Medical Statistics: Survival Data

Nick Fieller

(this page left blank for notes)

Contents

0. Introduction 1
0.1 Books 1
0.2 Objectives 2
0.3 Organization of course material 2
0.4 A Note on R, S-PLUS and MINITAB 3
0.5 Data sets 4
0.5.1 R data sets 4
0.5.2 Data sets in other formats 5
0.6 R libraries required 5
0.7 Outline of Course 6
1 Background and Basic Concepts 6
1.1 Preliminary Discussion 6
1.1.1 Example: Survival of angina pectoris 8
1.2 Censoring 10
1.2.1 Notes 11
1.3 Approaches 12
1.4 Summary 12
2 Single Sample Models 15
2.1 Basic concepts 15
2.1.1 Survivor function 15
2.1.2 Hazard function 16
2.1.2.1 Interpretation: 16
2.1.3 Cumulative hazard function 17
2.2 Typical patterns 18
2.2.2 Example : exponential 19
2.2.3 Example: Weibull 20
2.3 Lifetables 21
2.3.1 Example: no censoring 22
2.3.2 Example: lost to follow up 24
2.3.2.1 Notes: 26
Tasks 1 28
2.4 Kaplan–Meier product limit estimate of S(t) 29
2.4.1 Simple Case, no censoring: 29

2.4.2 Standard Case, censoring: **31**
2.4.3 Example: tumour remission timea **33**
2.4.4 Computer Implementation **35**
2.4.4.1 R 35
2.4.4.2 S-PLUS 37
2.4.4.3 MINITAB 39
2.4.4.4 SPSS 39
2.4.5 Summary of Non-Parametric Methods **39**
Tasks 2 40
2.6 Parametric Models 41
2.6.1 Introduction **41**
2.6.2 Exponential **42**
2.6.2.1 Uncensored data 42
2.6.2.2 Censored Data 43
2.6.3 Example: survival times of lung cancer **46**
2.6.4 Notes **48**
2.6.5★ Refinements: modelling the censoring distribution **49**
2.6.6 Other Distributions **51**
2.6.6.1 Weibull 51
2.6.6.2 Lognormal 51
2.6.6.3 Others 51
2.6.7 Computer Implementation in R **52**
2.6.7.1 Illustration on lung cancer survival times: 53
2.6.7.2 Illustration on tumour remission times: 54
2.8 Summary 56
3 Two-Sample Comparisons 59
3.1 Introduction 59
3.2 Logrank Test (non-parametric) 59
3.2.1 Example: brain tumour survival times **59**
3.2.2 Notes **61**
3.2.3 Computer Implementation **61**
3.2.3.1 R 61
3.2.3.2 S-PLUS 62
3.2.3.3 MINITAB 63
3.2.3.4 SPSS 63
Exercises 1 63
3.3 Parametric Tests 64
3.3.1. M.L.E. Test **64**
3.3.1.1 Example: brain tumour survival times 65
3.3.2 Likelihood Ratio Test **65**
3.3.2.1 Example: brain tumour survival times 66
3.4 Computer Implementation 66

3.4.1 Illustration on brain Tumour times in R 66
3.5 Notes 68
3.6 Summary 70
Tasks 3 71
4 Regression Models 73
4.1 Introduction 73
4.2 Parameteric Regression Models 73
4.2.1 Exponential Regression Model 73
4.2.1.1 Notes 74
4.2.1.2 Example: myelogenous leukemia 75
4.2.2 Computer Implementation 76
4.2.2.1 R 76
4.2.2.2 S-PLUS 78
4.2.2.3 MINITAB 80
4.2.2.4 SPSS 80
4.2.3 Two-Sample Example 80
4.2.4 Notes 81
4.3 Covariates and Prognostic Factors 82
4.3.1 Notes 82
4.3.2 Modelling 83
4.3.3 Exponential Model 84
4.3.4 Other Models 85
4.4 Proportional Hazards Model 86
4.4.1 Notes 86
4.4.2 Parameter Estimation 88
4.4.3 Partial Likelihood Approach 89
4.4.3.1 Notes 92
4.4.4 Example: atrial fibrillation 94
4.4.5 Computer Implementation 98
4.4.5.1 R 98
4.4.5.2 S-PLUS 99
4.4.5.3 MINITAB 100
4.4.5.4 SPSS 100
4.4.6 Estimation of h(t) 100
4.4.7 Model checking 103
4.4.7.1 log–log plots 103
4.4.7.2 Residuals 104
4.4.7.3 Implementation in R 105
4.4.7.4 S-PLUS implementation 106
4.4.8⋆ Time-dependent covariates 106
Tasks 4 108

Exercises 2 109
4.5⋆ Accelerated Failure Time Regression Models 113
4.5.1⋆ Introduction 113
4.5.2⋆ Implementation in R 115
4.5.3⋆ Example 115
4.6 Summary & Conclusions 117
Tasks 5 119
5⋆ Competing Risks 121
5.1⋆ Introduction 121
5.2⋆ Basic terminology 122
5.3★ Estimation of hazard and survivor function 124
5.4⋆ Analysis of effects of covariates 125
5.5⋆ Implementation in R 126
5.5.1 Example on organ transplants 126
Exercises 3 130
Notes & Solutions for Tasks 1 137
Notes & Solutions for Tasks 2 139
Notes & Solutions for Tasks 3 147
Notes & Solutions for Tasks 4 150
Notes & Solutions for Tasks 5 152
Notes & Solutions for Exercises 1 155
Notes & Solutions for Exercises 2 160
Notes & Solutions for Exercises 3 179
APPENDIX 0: Maximum Likelihood Estimation 181
A0.0 Estimation 181
A0.1 Definition 181
A0.2 Examples 182
A0.3 Further properties of MLEs 184
A0.4 Examples: 185
A0.5 [Generalized] Likelihood Ratio Tests 187
A0.5.1 Examples 188

Analysis of Survival Data

0. Introduction

0.1 Books

★Altman, D.G. (1991) *Practical Statistics for Medical Research.* Chapman and Hall.

★Campbell, M. J. (2001) *Statistics at Square Two.* BMJ

★Collett, D. (2014) *Modelling Survival Data in Medical Research (3rd Ed.).* Chapman and Hall.

Cox, D.R. & Oakes, D. *Analysis of Survival Data*. Chapman and Hall.

Crowder, M.J., Kimber, A.C., Sweeting, T.J., & Smith, R.L. (1991) *Statistical Analysis of Reliability Data*. Chapman and Hall.

Everitt, Brian & Rabe-Heskith, Sophia (2001) *Analyzing Medical Data Using S-PLUS*. Springer.

Gross, A.J. & Clark, Y.A. (1975) *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley.

Kalbfleisch, J.D. & Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* Wiley.

Lee, E.T. (1980) *Statistical Methods for Survival Data Analysis.* Wadsworth.

Marubini, E. and Valsecchi, M. G. (1995) *Analysing Survival Data from Clinical Trials and Observational Studies.* Wiley.

Miller, R. Jr. (1984) *Survival Analysis*. Wiley.

Swinscow, T. D. V. (1996) *Statistics at Square One (9th Ed.).* BMJ

★ Indicates texts at the appropriate level for this course

0.2 Objectives

The objective of this book is to provide an introduction to the statistical modelling and analysis of *lifetime data*. Lifetime data arise especially in medical statistics as well as in studies of reliability. A lifetime might refer to a *survival time*, (i.e. time to death of a patient from diagnosis) or *time to recovery* or *remission* of a patient or *time to failure* of an electronic component.

0.3 Organization of course material

The notes in the main Chapters 1– 4 are largely covered in the two highlighted books in the list of recommended texts above and are supplemented by various examples and illustrations. Some background mathematical details and computational aspects (maximum likelihood estimation etc) are outlined in Appendices at the end. A few individual sections are marked by a star,*, which indicates that although they are part of the course they are not central to its main theme.

The expository material is supplemented by simple 'quick problems' (*task sheets*) issued each week and more substantial *exercises.* These task sheets are designed for you to test your own understanding of the course material. If you are not able to complete the tasks then you should go back to the lecture notes (and other course material) and re-read the relevant section (and if necessary re-read again & …). Solutions are provided at the end of the book. Solutions are provided at the end of the book.

0.4 A Note on R, S-PLUS and MINITAB

The main statistical package for this course is **R.** It is very similar to the copyright package S-PLUS and the command line commands of S-PLUS are [almost] interchangeable with those of **R**. Unlike S-PLUS, **R** has only a very limited menu system which covers some operational aspect but no statistical analyses. A brief guide to getting started in **R** is available from the course homepage.

R is a freely available programme which can be downloaded over the web from http://cran.r-project.org/ or any of the mirror sites linked from there for installation on your own machine. It is available on University networks. **R** and S-PLUS are almost identical except that **R** can only be operated from the command line apart from operational aspects such as loading libraries and opening files. Almost all commands and functions used in one package will work in the other. However, there are some differences between them. In particular, there are some options and parameters available in **R** functions which are not available in S-PLUS. Both S-PLUS and **R** have excellent help systems and a quick check with help(*function*) will pinpoint any differences that are causing difficulties. A key advantage of **R** over S-PLUS is the large number of libraries contributed by users to perform many sophisticated analyses. These are updated very frequently and extend the capabilities substantially. If you are considering using multivariate techniques outside this course (e.g. for some other substantial project) then you would be well advised to use **R** in preference to S-PLUS. Command-line code for he more substantial analyses given in the notes for this course have been tested in **R**. In general, they will work in S-PLUS as well but there could be some minor difficulties which are easily resolved using the help system.

0.5 Data sets

Data sets used in this course are available in a variety of formats on the associated course web page available here.

