2 0.783      0.730 1.07  0.28

Scale fixed at 1

Exponential distribution
Loglik(model)= -37.4   Loglik(intercept only)= -38
Chisq= 1.2 on 1 degrees of freedom, p= 0.27
Number of Newton-Raphson Iterations: 4
n= 12

br_regexp$var # gives the covariance matrix
                  (Intercept) as.factor(group)2
(Intercept)               0.2        -0.2000000
as.factor(group)2        -0.2         0.5333333
```

Note that:

- $\hat{\lambda}_1 = 1/\exp(3.381) = 0.034$
- $\hat{\lambda}_2 = 1/\exp(3.381 + 0.783) = 0.015$
- Equation (58) gives $\text{se}(\hat{\lambda}_1) \approx \exp(-3.381)/\sqrt{5}$
- $\text{se}(\lambda_2) \approx (0.447^2 + 0.7302^2 + 2(-0.2))^{\frac{1}{2}}/\exp(3.381 + 0.783)$
- It is simpler to note in the `R` output that the p-value for testing whether the parameter indicating the group is zero is 0.284 (this tests whether the groups differ in their survival curves)
- This is close to the p-value of 0.308 for the MLE test statistic of 1.05 calculated in Section 3.3.2 [given by `2*(1-pnorm(1.05)`]

- The χ_1^2 value of 1.2 (giving a p-value of 0.27) given above is the value of the likelihood ratio test statistic.

It can be shown that the MLE and likelihood ratio tests are asymptotically equivalent so for large sample sizes they should lead to similar conclusions. There is some research showing that the LRT is more appropriate for small samples although this is not a hard rule. In large samples, little power is gained by using parametric methods instead of the log-rank test.

4 Regression Models

So far we have just looked at problems when all the failure times of all the individuals are assumed IID or at most there are a fixed number of distinct groups with a common failure time distribution for each group. However, each individual might have their own failure time distribution dependent upon a number of characteristics about either them or their treatment. For a more sensitive analysis, we might want to include in our model (of the failure time) the possible effect of these explanatory variables. Examples of such variables might be:

- Treatment given (typically binary or categorical if multiple treatments are considered)
- Age of individual (continuous)
- Sex of individual (categorical)
- Measured variables describing prior medical history

For each individual, with explanatory variables $\mathbf{x}$ we may wish to allow their survival time to depend upon these variables. When we do so, we typically define $\mathbf{x} = \mathbf{0}$ to correspond to a baseline and then measure $\mathbf{x}$ relative to this baseline.

We will look at two different ways that such information can be incorporated into a model for the failure time:

1. Accelerated Failure Time Models (or Accelerated Life Models)
2. Proportional Hazards

4.1 Accelerated Failure Time (AFT) Models

4.1.1 Two-Sample Example

Consider initially the example where we have n individuals belonging to one of two groups dependent upon whether they are controls (Group 0) or receive a treatment (Group 1). In this case the binary variable

$$\mathbf{x_i} = \begin{cases} 0 & \text{if person } i \text{ is in the control group} \\ 1 & \text{if person } i \text{ is in the treatment group} \end{cases}$$

denotes which group the i^{th} individual belongs to. Suppose the failure times of the individuals in the control and treatment groups are labelled as T_0 and T_1 respectively. If $S_0(t)$ and $S_1(t)$ are the survivor functions then we could choose to model them as

$$\begin{aligned} &\text{Group 0:} \quad S_0(t) = P(T_0 > t) && (94) \\ &\text{Group 1:} \quad S_1(t) = P(T_1 > t) = P(T_0 > \psi t) = S_0(\psi t) && (95) \end{aligned}$$

so that, in terms of random variables $T_1 = T_0/\psi$. Intuitively this means that lifetime is accelerated/decelerated in group 1 compared with group 0. Any individual with survival time t in group 0 would have survival time t/ψ in group 1. If $\psi > 1$ then individuals will fail more quickly in group 1 compared with group 0 while if $\psi < 1$ then they will fail more slowly.

4.2 General AFT model Framework for including Explanatory Variables

Suppose instead that we have a vector of explanatory variables ($\mathbf{x}$) for each individual. We can extend this idea by modelling the survivor function of an individual with variables $\mathbf{x}$ as

$$S(t; \mathbf{x}) = S_0(t\psi(\mathbf{x})) \tag{96}$$

where $S_0(t)$ corresponds to the survivor function of the chosen baseline $\mathbf{x} = \mathbf{0}$ for any positive function $\psi(\mathbf{x})$ (with $\psi(\mathbf{0}) = 1$ to satisfy $S(t; \mathbf{x} = \mathbf{0}) = S_0(t)$). With this parameterisation we have the density and hazard functions (as functions of t and $\mathbf{x}$) given by

$$f(t; \mathbf{x}) = f_0(t\psi(\mathbf{x}))\psi(\mathbf{x}) \tag{97}$$

$$h(t; \mathbf{x}) = h_0(t\psi(\mathbf{x}))\psi(\mathbf{x}) \tag{98}$$

In terms of random variables this is equivalent to defining

$$T = T_0/\psi(\mathbf{x}) \tag{99}$$

where T_0 has survivor function $S_0(\cdot)$. Since we require $\psi(\mathbf{x}) \geq 0$ and $\psi(\mathbf{0}) = 1$ a natural choice is

$$\psi(\mathbf{x}) = e^{-\boldsymbol{\beta}'\mathbf{x}}. \tag{100}$$

Using Equations (99) and (100) we have

$$E[T] = e^{\boldsymbol{\beta}'\mathbf{x}} E[T_0]. \tag{101}$$

Sometimes this model is called a log-linear type model since, taking logs of both sides in Equation (101) we find

$$\log(E[T]) = \mu_0 + \boldsymbol{\beta}'\mathbf{x} \tag{102}$$

where $\mu_0 = \log(E[T_0])$.

4.2.1 Using this Model

To employ this model in practice we would assume a parametric distribution (e.g. Exponential, Weibull, ...) for T_0. We can then estimate the parameters needed for that distribution and $\boldsymbol{\beta}$ by maximum likelihood. This can all be done in *R* which will also find confidence intervals for the different parameters.

4.2.2 An Example — Exponential Model

Suppose that we have n individuals where for each we have observed

$$\begin{aligned} t_i &= \min(\text{Failure time}, \text{Time of right censoring}) \\ \delta_i &= \begin{cases} 1 & \text{if } i \text{ is observed to fail} \\ 0 & \text{if } i \text{ is censored} \end{cases} \\ \mathbf{x_i} &= p\text{-dimensional vector of explanatory variables} \end{aligned}$$

For an exponential AFT model, Equations (96) to (98) give

$$S_0(t) = e^{-\lambda t} \quad \Rightarrow \quad S(t;\mathbf{x}) = e^{-\lambda t e^{-\boldsymbol{\beta}'\mathbf{x}}} \tag{103}$$

$$h_0(t) = \lambda \quad \Rightarrow \quad h(t;\mathbf{x}) = \lambda e^{-\boldsymbol{\beta}'\mathbf{x}} \tag{104}$$

$$f_0(t) = \lambda e^{-\lambda t} \quad \Rightarrow \quad f(t;\mathbf{x}) = \lambda e^{-\boldsymbol{\beta}'\mathbf{x}} e^{-\lambda t e^{-\boldsymbol{\beta}'\mathbf{x}}}. \tag{105}$$

Note that this is equivalent to assuming that with explanatory variables $\mathbf{x}$ the failure time $T \sim Exp(\lambda e^{-\boldsymbol{\beta}'\mathbf{x}})$. We estimate the $p+1$ parameters $\lambda, \beta_1, \beta_2, \ldots, \beta_p$ by maximum likelihood. We write the likelihood as a product of terms representing the observed failures and a product of terms representing the censored observations.

$$L(\lambda,\boldsymbol{\beta}) = \prod_{i=1}^{n} f(t_i;\mathbf{x_i})^{\delta_i} S(t_i;\mathbf{x_i})^{1-\delta_i} \tag{106}$$

$$= \prod_{i=1}^{n} \left[\lambda e^{-\boldsymbol{\beta}'\mathbf{x_i}} e^{-\lambda t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}}}\right]^{\delta_i} \left[e^{-\lambda t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}}}\right]^{1-\delta_i} \tag{107}$$

$$= \prod_{i=1}^{n} \left[\lambda e^{-\boldsymbol{\beta}'\mathbf{x_i}}\right]^{\delta_i} e^{-\lambda t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}}}. \tag{108}$$

Taking logs gives

$$l(\lambda,\boldsymbol{\beta}) = \log\lambda \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i \boldsymbol{\beta}'\mathbf{x_i} - \lambda \sum_{i=1}^{n} t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}}. \tag{109}$$

Differentiating gives

$$\frac{\partial l}{\partial \lambda} = \frac{\Delta}{\lambda} - \sum t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}} \quad \text{where } \Delta = \sum_{i=1}^{n} \delta_i \tag{110}$$

$$\frac{\partial l}{\partial \beta_j} = -\sum_{i=1}^{n} \delta_i x_{ij} + \lambda \sum_{i=1}^{n} x_{ij} t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}} \quad \text{for } j = 1, \ldots, p. \tag{111}$$

We can set these two derivatives equal to zero and solve iteratively to find maximum likelihood estimates for λ and $\boldsymbol{\beta}$. For estimates of variance and standard errors we can use the asymptotic properties of the mle and note that

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\Delta}{\lambda^2} \tag{112}$$

$$\frac{\partial^2 l}{\partial \lambda \partial \beta_j} = \sum_{i=1}^{n} x_{ij} t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}} \qquad \text{for } j = 1, \ldots, p \tag{113}$$

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\lambda \sum_{i=1}^{n} x_{ij} x_{ik} t_i e^{-\boldsymbol{\beta}'\mathbf{x_i}} \qquad \text{for } j = 1, \ldots, p; \quad k = 1, \ldots, p \tag{114}$$

4.3 Fitting an Exponential AFT Model to the Myelogenous Leukemia Data in R

Parametric models can be fitted using `survreg()` as described in Section 2.5.7. We illustrate this procedure on the survival times of patients with myelogenous leukemia found in `myleuk.Rdata`. Here the survival times (from date of diagnosis) of patients with acute myelogenous leukemia were observed. This time was thought to depend upon the white blood cell count. Preliminary analysis (plotting survival time against covariate) suggested that a log transformation of this white blood cell count was needed and so we work with `log.wbc.`. (Note that in this particular case none of the observations is censored so in the initial `Surv()` step the censoring variable can be omitted).