0.5.1 R data sets

Those in **R** are given first and they have extensions **.Rdata**; to use them it is necessary to copy them to your own hard disk. This is done by using a web browser to navigate to the course web, clicking with the right-hand button and selecting 'save target as…' or similar which opens a dialog box for you to specify which folder to save them to. Keeping the default **.Rdata** extension is recommended and then if you use Windows explorer to locate the file a double click on it will open **R** with the data set loaded and it will change the working directory to the folder where the file is located. For convenience all the **R** data sets for Medical Statistics are also given in a WinZip file.

NOTE: It is not possible to use a web browser to locate the data set on a web server and then open R by double clicking. The reason is that you only have read access rights to the web page and since **R** changes the working directory to the folder containing the data set write access is required.

0.5.2 Data sets in other formats

Most of the data sets are available in other formats (Minitab, SPSS etc). It is recommended that the files be downloaded to your own hard disk before loading them into any package but in most cases it is possible to open them in the package *in situ* by double clicking on them in a web browser. However, this is not possible with **R.**

0.6 R libraries required

Most of the statistical analyses described in this book use functions within the `base and stats` packages and the `MASS` package. It is recommended that each **R** session should start with

```
library(MASS)
```

The `MASS` library is installed with the base system of **R** and the `stats` package is automatically loaded.

0.7 Outline of Course

1. Introduction:– types of survival data, censoring, outline of parametric and non-parametric approaches.
2. Single sample methods:– Basic concepts of survivor & hazard functions. Lifetables, (population, cohort and clinical). Kaplan-Meier product limit estimator of the survivor function, including censored data. Parametric models (exponential, Weibull and log-Normal).
3. Two Sample Comparisons:– Log rank test, parametric tests (maximum likelihood and likelihood ratio tests), proportional hazards.
4. Regression Models:– exponential regression, covariates and prognostic factors, exponential and Weibull models. Proportional Hazards Model, outline of estimation procedures by partial likelihood.

1 Background and Basic Concepts

1.1 Preliminary Discussion

The objective of a survival data analysis may be just to describe (and model) a single sample of data to 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***: often the event of interest (e.g. death, of patient, failure of component, recovery of patient) has not occurred by the time of

recording so that all is known is that the lifetime for that subject is *at least* some value (and may well be greater than this value). Such censoring cannot be ignored (i.e. the censored observations cannot just be omitted) since they carry important information about the effectiveness of the treatment (and indeed one hopes that many patients are alive at the end of a medical study!). This introduces a complication in the statistical description and analysis of the data.

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). These patients form part of a large group of patients examined at the Mayo Clinic during the 15 year period January 1, 1927 — December 31, 1941.

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	

This example illustrates: Follow-up study

Grouping

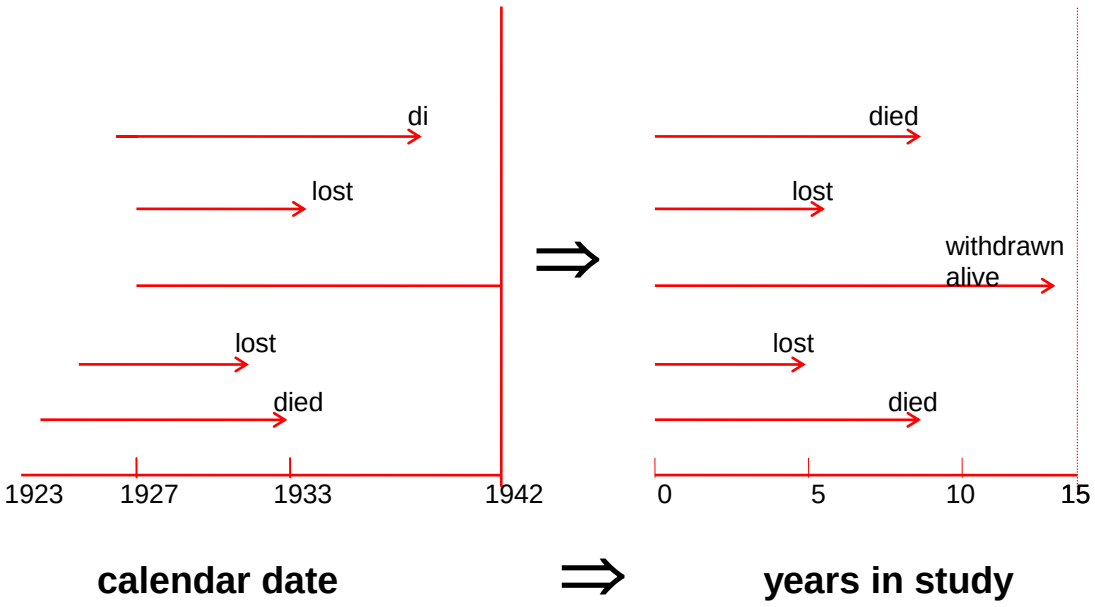
'Lost' patients (censoring)

Change in time axis (which is hidden)

— measure from entry into study.

We might also consider (i) several possible causes of death

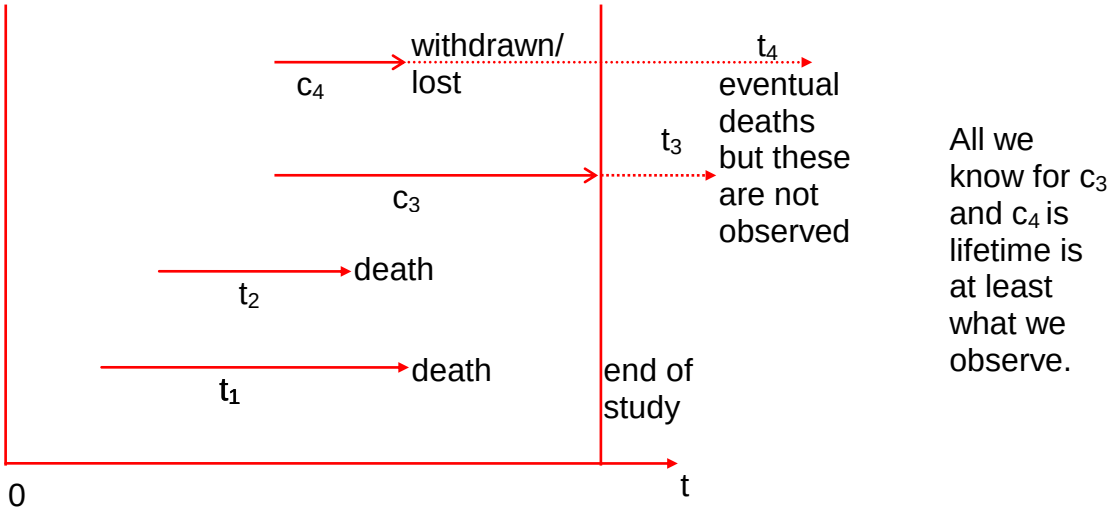
(ii) use of covariates which influence lifetime distributions, e.g. age, sex



1.2 Censoring

This is the complicating feature which identifies the need for a different type of analyses. Sometimes we do not observe the exact lifetime but only know that it exceeds some value (***right censoring***), or even, rarely, only that it is less than some value (***left censoring*)**. There are several common censoring schemes.

For example, 'Time' or 'Progressive' Censoring (as in Example 1)



i.e. individuals are subjected to limited periods $c_1,c_2,....,c_n$ of observation and the i^{th} individual lifetime t_i is observed only if $T_i \leq c_i$.

1.2.1 Notes

(i) **Type I censoring** if identical starting points and subjects are observed for a fixed time c_i (then typically $c_1=c_2=, ...,=c_n$). The number of censorings is then random.

(ii) **Type II censoring** assumes n patients in study at the start and the trial finishes after r deaths (do not specify the end of the trial initially — carry on until r out of n patients are dead, i.e. record $t_{(1)},t_{(2)},...,t_{(r)}$ up to a time $t_{(r)}$ when r have died). This type of censoring is less common in medical studies but is widely used in electronic component testing and reliability studies. The numbering of censorings is not random in Type II censoring but is fixed in advance.

(iii) ***Left censoring*** might occur if subjects are only observed at fixed appointments, and only then is it discovered that death occurred sometime before then, so survival time is *less* than the period of observation. Another example is when the 'lifetime observed' is the time to recurrence of a tumour observable only during surgery.

(iv) ***Interval censoring*** occurs when failure is only known to have occurred during an interval.

(This course will consider only *right censoring* in detail)

1.3 Approaches

Aims:

- estimate lifetime distributions
- predict survival times:–
  - non-parametric — lifetables, Kaplan-Meier
  - parametric — exponential, Weibull.

1.4 Summary

- **Censoring:** lifetime not observed exactly
  - *Right Censoring:* actual lifetime exceeds observation
  - *Left Censoring:* actual lifetime less than observation
  - *Interval Censoring:* observation gives upper and lower bound on lifetime
  - *Type I Censoring:* observation time fixed, # censorings random
  - *Type II Censoring:* # censorings fixed observation time random

This course considers primarily Type 1 right censoring and assumes generally that censoring is ***independent*** of lifetime.

2 Single Sample Models

2.1 Basic concepts

The random variable T measures survival time;

T>0, a continuous variable.

The actual survival time, t, of an individual is the value of the variable T.

T has **p.d.f.** (probability density function) f(t) (t>0)

and **d.f.** (distribution function) $F(t) = P[T \leq t]$

(so $f(t) = F'(t)$ and $F(t) = \int_0^t f(u)du$)

2.1.1 Survivor function

$$S(t) = P[T \geq t] = 1 - F(t).$$

(so $S'(t) = -f(t)$, $S(t) = \int_t^\infty f(u)du$)

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. It is the probability that an individual dies at time t, conditional on (s)he having survived to that time. The hazard function represents the instantaneous death rate for an individual surviving to time t.

$$h(t) = \lim_{\delta t \to 0}\left[\frac{P[t \le T < t + \delta t \mid T \ge t]}{\delta t}\right]$$

(This is also known as the **hazard rate** or **failure rate**.)

2.1.2.1 Interpretation:

Suppose time units are days, Take $\delta t=1$ (small in relation to times considered), $h(t)=P[t \le T<t+1]$, the probability of dying on day t.

2.1.3 Cumulative hazard function

f(t), S(t), H(t) and h(t) are equivalent ways of defining a specific survival pattern uniquely and they are all inter-related.

Clearly, $h(t) = \lim_{\delta t \to 0}\left[\frac{P[t \le T < t + \delta t, T \ge t]}{P[T \ge t]\delta t}\right]$

$$= \lim_{\delta t \to 0}\left[\frac{P[t \le T < t + \delta t}{P[T \ge t]\delta t}\right] = \lim_{\delta t \to 0}\left[\frac{F(t + \delta t) - F(t)}{S(t)\delta t}\right]$$

$$= \lim_{\delta t \to 0}\left[\frac{f(t)\delta t}{S(t)\delta t}\right]$$

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

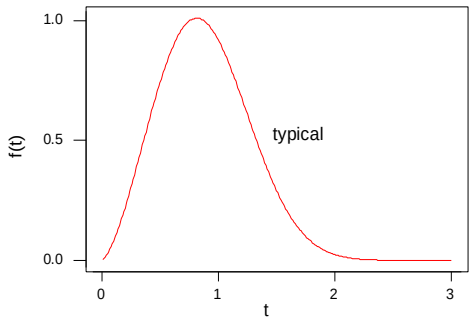
Also $f(t) = F'(t) = -S'(t)$, so $h(t) = -S'(t)/S(t) = -{}^{d\log S(t)}\!/_{dt}$

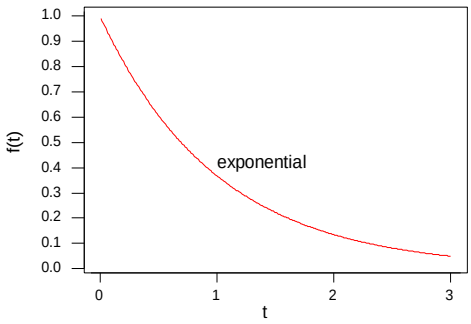
Thus $S(t) = \exp\{-\int_0^t h(u)du\} = \exp\{-H(t)\}$ and

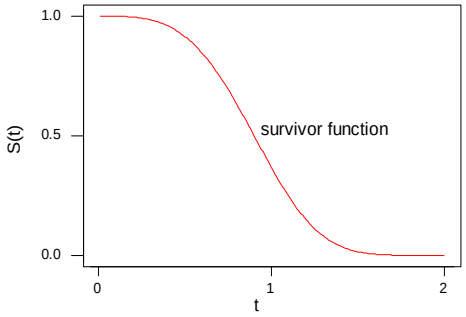
$$H(t) = \int_0^t h(u)du = -\log_e S(t)$$

The survivor function and hazard function are estimated from the observed survival times.

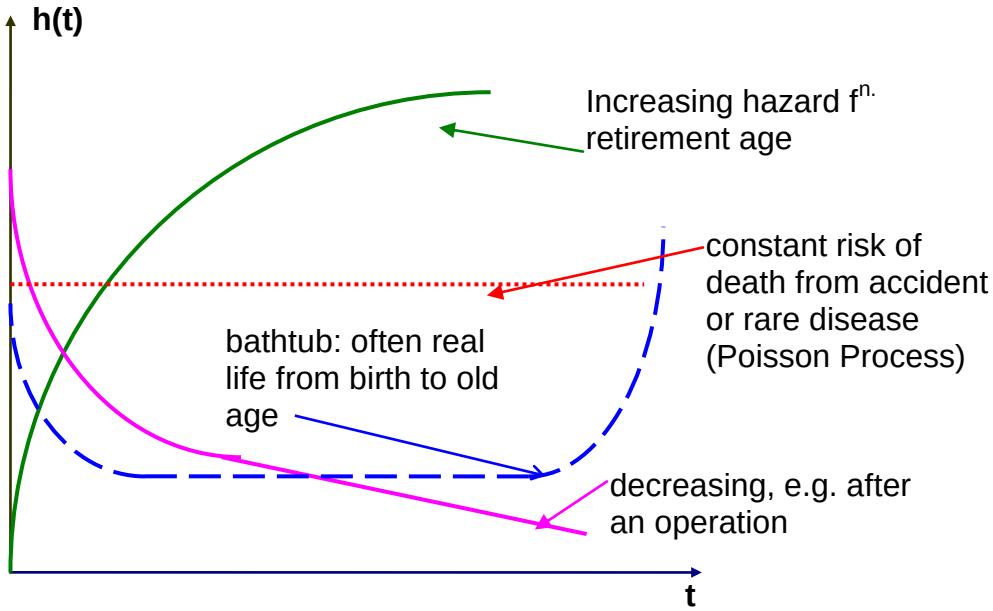
2.2 Typical patterns







Hazard functions:



Often in a practical situation we can guess at the form of the hazard function and so recognise an appropriate family of models to try to

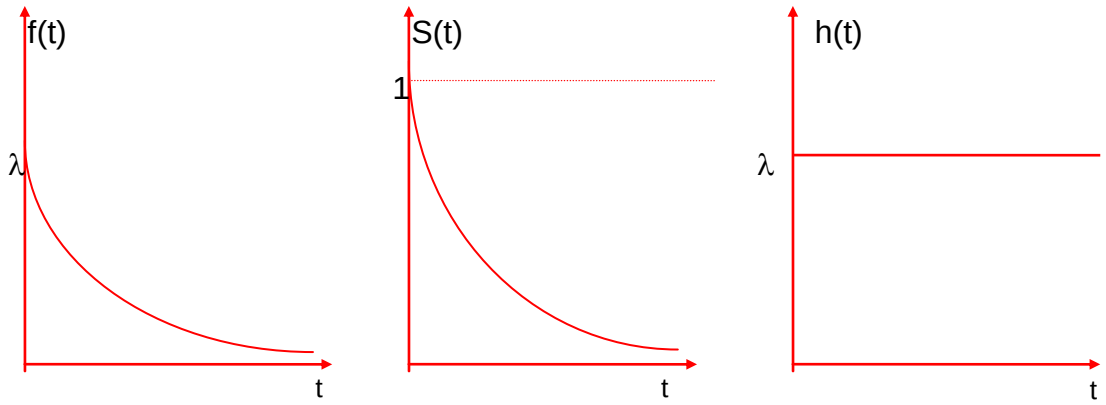
estimate. Sometimes we can determine the general form of the hazard function from an initial investigation of the data.

2.2.2 Example : exponential

Exponential $f(t) = \lambda e^{-\lambda t}$

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

$h(t) = \lambda$



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

2.2.3 Example: Weibull

Weibull: $h(t) = \lambda\gamma t^{\gamma-1}$

$\gamma > 1$: increasing

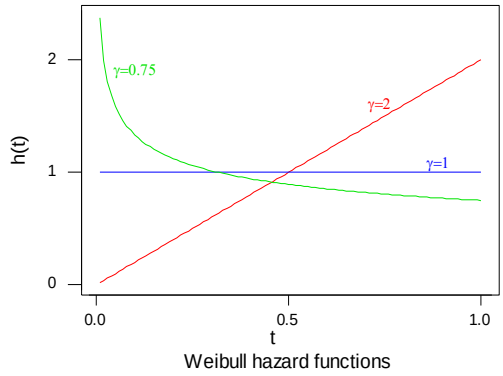
$\gamma = 1$: constant (exponential)

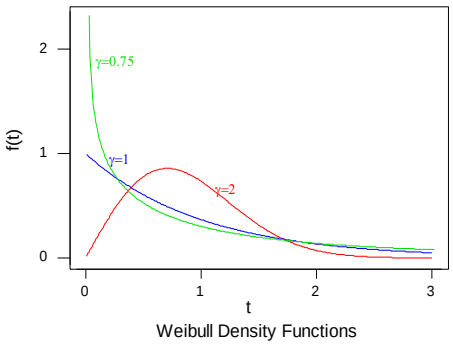
$\gamma < 1$: decreasing

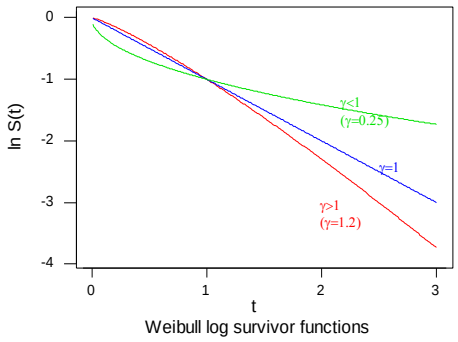
$\therefore \quad S(t) = \exp(-\lambda t^{\gamma})$

$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$

e.g. $\lambda = 1$







Alternative parameterization:

$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$ i.e. $\lambda^{\gamma}\gamma t^{\gamma-1}$

The Weibull distribution provides a very flexible family of survival distributions with both increasing ($\gamma>1$) and decreasing ($\gamma<1$) hazard functions. It can be difficult to estimate, particularly if γ is close to 1.

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.

There are 3 types of lifetable:—

(i) Population (or current)

Obtained from a census or survey. It gives the survival pattern of a group of individuals subject to the age–specific death rates currently observed in the population. It is an artificial population — it gives the pattern of mortality or what would happen if individuals were subjected throughout their lifetime to the present death rates.

(ii) Cohort

Follow a group of individuals throughout their lifetimes.

(iii) Clinical (or follow-up)

Of more relevance in clinical studies — survival pattern of a specific group of individuals. Source of data is usually from a follow-up study.

2.3.1 Example: no censoring

In this study every patient has been followed up after treatment, either until death or up to the end of 1992.