```
> library(survival)
> load("wbcleuk.Rdata")
> head(wbcleuk)
  patient   wbc survival log.wbc.
1       1  23.0       65 3.135494
2       2   7.5      156 2.014903
3       3  43.0      100 3.761200
4       4  26.0      134 3.258097
5       5  60.0       16 4.094345
6       6 105.0      108 4.653960
> # Note that there is no censoring so no need for 2nd argument in Surv
> wbcleuk_sv <- Surv(wbcleuk$survival)
> wbcleuk_regexp <- survreg(wbcleuk_sv ~ wbcleuk$log.wbc., dist="exponential")
> summary(wbcleuk_regexp)

Call:
survreg(formula = wbcleuk.sv ~ log.wbc., dist = "exponential")
             Value Std. Error     z        p
(Intercept)  7.845       1.05  7.45 9.05e-14
log.wbc.    -0.889       0.22 -4.05 5.15e-05
```

```
Scale fixed at 1

Exponential distribution
Loglik(model)= -84.5   Loglik(intercept only)= -92.9
Chisq= 16.86 on 1 degrees of freedom, p= 4e-05
Number of Newton-Raphson Iterations: 4
n= 18
```

From Equation (103), an exponential AFT model with explanatory variable vector $\mathbf{x}$ has survivor function

$$S(t;\mathbf{x}) = e^{-\lambda t e^{-\boldsymbol{\beta}'\mathbf{x}}} \tag{115}$$

$$= e^{-t\exp(-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p))} \quad \text{if we let} \quad \lambda = \exp(-\beta_0). \tag{116}$$

This implies $T \sim Exp(\lambda(\mathbf{x}))$ where

$$\lambda(\mathbf{x}) = \exp\left(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)\right). \tag{117}$$

The R output provides us with estimates of $\beta_0, \beta_1, \ldots, \beta_p$ and their standard errors. These can be used to perform [partial] z-tests of hypotheses that the separate covariates have no effect on the survival distribution. The tests are strictly conditional on all the remaining covariates being included in the model and hence are properly termed partial z-tests.

In the myelogenous leukemia data we have only one explanatory variable x_1 representing the log wbc. From the intercept term we can work out the mle of the rate parameter $\lambda(0)$ for the baseline failure time distribution T_0 (with $x = 0$) using

$$\hat{\lambda}(0) = e^{-7.845}. \tag{118}$$

The parameter estimate $\hat{\beta}_1 = -0.889$ corresponds to effect of the log(wbc). If we want to know the mean survival time for a patient with a `wbc` of 54, the mle is given by

$$\hat{\lambda}(\log(54)) = \exp(-7.845 + 0.889\log(54)). \tag{119}$$

and the estimate of the mean survival time for such an individual is therefore

$$\frac{1}{\hat{\lambda}(\log(54))} = e^{7.845-0.889\log(54)} = 73.61 \text{ days}. \tag{120}$$

An approximate standard error for this estimate can be calculated from the standard errors for the intercept and coefficient and using the formula for the variance of a function of a MLE given in Section 2.5.3. Since this model is an exponential model, the estimated median survival time for a patient with a wbc of 54 is

$$\frac{1}{\hat{\lambda}(\log(54))}\log(2) = 50.57\text{days}. \tag{121}$$

4.4 Fitting a Weibull AFT Model to the Myelogenous Leukemia Data in R

As a comparison, we fit a Weibull model to the same data. Since the Weibull is the default distribution for `survreg()` it is not necessary to specify the distribution.

```
> wbcleuk_regweib <- survreg(wbcleuk_sv ~ wbcleuk$log.wbc.)
> summary(wbcleuk_regweib)

Call:
survreg(formula = wbcleuk.sv ~ log.wbc.)
             Value Std. Error      z        p
(Intercept)  7.849      0.986  7.957 1.76e-15
log.wbc.    -0.882      0.207 -4.265 2.00e-05
Log(scale)  -0.105      0.187 -0.562 5.74e-01

Scale= 0.901

Weibull distribution
Loglik(model)= -84.3   Loglik(intercept only)= -92
Chisq= 15.4 on 1 degrees of freedom, p= 8.7e-05
Number of Newton-Raphson Iterations: 5
n= 18
```

Recall our Weibull parameterisation in Equations (14) - (16):

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1} \exp\left[-(\lambda t)^{\gamma}\right] \tag{122}$$

$$S(t) = \exp\left[-(\lambda t)^{\gamma}\right] \tag{123}$$

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \tag{124}$$

If we model T as an AFT Weibull model with baseline $T_0 \sim Weibull(\lambda_0, \gamma)$ then

$$S(t; \mathbf{x}) = \exp\left(-\left(\lambda_0 t \exp(-\boldsymbol{\beta}'\mathbf{x})\right)^{\gamma}\right) \tag{125}$$

$$= \exp\left(-\left(t \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))\right)^{\gamma}\right) \tag{126}$$

where $\lambda_0 = \exp(-\beta_0)$. This is equivalent to $T \sim Weibull(\lambda(\mathbf{x}), \gamma)$ where

$$\lambda(\mathbf{x}) = \exp\left(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)\right) \tag{127}$$

In an AFT Weibull model, only $\lambda(\mathbf{x})$ depends on $\mathbf{x}$. In the `R` output, the first thing to note is that the estimated rate parameter is $\hat{\gamma} = 1/0.901 = 1.1$ which is very close to 1.0 (equivalent to the exponential model). In fact noting that the log(scale) is estimated as -0.105 with standard error 0.187 it is clear that this estimate is not significantly different from zero so there is little evidence provided by these data that the Weibull model fits better than the simpler exponential model. Under this model, the estimated mean survival time, assuming a Weibull distribution for the survival time, is

$$\hat{\lambda}(\mathbf{x})^{-1}\Gamma\left(1+\frac{1}{\hat{\gamma}}\right) \tag{128}$$

If we approximate $\hat{\gamma}=1$ then we find the predicted mean survival time for a patient with a `wbc` of 54 is

$$e^{7.849-0.882\log(54)} = 76 \text{ days},$$

little different from the estimate based on the exponential model.

4.5 Proportional Hazards

An alternative approach to modelling the dependence of the failure time distribution on explanatory variables $\mathbf{x}$ is through the hazard function with

$$h(t;\mathbf{x}) = \psi(\mathbf{x};\boldsymbol{\beta})h_0(t) \tag{129}$$

where $h_0(t)$ corresponds to the hazard for an individual under the baseline conditions where $\mathbf{x}=\mathbf{0}$ and we require $\psi(\mathbf{x}=\mathbf{0};\boldsymbol{\beta})=1$. Typically we choose

$$\psi(\mathbf{x};\boldsymbol{\beta}) = e^{\boldsymbol{\beta}'\mathbf{x}} \tag{130}$$

so that

$$h(t;\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}h_0(t). \tag{131}$$

This is a semi-parametric model proposed by Cox (1972). The work he presented has become so widely used that his original paper is one of the most cited mathematical papers in history and recently won him the equivalent of the Nobel prize for statistics.

Proportional-hazards is called semi-parametric since the baseline hazard $h_0(t)$ does not need to be specified. The dependence of the failure time T on explanatory variables $\mathbf{x}$ is precisely modelled; but actual distribution of failure is not parametrically specified. It is particularly useful in medical situations where

- it is important to know which prognostic variables have an effect and to what extent
- **but** knowing the actual distribution of survival time is not as important.

One should also note that for two patients with covariates $\mathbf{x_1}$ and $\mathbf{x_2}$ we have

$$\frac{h(t;\mathbf{x_1})}{h(t;\mathbf{x_2})} = \frac{e^{\boldsymbol{\beta}'\mathbf{x_1}}}{e^{\boldsymbol{\beta}'\mathbf{x_2}}} \tag{132}$$

$$= e^{\boldsymbol{\beta}'(\mathbf{x_1}-\mathbf{x_2})} \tag{133}$$

which is **independent of time**. Hence hazard functions for any two patients are proportional over time as the linear component of the model does not vary with time.

4.5.1 Which AFT models exhibit the proportional hazards property?

Suppose $T \sim Exp(\lambda)$ then for an AFT with an exponential survival time, Equation (104) gives $h(t;\mathbf{x}) = \lambda e^{-\boldsymbol{\beta}'\mathbf{x}}$. It follows that the ratio of any two hazards with different values of $\boldsymbol{x}$ is independent of time and therefore has the proportional hazards property. Similarly if $T \sim Weibull(\lambda, \gamma)$ then for an AFT with a Weibull survival time, the baseline hazard is $h_0(t) = \lambda\gamma(\lambda t)^{\gamma-1}$. Using Equations (98) and (100) we have that $h(t;\mathbf{x}) = \lambda^{\gamma}\gamma t^{\gamma-1}exp(-\gamma\boldsymbol{\beta}'\boldsymbol{x})$. It follows that the ratio of any two hazards with different values of $\boldsymbol{x}$ is independent of time and therefore has the proportional hazards property. Note that because the exponential is a special case of the Weibull we only really needed to demonstrate this property for the Weibull distribution. The Weibull distribution (and hence Exponential as a special case) is the only distribution for which the corresponding AFT model has the proportional hazards property.

4.5.2 An Initial Attempt at Parameter Estimation

Consider our typical situation where we have n individuals and for each we have observed

$$t_i = \min(\text{Failure time}, \text{Time of right censoring})$$
$$\delta_i = \begin{cases} 1 & \text{if } i \text{ is observed to fail} \\ 0 & \text{if } i \text{ is censored} \end{cases}$$
$$\mathbf{x_i} = p \text{ dimensional vector of explanatory variables}$$

If we wish to model this with a proportional hazards model then

$$\begin{aligned} h(t;x) &= e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t) && (134)\\ S(t;x) &= \exp\left\{-\int_0^t h(t)dt\right\} && (135)\\ &= \exp\left\{-\int_0^t e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t)dt\right\} && (136)\\ &= \exp\left\{-e^{\boldsymbol{\beta}'\mathbf{x}}\int_0^t h_0(t)dt\right\} && (137)\\ &= [S_0(t)]^{e^{\boldsymbol{\beta}'\mathbf{x}}} && (138)\\ f(t;\mathbf{x}) &= h(t;\mathbf{x})S(t;\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t)S(t). && (139)\end{aligned}$$

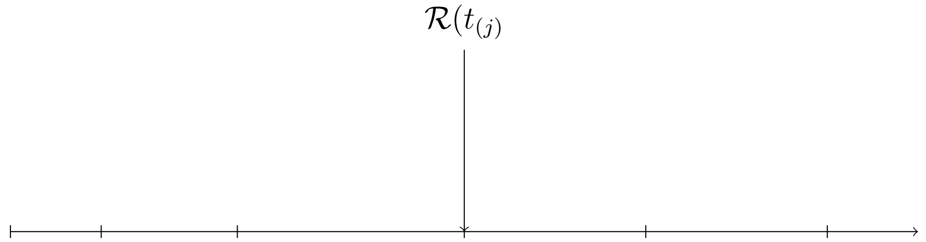
We can do our usual approach to writing down the likelihood of the censored and observed data