Survival after treatment:

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	
1989	47	45	41	38	73	
1990	58	53	48	115		
1991	42	40	173			
	248	232				

Year after treatment	Prob. of surviving each year	Prob. of dying each year	Lifetable (per 1000) Number alive on each anniversary	Number dying during each year
x	**p_x**	**q_x**	**l_x**	**d_x**
0	0.936	0.064	1000	64
1	0.901	0.099	936	93
2	0.920	0.080	843	67
3	0.948	0.052	776	40
4	0.933	0.067	736	49
5			687	

Censoring: 'withdrawn alive' at the end of the study

Notes: 0.936 = 232/248

0.901 = 173/(232–40)

0.920 = 115/(173–48) etc.

The p_x are calculated from the numbers surviving from one year to the next and strictly are [estimates of] conditional probabilities of surviving for that year, conditional on surviving up until the start of the year.

2.3.2 Example: lost to follow up

Complications begin to arise when patients are lost to follow-up — and we do not know if they have died or not ⇒ considered as 'withdrawn'. (From Armitage, 1971)

Interval since operation years x to x+1	Last reported during this interval: Died d_x	withdrawn w_x	Living at start of interval n_x	Adjusted number at risk n_x'	Estimated probability of death q_x	Estimated probability of survival p_x	% of survivors after x years l_x	Estimate of p.d.f. $\hat{f}_{x+\frac{1}{2}}$	Estimate of hazard function $\hat{h}_{x+\frac{1}{2}}$
0 – 1	90	0	374	374.0	0.2406	0.7594	100	0.241	0.274
1 – 2	76	0	284	284.0	0.2676	0.7324	75.9	0.203	0.309
2 – 3	51	0	208	208.0	0.2452	0.7548	55.6	0.136	0.279
3 – 4	25	12	157	151.0	0.1656	0.8344	42.0	0.070	0.181
4 – 5	20	5	120	117.5	0.1702	0.8298	35.0	0.059	0.186
5 – 6	7	9	95	90.5	0.0773	0.9227	29.1	0.023	0.080
6 – 7	4	9	79	74.5	0.0537	0.9463	26.8	0.014	0.055
7 – 8	1	3	66	64.5	0.0155	0.9845	25.4	0004	0.016
8 – 9	3	5	62	59.5	0.0504	0.9496	25.0	0.013	0.052
9 – 10	2	5	54	51.5	0.0388	0.9612	23.7	0.009	0.040
10 –	21	26	47	—	—	—	22.8		

d_x: number died during (x, x+1)

w_x: includes those who have disappeared (i.e. last report last year)
+'withdrawn alive'

n_x: number living at start of interval (x,x+1) $\rightarrow$ accumulate d_x+w_x
from bottom.

n_x': assume withdrawals are uniformly spread over interval
$\Rightarrow n_x' = n_x - \frac{1}{2}w_x$

q_x: conditional probability of dying in (x,x+1), $q_x=d_x/n_x'$

p_x: $=1-q_x$

l_x: life survival rates, $l_0=100$; $l_x=l_0p_0p_1.....p_{x-1}$

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

$$\hat{f}_{x+\frac{1}{2}} := \hat{S}_x - \hat{S}_{x+1} = \hat{S}_x q_x \quad x=0,1,2,..$$

$$\hat{h}_{x+\frac{1}{2}} := \frac{2q_x}{1+p_x}$$

2.3.2.1 Notes:

(i) 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).

(ii) 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 ½w_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.

(iii) We have assumed that p_x and q_x remain constant over the study. This is a relatively short period.

(iv) Difficult to calculate expectations of life with censored data and we begin to think in terms of a parametric model.

$$\hat{S}_x = {}^{l_x}\!/_{100}; \quad \hat{f}_{x+\frac{1}{2}} = \hat{S}_x - \hat{S}_{x+1}; \quad \hat{h}_{x+\frac{1}{2}} = \frac{\hat{f}_{x+\frac{1}{2}}}{P[T \geq x + \frac{1}{2}]} = \frac{\hat{f}_{x+\frac{1}{2}}}{\frac{1}{2}(\hat{S}_x + \hat{S}_{x+1})}$$

Clinical life tables can help us decide what the hazard function might look like

Note: These estimates are subject to sampling error:

Greenwood (1926) showed that approximately

$$\text{Var}(\hat{S}_x) = \hat{S}_x^2 \sum_{j=1}^{x-1} \frac{d_j}{n_j(n_j - d_j)}$$

Tasks 1

1) Derive a clinical life table for [at least the first five years of] the survival data of patients with angina pectoris given in Example 1 in the notes and reproduced below.

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	

2) Show that the probability that an individual lives longer than t1+t2 years given he has attained t1 years is equal to the unconditional probability that he survives at least t2 years if and only if the survival distribution is of exponential form.

   [Suggestion: obtain the equation P[T>t1+t2]=P[T>t1]P[T>t2] and then take logs, noting the well-known the result that:—

   if g(t1+t2)=g(t1)+g(t2) for all t1 and t2, then g(t)=αt for some real α.]

2.4 Kaplan–Meier product limit estimate of S(t)

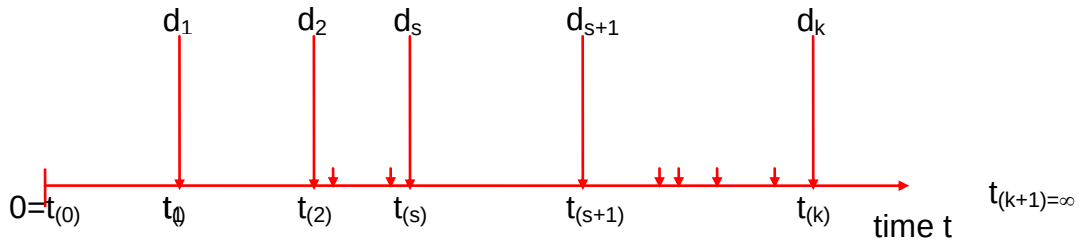
The lifetable methods all consider the data in groups. If the actual lifetimes (perhaps censored) are available, grouping will lose information.

2.4.1 Simple Case, no censoring:

n observations of lifetimes at $t_1,t_2,\ldots,t_n$,

⇨ order: $t_{(1)}<t_{(2)}<..<t_{(k)}$ (assuming k distinct lifetimes)

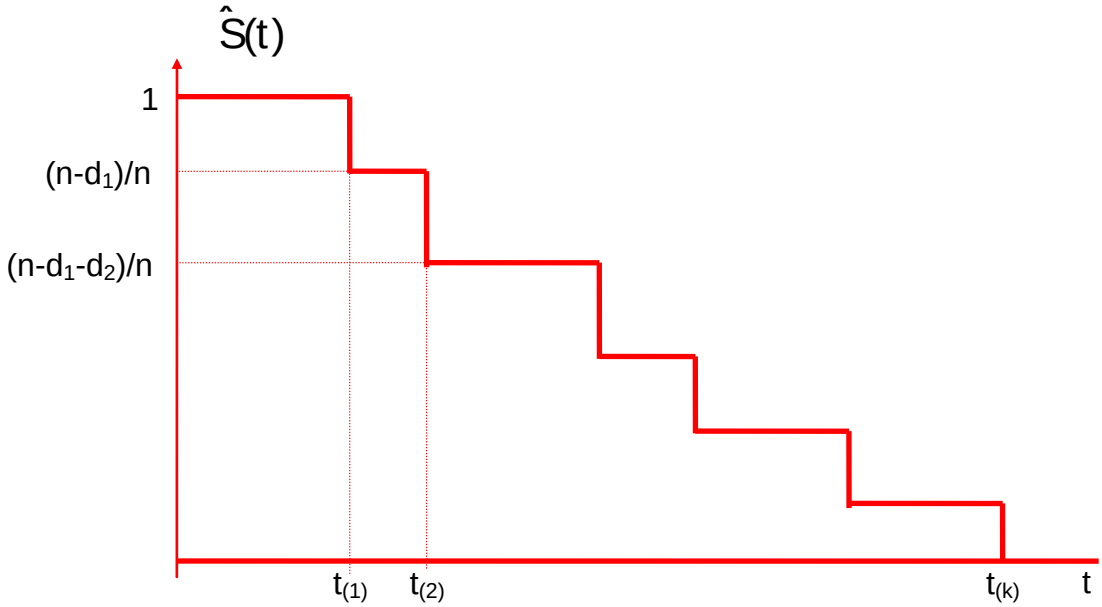
Let d_i be the number of deaths at $t_{(i)}$ (so $\Sigma d_i=n$)



$\hat{F}(t)$ = proportion of lifetimes < t

$$= \frac{1}{n}\sum_{j=1}^{s} d_j \quad \text{for} \quad t_{(s)} \le t < t_{(s+1)}$$

$$\hat{S}(t) = 1 - \hat{F}(t) = \frac{n - \sum_1^s d_j}{n} \quad \text{for} \quad t_{(s)} \le t < t_{(s+1)}$$



Let r_j be the number at risk (≡number alive) just before $t_{(j)}$,

Then $r_{j+1}=r_j-d_j$,

So $$\hat{S}(t) = \frac{n-d_1}{n}.\frac{n-d_1-d_2}{n-d_1}.\frac{n-d_1-d_2-d_3}{n-d_1-d_2}\ldots\frac{n-d_1-d_2-\ldots-d_s}{n-d_1-\ldots-d_{s-1}}$$

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

$$= = \prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \text{ for } t_{(s)} \le t < t_{(s+1)}$$

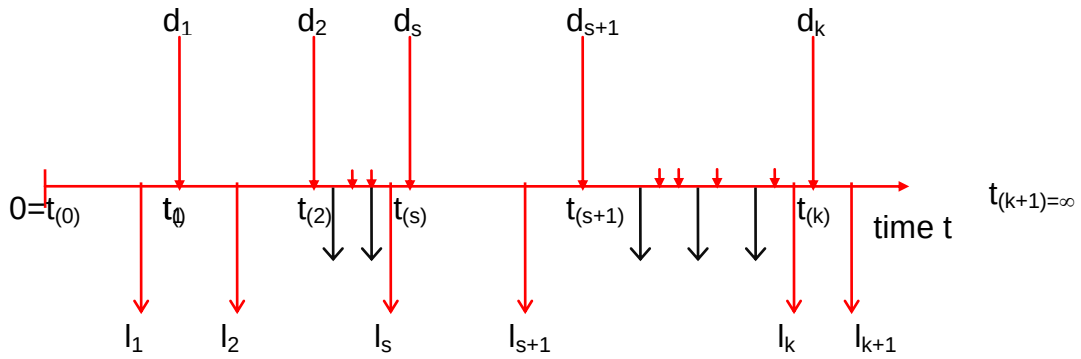
2.4.2 Standard Case, censoring:

Kaplan & Meier suggested using the same type of estimate based on a product when we have censoring.