$$\begin{aligned} \text{Likelihood} &= \prod_{i=1}^{n} f(t_i;\mathbf{x_i})^{\delta_i} S(c_i;\mathbf{x_i})^{1-\delta_i} && (140)\\ &= \prod_{i=1}^{n} \left[\frac{h(t_i;\mathbf{x_i})}{S(t_i;\mathbf{x_i})}\right]^{\delta_i} S(c_i;\mathbf{x_i})^{1-\delta_i}. && (141)\end{aligned}$$

However, this involves the unspecified baseline hazard $h_0(t)$ and so to proceed further we would need to specify its parametric form. However we don't want to do that and so we will proceed by working with the partial likelihood.

4.5.3 Partial Likelihood

Without specifying the form of the baseline hazard $h_0(t)$, the actual failure times t_i are not very informative. Instead we will work with the order that people fail i.e. who out of the people at risk of failing actually does so at each time. We define the random variable ζ_j to be the index of the individual who fails at the j^{th} failure time, conditional on that failure time being $t_{(j)}$.



Time	0	$t_{(1)}$	$t_{(2)}$	$t_{(j)}$	$t_{(n-1)}$	$t_{(n)}$	t
Failed Individual		$i_{(1)}$	$i_{(2)}$	$i_{(j)}$	$i_{(n-1)}$	$i_{(n)}$	
Censor Indicator		$\delta_{(1)}$	$\delta_{(2)}$	$\delta_{(j)}$	$\delta_{(n-1)}$	$\delta_{(n)}$	
Covariates		$\mathbf{x_{(1)}}$	$\mathbf{x_{(2)}}$	$\mathbf{x_{(j)}}$	$\mathbf{x_{(n-1)}}$	$\mathbf{x_{(n)}}$	

Figure 9: *Illustration of labelling for failure order for calculation of proportional hazards times with no censoring. We first order the failure times $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$. This creates a corresponding ordering on the rest of the variables so that the j^{th} individual who fails (at time $t_{(j)}$) is labelled $i_{(j)}$. This individual has censoring indicator value $\delta_{(j)}$ and explanatory covariates* $\mathbf{x_{(j)}}$

Simple Example — No censoring Suppose that we have no censoring in the data and that there are no ties in failure times i.e. only one individual dies at each death time. Let us order the observed failure times and denote them by $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$. Also let us create an order amongst the individuals so that the individual who dies at failure time $t_{(j)}$ is $i_{(j)}$ i.e. $i_{(j)} = k$ if and only if $t_k = t_{(j)}$. Finally let define a risk set

$$\mathcal{R}(t) = \{\text{The set of individuals alive and in the trial just before time } t\}.$$

This labelling is illustrated in Figure 9. Let $t^\dagger = \{t_{(1)}, \ldots, t_{(n)}\}$. We want to work with the likelihood or probability of the observed order $(i_{(1)}, \ldots, i_{(n)})$ in which the individuals fail conditional on those failure times i.e.

$$P(\zeta_1 = i_{(1)}, \zeta_2 = i_{(2)}, \ldots, \zeta_n = i_{(n)}|t^\dagger) \tag{142}$$

Using repeated conditioning we can split this probability as

$$P(\zeta_1 = i_{(1)}, \zeta_2 = i_{(2)}, \ldots, \zeta_n = i_{(n)}|t^\dagger) \tag{143}$$

$$=P(\zeta_1 = i_{(1)}|t^\dagger) \prod_{j=2}^{n} P(\zeta_j = i_{(j)}|\zeta_1 = i_{(1)}, \ldots, \zeta_{j-1} = i_{(j-1)}, t^\dagger) \tag{144}$$

and we can deal with each of these probabilities separately. Now

$$P(\zeta_j = i|\zeta_1 = i_{(1)}, \zeta_2 = i_{(2)}, \ldots, \zeta_{j-1} = i_{(j-1)}, t^\dagger)) \tag{145}$$

is the probability that individual i fails at time $t_{(j)}$ given that we know one individual does fail at that time from those individuals at risk (i.e. in the risk set $\mathcal{R}(t_{(j)})$). This can be thought of as picking an individual at random from the risk set where the probability you pick any individual k is proportional to that individuals hazard $h(t_{(j)}; \mathbf{x_k})$. As such it can be written as

$$\frac{h(t_{(j)}; \mathbf{x_i})}{\sum_{k\in\mathcal{R}(t_{(j)})} h(t_{(j)}; \mathbf{x_k})} = \frac{e^{\boldsymbol{\beta}'\mathbf{x_i}} h_0(t_{(j)})}{\sum_{k\in\mathcal{R}(t_{(j)})} e^{\boldsymbol{\beta}'\mathbf{x_k}} h_0(t_{(j)})} \tag{146}$$

$$= \frac{e^{\boldsymbol{\beta}'\mathbf{x_i}}}{\sum_{k\in\mathcal{R}(t_{(j)})} e^{\boldsymbol{\beta}'\mathbf{x_k}}} \tag{147}$$

after cancelling the baseline hazards. Putting all these terms together, we can write down the probability of the observed order $(i_{(1)}, \ldots, i_{(n)})$ as

$$P(\zeta_1 = i_{(1)}, \zeta_2 = i_{(2)}, \ldots, \zeta_n = i_{(n)}|t_{(1)}, \ldots, t_{(n)}) = \prod_{j=1}^{n} \frac{e^{\boldsymbol{\beta}'\mathbf{x_{(j)}}}}{\sum_{k\in\mathcal{R}(t_{(j)})} e^{\boldsymbol{\beta}'\mathbf{x_k}}} \tag{148}$$

where $\mathbf{x_{(j)}}$ is the vector of explanatory variables for individual $i_{(j)}$ i.e. the j^{th} individual to fail.

With censoring We can extend the above idea to also include censored observation by adjusting the risk set $\mathcal{R}(t)$ accordingly to take account of those individuals who have been censored by time t. We also only need to take account of the actual failure times i.e. when $\delta_i = 1$ for each term in the product. As such we can write down a partial likelihood for the observed data and parameter $\boldsymbol{\beta}$ as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \left(\frac{e^{\boldsymbol{\beta}'\mathbf{x_{(j)}}}}{\sum_{k\in\mathcal{R}(t_{(j)})} e^{\boldsymbol{\beta}'\mathbf{x_k}}} \right)^{\delta_{(j)}} \tag{149}$$

Note that individuals for whom the survival time is censored do not contribute to the numerator but they do enter the summation over the risk set at death times of subject less than the censored time.

4.5.4 Notes on Partial Likelihood Approach

- If there are no censored observations this is a conditional likelihood, conditional on the observed failure times $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$.
- With censored observations this is instead known as a partial likelihood. The use of such a partial likelihood was justified by Cox (1975). He showed that the usual likelihood methods apply in this case. So we maximize the (partial) likelihood to estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and asymptotically

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left[-\frac{\partial^2 l}{\partial\beta_i\partial\beta_j}\right]^{-1}_{\hat{\boldsymbol{\beta}}}\right) \tag{150}$$

where l is the log likelihood.

- **Ties (Peto's adjustment)**
  If we have data with ties in the failure times i.e.

$$\begin{array}{rl} t_{(1)} < t_{(2)} < \ldots < t_{(k)} & \text{the } k \text{ distinct survival times} \\ d_{(1)}, d_{(2)}, \ldots, d_{(k)} & \text{numbers of deaths at these times} \\ \mathcal{D}(t_{(1)}), \mathcal{D}(t_{(2)}), \ldots, \mathcal{D}(t_{(k)}) & \text{death set at } t_{(j)} \end{array}$$

  we can allow each of $d_{(j)}$ deaths at $t_{(j)}$ to contribute a factor (as before) to partial likelihood, each with the same risk set $\mathcal{R}(t_{(j)})$ so that

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \left(\frac{e^{\sum_{k \in \mathcal{D}(t_{(j)})} \boldsymbol{\beta}' \mathbf{x_k}}}{\sum_{k \in \mathcal{R}(t_{(j)})} e^{\boldsymbol{\beta}' \mathbf{x_k}}} \right)^{\delta_{(j)}}$$

  This is satisfactory provided $d_{(j)}/n_{(j)}$ is small, where $n_{(j)}$ is the number of individuals at risk at $t_{(j)}$.

4.5.5 Interpreting an Example: Atrial Fibrillation

Table 9 shows the results of fitting a proportional hazards model to data obtained from patients being treated for atrial fibrillation. Only main effects are considered in the model (no interactions). The purpose of the treatment is to maintain normal heart rhythm and interest is in the time to relapse.

Variable	Coefficient	Standard Error	χ^2 statistic	coeff./s.e.
Treatment:				
B	-1.42	0.64	4.92	-2.22
A (baseline)	-	-	-	-
Age (years)	-0.01	0.02	0.25	-0.12
Heart volume	0.11	0.05	4.84	2.20
Diet:				
Meat eater	1.21	0.92	1.73	1.32
Vegetarian	0.25	0.14	3.19	1.79
Vegan (baseline)	-	-	-	-
Digitalisation	-0.59	0.73	0.66	-0.81
Sex	0.31	0.72	0.19	0.43

Table 9: *Results of fitting a proportional hazards models to data from patients being treated for atrial fibrillation*

There are two questions of interest:

1. How can we quantify the effects of treatment and other covariates on time to relapse?

2. How could the analysis consider treatment-covariate interactions?

The effects of treatments

The model we are fitting for an individual with covariates $\mathbf{x} = (x_1, \ldots, x_7)'$ is

$$h(t; x) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_7 x_7}. \tag{151}$$

Variables $x_1, \ldots, x_7$ correspond to the seven variables in Table 9 in the order they appear. Note that we need two variables x_4 and x_5 to represnt 'diet' since this is a three-level factor variable (having a vegan diet is the baseline so its effect is captured by the intercept). $x_1 = 0$ for treatment A (baseline treatment), and $x_1 = 1$ for treatment B so that

$$h(t; x) = h_0(t)e^{\beta_2 x_2 + \ldots + \beta_7 x_7} \qquad \text{for treatment A} \tag{152}$$

$$h(t; x) = h_0(t)e^{\beta_1 + \beta_2 x_2 + \ldots + \beta_7 x_7} \qquad \text{for treatment B} \tag{153}$$

Therefore e^{β_1} has an interpretation of the **hazard ratio** of treatment B to treatment A as can be seen by dividing the last two expressions above. The table gives us the output $\hat{\boldsymbol{\beta}}$ along with the estimated standard error found using the approximation

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left[-\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}\right]^{-1}_{\hat{\boldsymbol{\beta}}}\right). \tag{154}$$

To see which factors are shown to affect the hazard we examine the estimated coefficients and their standard errors from the output.

Continuous or factor variables with only two levels

For these covariates we use either the χ^2 value (which we compare with a χ^2_1) or the value of coeff/s.e. (which we compare with a $N(0, 1)$).

Factor variable with k-levels ($k \geq 3$)

The these variables the χ^2 statistics for each level (excluding the baseline) of the variable should be added together and compared with a χ^2_{k-1} distribution. Note that $\chi^2 \approx$ (coeff/s.e.)^2, so it is possible to calculate the χ^2 value from the coefficient and its standard error if the χ^2 value is not given. It is **wrong** to consider the parameters for each of the levels of a factor in isolation. You need to make an **aggregated** assessment of the overall effect of the variable.

In our example we see that:

- **Treatment:** $\chi^2 = 4.92 > 3.84$ so we have significant evidence, at the 5% level of test, that treatment has an effect on relapse time. $\hat{\beta}_1 < 0$ so treatment B (coded as 1) decreases the hazard. Treatment B is better.

- **Heart volume:** $\hat{\beta}_3/s.e.(\hat{\beta}_3) = 2.11 > 1.96$ so we have significant evidence, at the 5% level of test, that increased heart volume decreases relapse time.

- **Diet:** We add up the two χ^2 values and compare it to the 0.95 quantile of the χ^2_2 distribution (which is 5.99). $1.73 + 3.19 = 4.92 < 5.99$ so there isn't significant evidence, at the 5% level of test, that diet affects relapse time.

Neither age, digitalisation nor sex appear to affect relapse time. This is **not** the same as showing that these two factors have no effect. It is useful to calculate confidence intervals for all parameters, not just those where there is significant evidence that the factor is affecting survival. This allows assessment of the **size** of the effect. For example, the 95% CI for β_6 (M/F) is $0.31 \pm 1.96 \times 0.72 = (-1.1, 1.72)$ so there **could** be a large difference in hazards between M & F. The 95% confidence interval for e^{β_6} is $(0.33, 5.58)$ so that the hazard for men could be a third that for women or nearly six times as much; these data exclude neither possibility at the 95% confidence level. Ideally we would like more data to investigate further.

Including interaction terms

Suppose that we wanted to investigate the interaction between the two binary variables **treatment** and **sex**. Then we would handle this by creating a new variable e.g.

$$x_7 = x_1 \times x_6 \tag{155}$$

the product of the treatment and sex variables. Since $x_1 = 0$ for treatment A and $x_1 = 1$ for treatment B, if follows that $x_7 = 0$ for all subjects receiving A and $x_7 = x_6$ for those receiving treatment B. The hazards for the interaction model, by treatment, are

$$h(t;x) = h_0(t)e^{\beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6} \qquad \text{for treatment A} \tag{156}$$

$$h(t;x) = h_0(t)e^{\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + (\beta_6 + \beta_7) x_6} \qquad \text{for treatment B} \tag{157}$$

where β_7 reflects the interaction effect. Interactions involving a k-level factor are handled by converting the factor into $k-1$ dummy binary variables. There are

- $k-1$ degrees of freedom for an interaction between a k-level factor and a continuous covariate
- $(k-1)(j-1)$ degrees of freedom for an interaction between a k-level and a j-level factor.

The separate parts of the χ^2 statistic are added to assess significance.

4.5.6 R Implementation

Cox proportional hazards models can be fitted using the function `coxph()`. The operation of this follows the familiar pattern of first needing to create a survival object combining the survival time with censoring information using `Surv()` and then regressing this on the explanatory variables. This is illustrated with the survival times of subjects with acute myelogenous leukaemia which has no censoring and only one explanatory variable (log white blood cell count):-