$t_{(1)}<t_{(2)}<...<t_{(k)}$ are the k distinct lifetimes

d_1 d_2 ... d_k number of lifetimes = $t_{(j)}$

l_1 l_2 l_3 .... l_k numbers censored between times $t_{(j-1)}$ and $t_{(j)}$.



l_j = number censored in the previous interval

= number with observed times $c_1,c_2,...,c_n$;

now $r_1=n-l_1$; $r_{j+1}=r_j-d_j-l_{j+1}$ for j=1,2,...,k-1.

[or $r_j=n-(d_1+d_2+...+d_{j-1})-(l_1+l_2+...+l_j)$ for $j\geq 2$]

So we have the Kaplan-Meier product limit

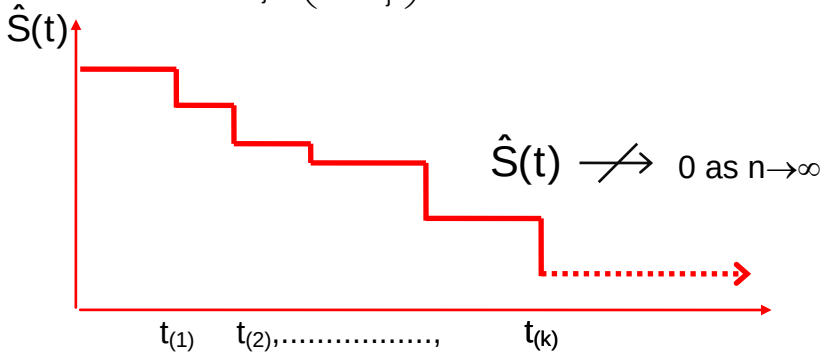
$$\hat{S}(t)=\prod_{j=1}^{s}\left(1-\frac{d_j}{r_j}\right) \text{ for } t_{(s)}\leq t<t_{(s+1)}$$

2.4.2.1 Notes

(i) Assumes that the l_j censorings survive up to $t_{(j)}$ and then are removed. Not that this is slightly different from the adjustment usually used in life tables. The difference is that 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.

(ii) Uncensored case is just a special case with $l_j=0$ all j

(iii) If $l_{k+1}>0$ then $\hat{S}(t) = \prod_{j=1}^{k}\left(1-\frac{d_j}{r_j}\right) > 0$ since $r_k>d_k$



(iv) Again $\hat{S}(t)$ is subject to sampling error.

Greenwood gives $\text{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j=1}^{s} \frac{d_j}{r_j(r_j-d_j)}$ for $t_{(s)}\le t<t_{(s+1)}$

(v) Similarly estimate H(t) by $\hat{H}(t)= -\log_e \hat{S}(t)$

slightly simpler estimate is to use $\tilde{H}(t) = \sum_{j=1}^{s} \frac{d_j}{r_j}$ for $t_{(s)}\le t<t_{(s+1)}$

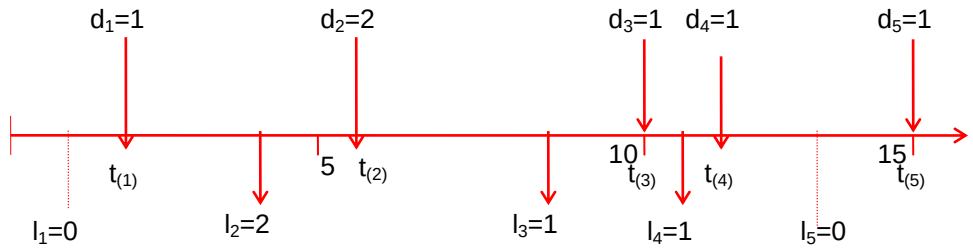
2.4.3 Example: tumour remission timea

Remission times for 10 patients with tumours

6 relapse after 3.0, 6.5,6.5,10,12,15 months

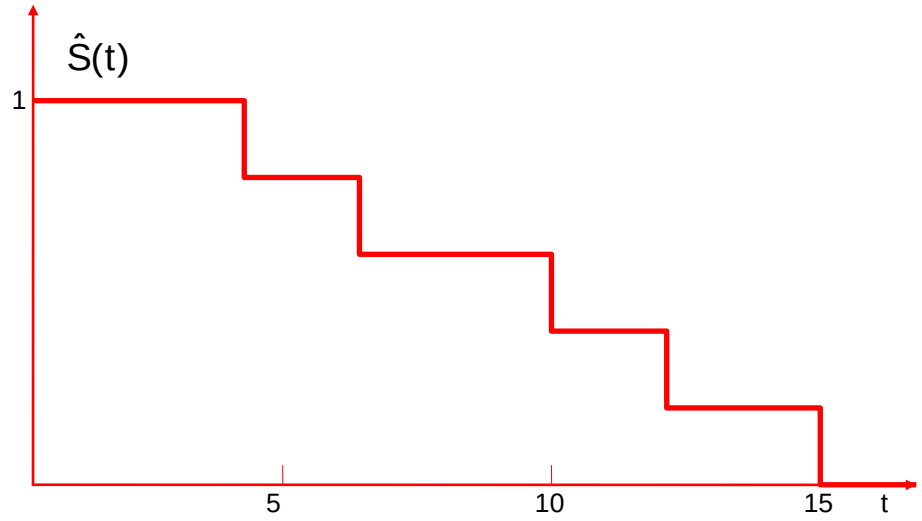
1 lost to follow-up at 8.4 months

3 still in remission at end of study after 4.0, 5.7, 10.1 months



(see data tumour.Rdata)

j	$t_{(j)}$	l_j	r_j	d_j	$\hat{S}(t)$		notes
					1	$0 \le t < 3.0$	
1	3.0	0	10	1	0.9	$3.0 \le t < 6.5$	9/10
2	6.5	2	7	2	0.643	$6.5 \le t < 10.0$	9/10x5/7
3	10.0	1	4	1	0.482	$10.0 \le t < 12.0$	9/10x5/7x3/4
4	12.0	1	2	1	0.241	$12.0 \le t < 15.0$	9/10x5/7x3/4x1/2
5	15.0	0	1	1	0	$15 \le t$	



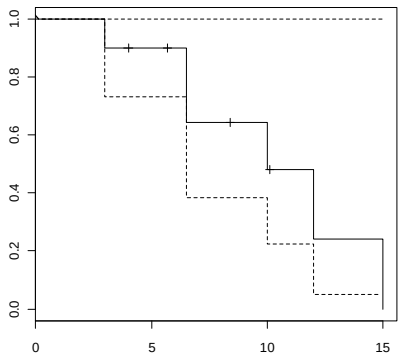
2.4.4 Computer Implementation

2.4.4.1 R

In **R** the functions for analysing survival data are provided in a package called `survival`. It is necessary to load this package with `library(survival)` before any of the commands below can be used. This package is bundled with the base system together with `MASS` etc so there is no need to download it separately from the CRAN site and install it. The first step is create a 'survival object' with the function `Surv()` (note the capitalization). The 'survival object' produced by `Surv()` will be used as the response variable in fitting a model. It contains the information on which observations are censored and which are fully observed events. The next step is to estimate the survivor curve with the function `survfit()`. Technically this step models the survival object produced by `Surv()` on a constant response. Essentially this is regressing the actual survival times on a constant response but making appropriate allowance for the censoring of some observations (information contained in the object produced by `Surv()`). This contains all the information needed for producing Kaplan-Meier plots (including the actual Kaplan-Meier estimate of the survivor function) and in assessing any model that has been fitted. Using the generic function `plot()` will produce the Kaplan-Meier plot, the function `summary()` will give the actual estimate of the survivor function and other details of the model fitting.

This is illustrated below on the data set of tumour remission times. It is necessary to make sure that you have downloaded the data set to a folder on your own hard disk and made this the working directory for your **R** session, perhaps navigating to it using `File>Change dir…` in the **R** menu. Alternatively you need to give the full pathname when loading the data set with something such as `load("C:\\...\\My Documents\\...\\tumour.Rdata")` or navigate to the file with `File>Load Workspace...` from the menus.

```
> library(survival)
Loading required package: splines
> load("tumour.Rdata")
> attach(tumour)
> tumour
        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(time, censor, type = "right")
> tumourSurv <-survfit(tumour.sv ~ 1, data=tumour)
> summary(tumourSurv)
Call: survfit(formula = tumour.sv, 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
> plot(tumourSurv)
```



It is not necessary to separate all the steps, they can be nested into one command:

```
tumourSurv<-survfit(Surv(time,censor,type = "right")~1, data=tumour)
```

The `help()` system will give more details. Line styles and colours can be changed in the `plot()` command. Type `help(par)` to find out more.