```
> library(survival)
> load("wbcleuk.Rdata")
> wbcleuk_sv <- Surv(wbcleuk$survival)
> wbcleuk_regcox <- coxph(wbcleuk_sv ~ wbcleuk$log.wbc.)
> summary(wbcleuk_regcox)
Call:
coxph(formula = wbcleuk_sv ~ log.wbc.)

  n = 18, number of events = 18

           coef exp(coef) se(coef)     z Pr(>|z|)
log.wbc. 1.1753    3.2392   0.3244 3.623 0.000292
---
         exp(coef) exp(-coef) lower .95 upper .95
log.wbc.     3.239     0.3087     1.715     6.118

Rsquare = 0.669   (max possible = 0.982 )
Likelihood ratio test = 19.89  on 1 df, p = 8.19e-06
Wald test            = 13.12  on 1 df,  p = 0.00029
Score (logrank) test = 17.39  on 1 df,  p = 3.04e-05
```

This summary provides the estimated coefficients and their standard errors to perform partial z-tests and estimate hazard ratios.

4.5.7 Model Checking

Whenever fitting a proportional hazards model, it is important to check the proportional hazards assumption i.e. that for covariates $\mathbf{x}$ relative to the baseline where $\mathbf{x} = 0$,

$$h(t; x) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t). \tag{158}$$

Typically this is done in two ways:

- Log-log plots
- Residual plots

If these diagnostic tests fails and it is concluded that a proportional hazards model is not appropriate then there are two options. One is in the special case where proportionality only breaks down for one particular factor in which case a stratified proportional hazards model might be considered. The other option is to consider a parametric AFT model which does not have the proportional hazards property i.e. not Weibull or Exponential.

Log-log plots for factor variables

Suppose that we have two treatments and we wish to test the proportional hazards assumption between these two treatments. We have defined an indicator variable x_1 so that

$$x_1 = \begin{cases} 0 & \text{if treatment A} \\ 1 & \text{if treatment B} \end{cases}$$

Then modelling via proportional hazards

$$h(t;x) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t). \tag{159}$$

so for treatment A we have hazard $h(t;0) = h_0(t)$ while for treatment B we have $h(t;1) = e^{\boldsymbol{\beta}} h_0(t)$. Now, noting that $S(t) = \exp\left\{-\int_0^t h(t)dt\right\}$ we find

$$-\log S_A(t) = \int_0^t h(t)\,dt \tag{160}$$

$$-\log S_B(t) = e^{\beta} \int_0^t h(t)\,dt \tag{161}$$

$$\Rightarrow \log[-\log S_B(t)] = \beta + \log[-\log S_A(t)] \tag{162}$$

so, if we plot an estimate of $\log[-\log S_j(t)]$ against t for both treatments we should get parallel curves a distance β apart. If the curves cross then a proportional hazards model is not appropriate.

Residuals

Various types of residuals can be defined for survival regression models. One common choice are the Schoenfeld Residuals which can be obtained with the *R* function `cox.zph()`. Schoenfeld residuals are defined only for non-censored observations and there is a separate set for each of the covariates. They are the difference between the value of the covariate of interest for the subject experiencing the event (say death) and the expectation over all members of the risk set of the covariate:

$$r_{ik} = x_{ik} - \sum_{j \in \mathcal{R}(t_i)} x_{jk}\hat{p}_j, \tag{163}$$

where r_{ik} is the Schoenfeld residual for individual i for the k^{th} covariate, x_{jk} are the values for that covariate for the individuals j in the risk set for $\mathcal{R}(t_i)$ of individuals at time t_i, the time of death for the i^{th} individual and $\hat{p}_j$ is the estimated probability that the j^{th} person dies by time t_i. These should be independent of time so a plot of these versus time should show no dependence. Deviation from independence will indicate inadequacy of the proportional hazards model.

Log-log Implementation in *R*

To produce separate log-log survival plots for different levels of a factor variable you can use the `strata` function within the model specification of the `coxph` function. The proportional hazards model is then fitted within each level of the factor. For example, for the lymphoma data where there are two levels of the `stage` variable we write

```
load("lymphoma.Rdata")
lym_sv <- Surv(lymphoma$time, lymphoma$censor, type = "right")
lym_ph <- coxph(lym_sv ~ strata(lymphoma$stage))

ggsurvplot(survfit(lym_ph), fun = "cloglog", data =  lymphoma,
           xlim = c(3,300), xlab = "Time (weeks)",
           legend.labs = c("Stage 3", "Stage 4"))
```

This produces the log-log plot shown in Figure 10. The evidence for proportional hazards here is a little inconclusive. There is some evidence of crossing but the extent is fairly marginal. This is a small dataset from which it is hard to draw meaningful conclusions.

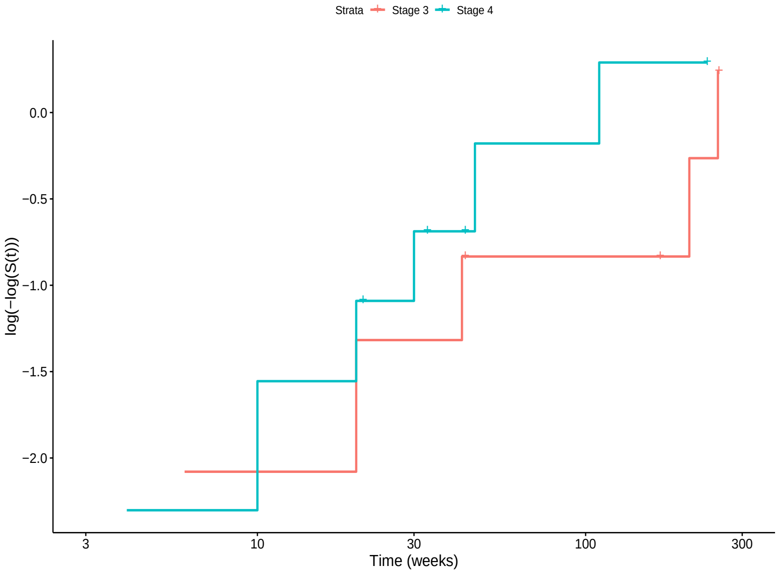


Figure 10: *Log-Log survival curves by lymphoma stage*

A Maximum Likelihood Estimation

A.1 Estimation

Suppose $x_1, \dots, x_n$ are n independent observations of a random variable X which has density function $f(\cdot;\theta)$ depending on an unknown parameter θ. There are various methods of estimating θ from the observations $x_1, \dots, x_n$ such as the method of least squares, the method of moments, the method of minimum chi-squared, ... The most central method in statistical work is the method of maximum likelihood. The procedure is to calculate the likelihood of θ for the data which will be a function of the unknown parameter θ. Then we maximize this w.r.t. θ — the value of θ which maximizes the likelihood is the maximum likelihood estimate of θ.

A.2 Definition

The likelihood of θ for data $x_1, \dots, x_n$ is

$$\begin{aligned} L(\theta; x_1, \dots, x_n) &= f(x_1;\theta)f(x_2;\theta)\dots f(x_n;\theta) \text{ if X is continuous} && (164) \\ &= P(X = x_1;\theta)P(X = x_2;\theta)\dots P(X = x_n;\theta) \text{ if X is discrete} && (165) \end{aligned}$$

(i.e. it is the product of the values of the density function or probability function evaluated at each of the observations). To obtain the maximum likelihood estimates of the parameters we maximize the calculated likelihoods w.r.t the unknown parameters.

Note: Generally it is often simpler to take the (natural) logarithms of the likelihood and maximize the log-likelihood — if the log is maximized then obviously the original likelihood will be maximized.

A.3 Examples

Suppose that we have data $x_1, \dots, x_n$ arising from the following models,

1. $X \sim N(\mu, 1)$ then

$$\begin{aligned} f(x;\mu) &= (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} && (166) \\ L(\mu; x_1, \dots, x_n) &= (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum(x_i-\mu)^2\right\} && (167) \\ \log(L(\mu)) = l(\mu) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(x_i-\mu)^2 && (168) \\ \frac{\partial l}{\partial \mu} &= \sum(x_i - \mu) && \textit{(Set to zero and solve)} \\ & && (169) \\ \Rightarrow \hat{\mu} &= \frac{\sum x_i}{n} && (170) \end{aligned}$$

2. $X \sim Exp(\lambda)$ then:

$$f(x;\lambda) = \lambda e^{-\lambda x} \tag{171}$$

$$L(\lambda; x_1, \ldots, x_n) = \lambda^n \exp\left\{-\lambda \sum x_i\right\} \tag{172}$$

$$\log(L(\lambda)) = l(\lambda) = n\log(\lambda) - \lambda \sum x_i \tag{173}$$

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i \qquad \textit{(Set to zero and solve)} \tag{174}$$

$$\Rightarrow \hat{\lambda} = \frac{n}{\sum x_i} \tag{175}$$

3. $X \sim Bin(m, p)$ then:

$$P\left(X = x\right) = \binom{m}{x} p^x \left(1-p\right)^{m-x} \tag{176}$$

$$L(p; x_1, \ldots, x_n) = \left(\prod_i \binom{m}{x_i}\right) p^{\sum x_i} \left(1-p\right)^{\sum(m-x_i)} \tag{177}$$

$$l(p) = \sum x_i \log(p) + \sum(m - x_i)\log(1-p) + C \quad \textit{(Constant C not involving p)} \tag{178}$$

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{\sum(m - x_i)}{1-p} \tag{179}$$

$$\Rightarrow \hat{p} = \frac{\bar{x}}{m} \tag{180}$$

4. $X \sim N(\mu, \sigma^2)$ then:

$$f(x;\mu,\sigma) = (2\pi)^{-\frac{1}{2}}\sigma^{-1}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{181}$$

$$L(\mu,\sigma; x_1, \ldots, x_n) = (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left\{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right\} \tag{182}$$

$$l(\mu,\sigma) = C - n\log\sigma - \frac{\sum(x_i-\mu)^2}{2\sigma^2} \tag{183}$$

$$\frac{\partial l}{\partial \mu} = \frac{\sum(x_i-\mu)}{\sigma^2} \tag{184}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum(x_i-\mu)^2}{\sigma^3} \tag{185}$$

$$\Rightarrow \hat{\mu} = \bar{x} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \hat{\mu})^2 = \frac{1}{n}\sum(x_i-\bar{x})^2 \tag{186}$$

(Note here that the parameter $\theta = (\mu, \sigma)$ has two components)

5. $X \sim Po(\lambda)$ then:

$$P(X = x) = \lambda^x \frac{e^{-\lambda}}{x!} \tag{187}$$

$$L(\lambda; x_1, \ldots, x_n) = \lambda^{\sum x_i}\frac{e^{-n\lambda}}{\prod x_i!} \tag{188}$$

$$l(\lambda) = \sum x_i \log\lambda - n\lambda + C \tag{189}$$

$$\frac{\partial l}{\partial \lambda} = \frac{\sum x_i}{\lambda} - n \tag{190}$$

$$\Rightarrow \hat{\lambda} = \bar{x} \tag{191}$$

A.4 Asymptotic Properties of MLEs

Maximum likelihood estimates (mles) have many useful properties. In particular they are asymptotically unbiased and asymptotically normally distributed (subject to some technical conditions), i.e. for large samples they are approximately normally distributed with mean equal to the [unknown] parameter and variance which can be calculated. This allows us to obtain standard errors of mles and so construct confidence intervals for them. In addition they can be used in the construction of [generalized] likelihood ratio tests.

Specifically,

$$\hat{\theta} \xrightarrow{d} N\left(\theta, I^{-1}\right), \tag{192}$$

where I is the Fisher information matrix

$$I = E\left[-\frac{\partial^2 l}{\partial \theta^2}\right] \tag{193}$$

i.e. the expected value of the second derivative of the log-likelihood. If θ is a vector parameter of dimension p then this Fisher Information matrix will be of dimension $p \times p$.

We can find the asymptotic distributions for the examples in the previous section as follows:

1. $X \sim N(\mu, 1)$ then

$$\frac{\partial l}{\partial \mu} = \sum (x_i - \mu) \tag{194}$$

$$\frac{\partial^2 l}{\partial \mu^2} = -n \tag{195}$$

$$\Rightarrow E\left[\frac{\partial^2 l}{\partial \mu^2}\right] = -n \tag{196}$$

$$\Rightarrow \hat{\mu} \xrightarrow{d} N(\mu, n^{-1}) \tag{197}$$

2. $X \sim Exp(\lambda)$ then:

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i \tag{198}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{n}{\lambda^2} \tag{199}$$