2.4.4.2 S-PLUS

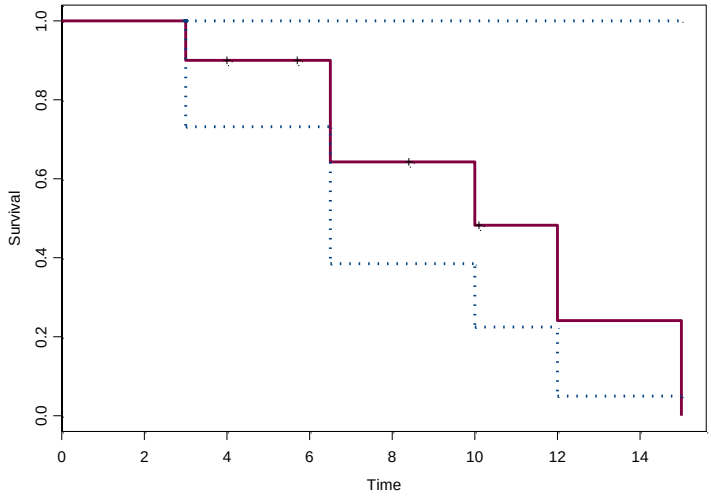
Kaplan-Meier plots are available via the menus in

Statistics>Survival>Nonparametric Survival…

First you need to create a formula by clicking the appropriate button and then select the appropriate variables for Time 1 and Censor codes by highlighting the variables in the Choose Variables box. Then click the Add Response button and you should have a formula of the form `Surv(time,censor,type='right')~1` in the formula box. Click OK and then choose appropriate Options, Results and Plots by clicking on these tabs. With Long Output in results you obtain

```
                *** Nonparametric Survival ***
Call: survfit(formula = Surv(time, censor, type = "right") ~ 1, data
= tumour, na.action = na.exclude, conf.int = 0.95,
     se.fit = T, type = "kaplan-meier", error = "greenwood",
conf.type = "log", conf.lower = "usual")

  n events mean se(mean) median 0.95LCL 0.95UCL
 10      6 10.1     1.39     10     6.5      NA
Call: survfit(formula = Surv(time, censor, type = "right") ~ 1, data
= tumour, na.action = na.exclude, conf.int = 0.95,
     se.fit = T, type = "kaplan-meier", error = "greenwood",
conf.type = "log", conf.lower = "usual")

 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      NA           NA           NA
```

Note that the default option is to provide confidence intervals which necessarily have upper and lower limits of 1 and 0 which are plotted in the graphical output (if selected) and that the censoring times are indicated.



Kaplan-Meier plot from S-PLUS

Note that you can produce the analysis with from the command line with the function `survfit(.)` and an example of the call statement is given above. This function does not itself produce the graph but it produces an *object* (in the S-PLUS sense) which can then be plotted using the generic `plot(.)` function.

```
> tumourSurv<-survfit(Surv(time, censor, type = "right") ~1,
                                                    data=tumour)
> plot(tumourSurv)
```

The `help()` system will give more details. Note that this is identical to the **R** command line version.

2.4.4.3 MINITAB

Kaplan-Meier plots are available through the menus:

`Stat>Reliability/Survival>Nonparametric Dist Analysis-Right Censoring`

It is necessary to specify the value indicating censored observations, i.e. which value indicates that the observation is censored.

2.4.4.4 SPSS

Kaplan-Meier plots are available through the menus:

`Analyze>Survival>Kaplan-Meier`

It is necessary to specify the value indicating uncensored values and the graphical output indicates the censored values as with S-PLUS.

2.4.5 Summary of Non-Parametric Methods

- Life tables
  - Grouped data
  - Allow for censoring by adjusting # at risk
- Kaplan-Meier
  - Individual data
  - Uncensored case = 1 – empirical CDF
  - Express as a product of terms $[1 - d_i/r_i]$
  - Censored case by adjusting # at risk r_i

Tasks 2

1) The data below give the times of remission (in weeks) of two groups of leukaemia patients randomized to a treatment or a control group.

1	drug-6-MP	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*. [* indicates a censored value]
2	control	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

   a) Obtain (by hand and by computer package) and plot the Kaplan-Meier survivor functions for the data (obtaining separate functions for control and drug patients).
   b) Estimate the median survival times for the two groups.

   (The data are given in file *leukaemia remission times*)

2) In an Institute for Medical Research and Public Health in Australia a study was reported in 2005 in which the survival of teaspoons was investigated. 102 teaspoons were purchased and discreetly numbered, 16 of these were of higher quality that the other 86. Equal numbers of teaspoons of each type were placed in eight tearooms around the institute, with equal numbers in communal rooms and programme-linked rooms. Audits were taken at various times during the following five months and the day on which a teaspoon went missing was recorded. The data are given in the dataset spoons.Rdata, with variables indicating day of disappearance, category of tearoom (1 for communal room) and type of teaspoon.

a) Plot the Kaplan-Meier estimates of the survival times of teaspoons and estimate the median survival times in the two categories of rooms.

2.6 Parametric Models

2.6.1 Introduction

Lifetime T, p.d.f. f(t) with t>0, d.f. F(t),
survival function S(t)=1 – F(t), 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.) whose justification relies on asymptotic properties (i.e. large samples). Some details of likelihoods, maximum likelihood estimation and likelihood ratio tests are given in the Appendix 0.

In summary, the likelihood of a parameter θ for data $x_1,\ldots,x_n$ is 'the probability of observing the data $x_1,\ldots,x_n$' . this probability is calculated in terms of the unknown quantity θ and so will be a function of it, $L(\theta)$ say. We can now maximize $L(\theta)$ wrt θ (by differentiating wrt θ and setting = 0) and the value that produces the maximum, $\hat{\theta}$ say, is the maximum likelihood estimate of θ. It can be thought of as the '*most probable'* value of θ in the light of the data just obtained.

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

2.6.2 Exponential

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

$S(t)=e^{-\lambda t}$ $(=1–F(t))=1–(1–e^{-\lambda t}))$

$h(t)=\lambda$

2.6.2.1 Uncensored data

Observe $t_1,t_2,...,t_n$

$Lik(\lambda;t_1,t_2,...,t_n)=L(\lambda)=\Pi f(t_i)=\lambda^n e^{-\lambda\sum_1^n t_i}$

$Log_e(L)=\ell(\lambda)=n\log(\lambda)-\lambda\Sigma t_i$

$$\Rightarrow\hat{\lambda}=\frac{n}{\sum_1^n t_i}$$

<u>Confidence Interval for λ</u>

$Y=\sum_1^n T_i \sim \Gamma(n,\lambda)$ with p.d.f. $f(y)=\lambda^n y^{n-1}e^{-\lambda y}/\Gamma(n)$.

If we let $Z=2\lambda Y$ then Z has p.d.f. $f(z)=(½)^n z^{n-1}e^{-½z}/\Gamma(n)$

$$\text{i.e. } Z \sim \chi^2_{2n}$$

So $P[\chi^2_{2n;\alpha/2} < 2\lambda\sum T_i < \chi^2_{2n;1-\alpha/2}]=1-\alpha$

and so a $100(1-\alpha)\%$ confidence interval for λ given by

$$\left(\frac{\chi^2_{2n;\alpha/2}}{2\sum t_i}, \frac{\chi^2_{2n;1-\alpha/2}}{2\sum t_i}\right)$$

Similarly, MLE of S(t) is $\hat{S}(t)=e^{-\hat{\lambda}t}$

2.6.2.2 Censored Data

'Time' censored, n patients, (potential lifetimes i.i.d. Ex(λ)).

We observe either the lifetime t_i or the fact that $t_i > c_i$, for each individual, (i=1,2,...,n). The simplest case is to assume that the c_i are fixed and given, i.e. non-random

Thus the contribution of each individual to the likelihood is either

$\lambda e^{-\lambda t_i}$ (if $t_i \le c_i$) ("=P[$T_i=t_i$]")

or

$e^{-\lambda c_i}$ (if $t_i > c_i$) ("=P[$T_i>c_i$]")

Define $\delta_i=1$ if $t_i \le c_i$ (i.e. uncensored) and $\delta_i=0$ if $t_i > c_i$ (i.e. censored)

Then Likelihood = $L(\lambda) = \prod_{i=1}^{n} [\lambda e^{-\lambda t_i}]^{\delta_i} [e^{-\lambda c_i}]^{1-\delta_i}$

so $\log_e$[lik (λ)]= $\ell(\lambda) = \log_e \lambda \Sigma \delta_i - \lambda \Sigma t_i \delta_i - \lambda \Sigma (1-\delta_i) c_i$

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{\sum \delta_i}{\lambda} - (\sum \delta_i t_i + (1-\delta_i) c_i)$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_1^n \delta_i}{\sum_1^n \{\delta_i t_i + (1-\delta_i) c_i\}} \quad \text{(noting } \frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} > 0 \text{)}$$

However, the exact distribution of $\hat{\lambda}$ is not now straightforward. Instead we have to use the asymptotic properties of maximum likelihood estimates, i.e.

$$\hat{\lambda} \approx N\left(\lambda, \left\{-E\left[\frac{\partial^2 \ell}{\partial \lambda^2}\right]_{\lambda=\hat{\lambda}}\right\}^{-1}\right)$$

Now $\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{-\sum \delta_i}{\lambda^2}$ and $E[\delta_i]=1.P[T_i \leq c_i] + 0.P[T_i>c_i]$

$$=1.(1- e^{-\lambda c_i}) + 0.\ e^{-\lambda c_i}$$

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

(it is implicit here that the c_i are considered non-random).

So $\text{var}(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_1^n (1-e^{-\hat{\lambda} c_i})}$.

[Alternatively, $\hat{\lambda} \approx N\left(\lambda, \left\{-\left[\frac{\partial^2 \ell}{\partial \lambda^2}\right]_{\lambda=\hat{\lambda}}\right\}^{-1}\right)$, giving $\text{var}(\hat{\lambda}) \approx \frac{\hat{\lambda}^2}{\sum_1^n \delta_i}$.]

A 100(1-α)% Confidence Interval for λ is $\hat{\lambda} \pm z_{1-\frac{1}{2}\alpha} \times$ s.e.($\hat{\lambda}$) where s.e.($\hat{\lambda}$) is the standard error of $\hat{\lambda}$, i.e. $\sqrt{}$var($\hat{\lambda}$).

Interest may be in other aspects

e.g. $\mu=\lambda^{-1}=E[T]$, the mean lifetime or

the age S_α beyond which $100\alpha\%$ survive.