$$\Rightarrow E\left[\frac{\partial^2 l}{\partial \lambda^2}\right] = -\frac{n}{\lambda^2} \tag{200}$$

$$\Rightarrow \hat{\lambda} \xrightarrow{d} N\left(\mu, \frac{\lambda^2}{n}\right) \tag{201}$$

In this case we would substitute the mle $\hat{\lambda}$ for λ in the variance term so that we could find confidence intervals i.e.

$$\hat{\lambda} \xrightarrow{d} N\left(\lambda, \frac{\hat{\lambda^2}}{n}\right) \tag{202}$$

3. $X \sim Bin(m,p)$ then:

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{\sum(m-x_i)}{1-p} \tag{203}$$

$$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum x_i}{p^2} - \frac{\sum(m-x_i)}{(1-p)^2} \tag{204}$$

$$\Rightarrow E\left[\frac{\partial^2 l}{\partial p^2}\right] = -\frac{nm}{p} - \frac{nm}{1-p} = -\frac{nm}{p(1-p)} \tag{205}$$

$$\Rightarrow \hat{p} \xrightarrow{d} N\left(p, \frac{p(1-p)}{nm}\right) \tag{206}$$

4. $X \sim N(\mu, \sigma^2)$ then, letting $\theta = (\mu, \sigma)$:

$$\frac{\partial l}{\partial \mu} = \frac{\sum(x_i-\mu)}{\sigma^2} \tag{207}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum(x_i-\mu)^2}{\sigma^3} \tag{208}$$

$$\frac{\partial^2 l}{\partial \theta^2} = \begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma} & \frac{\partial^2 l}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -2\frac{\sum(x_i-\mu)}{\sigma^3} \\ -2\frac{\sum(x_i-\mu)}{\sigma^3} & \frac{n}{\sigma^2} - 3\frac{\sum(x_i-\mu)^2}{\sigma^4} \end{pmatrix} \tag{209}$$

$$\Rightarrow E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = \begin{pmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{\sigma^2} - 3\frac{n}{\sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{2n}{\sigma^2} \end{pmatrix} \tag{210}$$

$$\Rightarrow \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}\right) \tag{211}$$

(Note again that we get a matrix as the parameter $\theta = (\mu, \sigma)$ has two components)

5. $X \sim Po(\lambda)$ then:

$$\frac{\partial l}{\partial \lambda} = \frac{\sum x_i}{\lambda} - n \tag{212}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum x_i}{\lambda^2} \tag{213}$$

$$\Rightarrow E\left[\frac{\partial^2 l}{\partial \lambda^2}\right] = -\frac{n\lambda}{\lambda^2} \tag{214}$$

$$\Rightarrow \hat{\lambda} \xrightarrow{d} N\left(\lambda, \frac{\lambda}{n}\right) \tag{215}$$

Note: In examples (3)-(5) we would substitute the mles for the unknown parameters in the expressions for the variances so that the asymptotic results are useful to find confidence intervals.

A.4.1 Confidence Intervals

Having found the asymptotic distributions of the mles, we can then obtain an approximate 95% confidence interval as

$$\hat{\theta} \pm 2 \times s.e.(\hat{\theta}) \tag{216}$$

A.5 [Generalized] Likelihood Ratio Tests

A useful procedure for constructing hypothesis tests about the value of parameters is the generalised likelihood ratio test. Under suitable technical conditions, the (asymptotically) most powerful test of a composite hypothesis (i.e. one involving unknown parameters) against another can be based on the ratio of the maximized likelihoods, where any unknown parameters are replaced by their mles. Specifically, suppose we have data $x_1, \ldots, x_n$ from a random variable X whose distribution depends on a parameter θ and we wish to test a hypothesis H_0 against an alternative H_A. The likelihood ratio statistic is

$$\zeta = 2\{l(\theta_A) - l(\theta_0)\} \tag{217}$$

where θ_A and θ_0 are the estimates of θ under the hypotheses H_A and H_0 respectively.

We would reject H_0 in favour of H_A if ζ is sufficiently large. For large sample sizes,

$$\zeta \sim \chi^2_r, \tag{218}$$

where r is the difference in numbers of parameters estimated under H_A and H_0.

A.5.1 Examples

Again, suppose we have observed data $x_1, \ldots, x_n$ from the following models:

1. $X \sim N(\mu, 1)$ and we wish to test $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$ then

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(x_i - \mu)^2. \tag{219}$$

Under H_0, μ is set and so $\hat{\mu}_0 = 0$ while under H_A we have no restriction on the value of μ and so $\hat{\mu_A} = \bar{x}$. Hence

$$\zeta = 2\{l(\bar{x}) - l(0)\} = \sum x_i^2 - \sum(x_i - \bar{x})^2 = n\bar{x}^2. \tag{220}$$

We would reject H_A if this is large compared with χ^2_r.

2. $X \sim Exp(\lambda)$ and we wish to test $H_0 : \lambda = \lambda_0$ vs. $H_A : \lambda \neq \lambda_0$ then

$$l(\lambda) = n\log(\lambda) - \lambda\sum x_i. \tag{221}$$

Under H_0, μ is set and so $\hat{\lambda}_0 = \lambda_0$ while under H_A, $\hat{\lambda_A} = \frac{1}{\bar{x}}$. Hence

$$\zeta = 2\left\{l(\frac{1}{\bar{x}}) - l(\lambda_0)\right\} = 2\left\{(-n\log(\bar{x}) - n) - \left(n\log(\lambda_0) - \lambda_0\sum x_i\right)\right\} \tag{222}$$

$$= 2\left\{\lambda_0\sum x_i - n\log(\bar{x}) - n - n\log(\lambda_0)\right\} \tag{223}$$

which we would compare to a χ^2_1 distribution.

3. $X \sim N(\mu, \sigma^2)$ and we wish to test $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$ and σ is unknown. Then:

$$l(\mu, \sigma) = C - n \log \sigma - \frac{\sum(x_i - \mu)^2}{2\sigma^2} \tag{224}$$

Here σ is treated as a nuisance parameter and must be maximised under both H_0 and H_A. Under H_0 we find that the mle for $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0^2) = (0, n^{-1} \sum x_i^2)$ while under H_A we have $\hat{\theta}_A = (\hat{\mu}_A, \hat{\sigma}_A^2) = (\bar{x}, n^{-1} \sum(x_i - \bar{x})^2)$. Following some manipulation we find that

$$\zeta = 2\{l(\hat{\theta}_A) - l(\hat{\theta}_0)\} = \left\{ n \log \left(\sum x_i^2 \right) - n \log \left(\sum (x_i - \bar{x})^2 \right) \right\} \tag{225}$$

to be compared with a χ_1^2 distribution.

A.6 Exercises

Try finding the likelihood ratio statistic for the following:

- $X \sim N(\mu, \sigma^2)$ and we wish to test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_A : \sigma^2 \neq \sigma_0^2$ with μ unknown.
- $X \sim N(\mu, \sigma^2)$ and we wish to test $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_A : \sigma^2 \neq \sigma_0^2$ with μ known.
- $X \sim Bin(m, p)$ and we wish to test $H_0 : p = p_0$ vs. $H_A : p \neq p_0$ with m known.
- $X \sim Po(\lambda)$ and we wish to test $H_0 : \lambda = \lambda_0$ vs. $H_A : \lambda \neq \lambda_0$