For these we use the result that

$$\text{var}\{g(\hat{\lambda})\} \approx \left[[g'(\lambda)]^2 \, \text{var}(\hat{\lambda}) \right]_{\lambda=\hat{\lambda}}$$

where g(.) is any [differentiable, monotonic] function.

So $\hat{\mu} = \dfrac{1}{\hat{\lambda}}$ and $\text{var}(\hat{\mu}) \approx \dfrac{\hat{\mu}^2}{\sum_1^n (1 - e^{-c_i/\hat{\mu}})}$ or $\text{var}(\hat{\mu}) \approx \dfrac{\hat{\mu}^2}{\sum_1^n \delta_i}$

and for S_α we have $\alpha = P[T \geq S_\alpha] = S(S_\alpha) = e^{-\lambda S_\alpha}$,

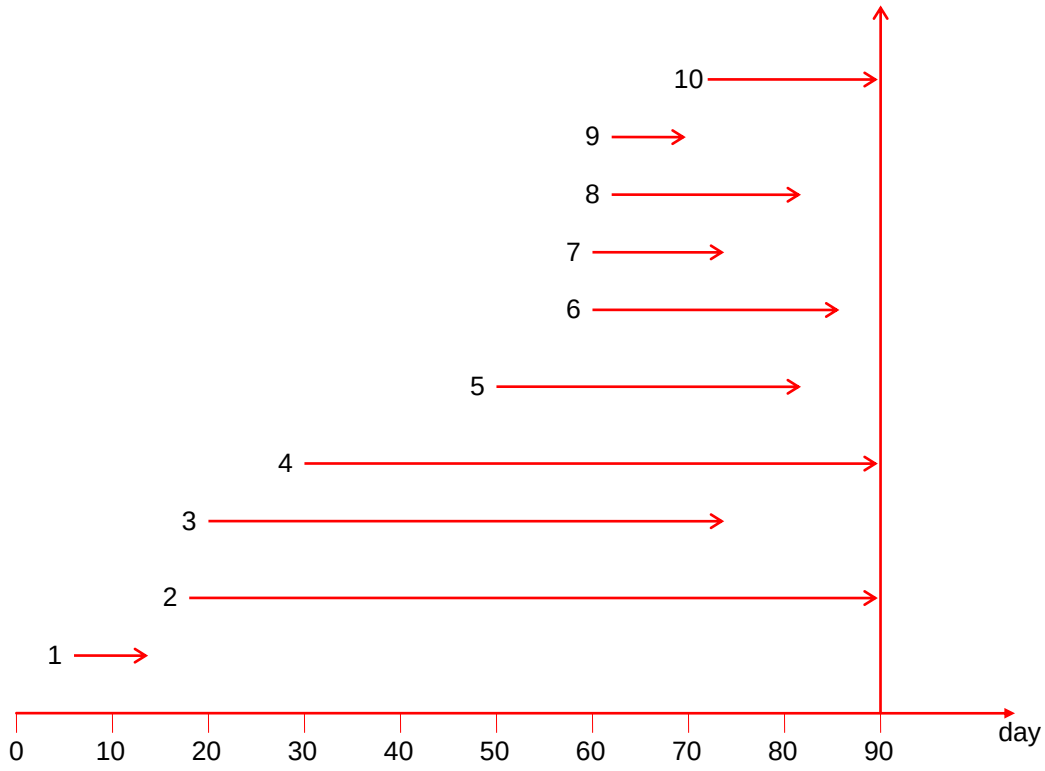
so $S_\alpha = -\lambda^{-1}\log_e(\alpha)$ and $\hat{S}_\alpha = -\hat{\lambda}^{-1}\log_e(\alpha)$

with $\text{var}(\hat{S}_\alpha) = \text{var}(-\hat{\lambda}^{-1}\log_e(\alpha)) = [-\log_e(\alpha)]^2 \, \text{var}(\hat{\lambda}^{-1})$

2.6.3 Example: survival times of lung cancer

Survival times (in days) of 10 patients with advanced lung cancer. Study terminated after 90 days, (see data set lcancer.Rdata).

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



Thus $\Sigma\delta_i = 7$ deaths on study.

$\Sigma\delta_i t_i = 155$, $\Sigma(1-\delta_i)c_i = 153$

thus $\hat{\lambda} = \frac{7}{308} = 0.0227$ per day

$\hat{\mu} = \frac{308}{7} = 44.0$ days

s.d.$(\hat{\lambda})$=0.00859

and so a 95% C.I. for λ is $\hat{\lambda} \pm 1.96\,\text{s.d.}(\hat{\lambda}) \Rightarrow$ (0.00586, 0.0395)

2.6.4 Notes

(i) If we have data $t_1,t_2,...,t_n$ (—observation times)

$\delta_1,\delta_2,...,\delta_n$ (—censoring indicators)

where the lifetime T has a distribution depending on a parameter θ

$$L(\theta) = \prod_{\text{deaths}} f(t_i) \prod_{\text{censored}} S(t_i)$$

$$= \prod_{\text{deaths}} h(t_i)S(t_i) \prod_{\text{censored}} S(t_i) \qquad \text{[since } h(t) = f(t)/S(t) \text{]}$$

$$= \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i) \text{ remembering that some of the } t_i\text{'s}$$

correspond to censoring.

In this notation, for exponential lifetimes, mean λ^{-1}, we have

$$\hat{\lambda} = \frac{\sum_1^n \delta_i}{\sum_1^n t_i} = \frac{\text{total number of deaths observed}}{\text{total time alive of all patients in the study}}$$

(ii) for Type II censoring (i.e. wait until r deaths) it can be shown that

$L_{II}(\theta) = \frac{n!}{(n-r)!} L(\theta)$ $\Rightarrow$ exactly the same estimates as before.

2.6.5★ Refinements: modelling the censoring distribution

A full treatment here is beyond the scope of these notes but a summary of the simplest approaches and results are stated with derivation. The simplest case to consider is the common situation where observations are collected over a fixed time 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.

As before, let δ_i be the censoring indicator taking values 0 or 1 but now it can be shewn that the maximum likelihood estimate of λ is the solution of the equation

$$\sum_i \left(\log(\lambda)\delta_i + \lambda t_i + \lambda(1-\delta_i)T_i\right) = 0\,,$$

where T_i is the total time on the study if lost to follow up.

$$E[\delta_i] = 1 - \frac{1-\exp(-\lambda T)}{\lambda T}$$

and that then the standard error of $\hat{\lambda}$ becomes (c.f.§2.6.2.2)

$$\frac{\hat{\lambda}}{n^{1/2}\left(1-\frac{1-\exp(-\hat{\lambda}T)}{\lambda T}\right)^{1/2}}$$

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

$$E[\delta_i] = 1 - \frac{1-\exp(-\lambda(T-R)) - \exp(-\lambda T)}{\lambda R}$$

with similar modifications to the likelihood equation and standard error.

If further, there is loss to follow up in addition to the censoring caused by a fixed length of study at an exponential rate, with rate η then we have

$$E[\delta_i] = 1 - \frac{\exp(-(\lambda+\eta)(T-R)) - \exp(-(\lambda+\eta T)}{(\lambda+\eta)R}$$

2.6.6 Other Distributions

2.6.6.1 Weibull

See §2.2.1: gives a flexible range of hazard functions.

2.6.6.2 Lognormal

(Beware that estimation can be unstable if there are short lifetimes)

$\log(T) \sim N(\mu,\sigma^2)$, $f(t)=(2\pi)^{-½}(\sigma t)^{-1}\exp[-½(\log(t)-\mu)^2/\sigma^2]$

$F(t)=\Phi\{(\log(t)-\mu)/\sigma\}$, $S(t)=1-\Phi\{(\log(t)-\mu)/\sigma\}$

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

2.6.6.3 Others

e.g. gamma, Gumbel, mixtures — most of these require numerical evaluation of the hazard function.

2.6.7 Computer Implementation in R

Estimation of these models an be performed in **R,** S-PLUS and MINITAB but not directly in SPSS. Here we give only guidance on implementation in **R.** The basic function is `survreg()` and one of the arguments specifies which distribution amongst 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. The help system describes how to do this.

Care needs to be taken with the parameterization in `survreg()`. Firstly, the time variable is incorporated as log(time), so for example when extracting the mean survival time the result from `survreg()` is actually the logarithm of the mean survival time. Next, in the Weibull model (in the usual parameterization), the shape parameter is given as the reciprocal of the `survreg()` intercept and the logarithm of the scale parameter is given as the intercept in `survreg()`. For the exponential model with rate parameter λ (or mean λ^{-1}) the maximum likelihood estimate of λ is given as 1/exp(intercept). It is perhaps easiest to calculate a confidence interval for the intercept (which actually is for the log(mean survival time) and then transform this to a confidence interval for the true value of λ.

2.6.7.1 Illustration on lung cancer survival times:

As with calculation of the non-parametric Kaplan-Meier estimate the first step is to calculate a survival object which contains the information on censoring. This object can then be used in fitting any of the available survival regression models. To illustrate this on fitting an exponential distribution to the lung cancer survival times given in lcancer.Rdata, first the data needs to be loaded and attached:

```
> load("lcancer.Rdata")
> attach(lcancer)
> time
 [1]  2  4 14 21 24 27 33 51 60 72
> sum(time)
[1] 308
> sum(censor)
[1] 7

> 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)  3.78      0.378 10.0 1.35e-23

Scale fixed at 1

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

Thus the estimate of λ is 1/exp(intercept) = 1/exp(3.78) = 0.0228 and a 95% confidence interval for λ is

$1/\{\exp(3.78 \pm 1.96 \times 0.378)\} = (0.0108, 0.0479)$ (compare §2.6.3)

These values are rather different from calculating the approximate standard error using the formula given towards the end of §2.6.2.2 .

This would give an approximate standard error of the estimate of λ as 0.378/exp(3.78) = 0.008627 and a confidence interval of (0.00589, 0.0389). These illustrate that calculations of standard errors and confidence intervals can only be approximate, most especially for such small illustrative data sets and none is particularly 'more accurate' than any other and the differences apparent here are of little practical importance. If using **R,** as in most practical cases would be the case, then the results from **R** are perfectly adequate.

2.6.7.2 Illustration on tumour remission times:

As with calculation the non-parametric Kaplan-Meier estimate to first step is to calculate a survival object which contains the information on censoring. This object can then be used in fitting any of the available survival regression models. Since the default is available it is not necessary to specify `dist="weibull"`.

```
> library(survival)
> load("tumour.Rdata")
>
> tumour.sv <-Surv(time, 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
```

The survival object tumour.sv can be used for fitting another model, for example the exponential:

```
> tumourSurvExp<-survreg(tumour.sv, 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
```

Note that it is possible to fix (or constrain to a specific value) the scale parameter of the Weibull distribution. This parameter is often the most difficult to estimate, especially if it is actually close to 1. Fixing it to 1 reduces the Weibull to an exponential model:

```
> tumourSurvWeib1<-survreg(tumour.svweib,data=tumour, scale=1)
>
> summary(tumourSurvWeib1)

Call:
survreg(formula = tumour.svweib, data = tumour, scale = 1)
            Value Std. Error    z        p
(Intercept)  2.61      0.408 6.38 1.76e-10

Scale fixed at 1

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

which gives identical estimates etc to fitting an exponential model.

2.8 Summary

- Parametric Models
  - Estimate parameters by MLE
  - Uncensored observations contribute $f(t_i)$
  - & censored contribute $S(t_i)$ to likelihood
  - Use MLE theory for standard errors
  - Plug in MLEs for other functions of θ
  - Use formula for s.e.[g(θ)]

- Noting $f(t_i) = h(t_i)S(t_i)$ allows the likelihood to be written concisely with a censoring indicator $\delta_i = 1$ for uncensored, 0 for censored, as $L(\theta) = \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i)$

- For exponential data with mean lifetime λ^{-1}

$$\hat{\lambda} = \frac{\sum_1^n \delta_i}{\sum_1^n t_i} = \frac{\text{total number of deaths observed}}{\text{total time alive of all patients in the study}}$$

- Other models mentioned:
  - Weibull
  - log-Normal
  - gamma
  - Gumbel

- All generally require numerical estimation

- Easy to do in **R**

3 Two-Sample Comparisons

3.1 Introduction

A common problem is the comparison of two (or more) survival distributions, e.g. which treatment is better? Is the pattern of survivals/deaths for Males different from that for Females?

A simple comparison is to plot the Kaplan-Meier estimates ***for each group***. Is there any difference?

3.2 Logrank Test (non-parametric)

3.2.1 Example: brain tumour survival times

12 brain tumour patients randomized to radiation or radiation+chemotherapy. One year after the start of the study the survival times in weeks are:

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

(* denotes censored)

Kaplan Meier (check)

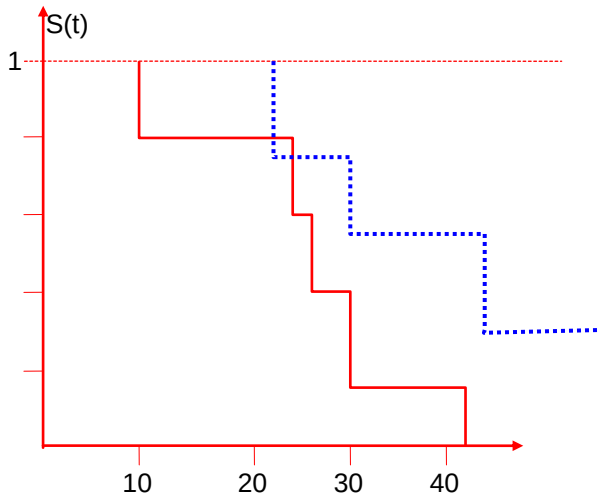
$\hat{S}_1(10)=0.833$ $\hat{S}_2(24)=0.800$

$\hat{S}_1(26)=0.625$ $\hat{S}_2(30)=0.600$

$\hat{S}_1(28)=0.417$ $\hat{S}_2(42)=0.300$

$\hat{S}_1(30)=0.208$

$\hat{S}_1(41)=0$



Test: H_0: $S_1(t)=S_2(t)$

H_1: $S_1(t) \neq S_2(t)$ (for some t)

Order times of death for two groups combined $t_{(1)}<t_{(2)}<$.....

Expected number of deaths: (i) Assume H_0 (no difference in groups) is true

(ii) first death at t=10; of 12 at risk, 1 died

If H_0 true we would expect $1\times{}^6/_{12}$ of these to be in group 1 and $1\times{}^6/_{12}$ of these to be in group 2.

Next death at t=24 , so $1\times{}^4/_9$ in group 1, $1\times{}^5/_9$ in group 2.

		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
					$O_1=5$	$O_2=3$		$E_1=2.87$	$E_2=5.13$

Log rank statistic: $(O_1–E_1)^2/E_1 + (O_2–E_2)^2/E_2 \sim \chi^2_1$ under H_0

Here $=2.46 < \chi^2_{1,0.95} = 3.84$

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

3.2.2 Notes

- Obvious generalization to 3 or more groups and then χ^2 has k–1 degrees of freedom
- Several other non-parametric tests are used — generalize the Wilcoxon-Mann-Whitney test ideas, e.g. Gehan, Cox-Mantel, Peto and Peto, Mantel-Haenszel etc. See references and computer packages.

3.2.3 Computer Implementation

3.2.3.1 R

In **R** the function for performing logrank tests is `survdiff()` . The procedure is similar to calculating a non-parametric (i.e. Kaplan-Meier) survival model. The first step, as always with censored survival times, is to create a survival object using `Surv()`. Next we need to indicate which group each observed or censored survival time comes from. This is done by regressing the `Surv` object on a factor which indicates the groups. The example below gives uses the brain tumour survival times used above.

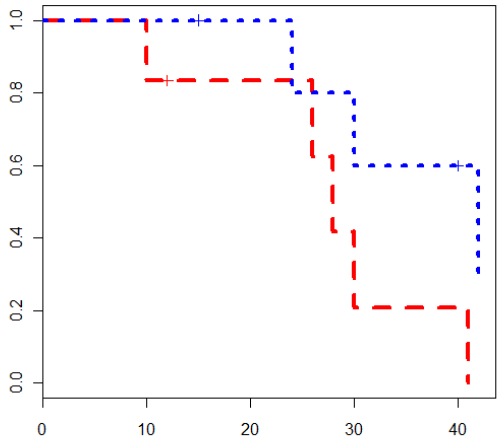
```
> library(survival)
Loading required package: splines
load("braintu.Rdata")

brain.sv<-Surv(time, censor, type = "right")
survdiff(brain.sv ~ group, data=braintu)
Call:
survdiff(formula = brain.sv, data = braintu)

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

 Chisq= 2.9  on 1 degrees of freedom, p= 0.0896
>
```

Note that the numerical value is slightly different from above since **R** handles ties in a more sophisticated way. Indeed other packages also differ slightly from each other since they may have different methods.

The survival object brain.sv can be used to fit non-parameteric (Kaplan-Meier) survival models to the data separately for each group and then produce Kaplan-Meier plots)

```
plot(survfit(brainsv,data=braintu),
lty=c(2,3),lwd=4,col=c("red","blue"))
```

Note that this uses line styles 2 and 3 (`lty` parameter) and colours red and blue for group 1 and group 2 (`col` parameter) and thickness 4 (`lwd` parameter). Type `help(par)` to find out more details.



3.2.3.2 S-PLUS

In S-PLUS there is ***no*** menu facility for log rank tests but they can be obtained from the command line in just the same way as in R:

```
survdiff(Surv(time,censor,type='right')~group, data=braintu)
```

(note the capitalization of `Surv`) which produces:

```
Call:
survdiff(formula = Surv(time, censor, type = "right")
~ group,
    data = braintu)

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

 Chisq= 2.9  on 1 degrees of freedom, p= 0.0896
```

It is possible to draw separate Kaplan-Meier plots from the menus by including the grouping variable as a main effect when creating the formula (see §2.4.4.2).

3.2.3.3 MINITAB

Log-rank tests are obtained as an option in the Kaplan-Meier plot above (§2.4.4.3): click **By variable** and alter default choice in **Result**s

3.2.3.4 SPSS

Log-rank tests are obtained as an option in the Kaplan-Meier plot above (§2.4.4.3): under Options click **Factor** and **Compare Factor**

Exercises 1

1) Returning to the Australian study on survival of spoons,
   i) Is there evidence that the disappearance of spoons is dependent upon either the category of tearoom or the value of the spoon?
   ii) What is the average rate of loss of teaspoons?
   iii) If the Institute where the study was conducted has 150 employees, how many teaspoons should be purchased annually to provide one spoon for every two people?

   (**N.B.** You should appreciate that the data given here are those observed at the Australian institution so you are advised to evaluate your answer to this question using common sense: the answer should be within the petty cash budget of the tea-room).

3.3 Parametric Tests

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. M.L.E. Test

Setup: n_1 observations on T_1~ Ex(λ_1), and n_2 on T_2~Ex(λ_2).

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n_j} \delta_{ji}}{\sum_{i=1}^{n_j} t_{ji}} = \frac{\Delta_j}{\Im_j} \text{ for j=1,2}$$

Δ_j=number of deaths in group j, $\Im_j$=total 'time' on test in group j

From 2.6.2.2 $\hat{\lambda}_j \approx N\left(\lambda_j, \lambda_j^2 / \Delta_j\right)$ for j=1,2

so $\hat{\lambda}_1 - \hat{\lambda}_2 \approx N\left(\lambda_1 - \lambda_2, \lambda_1^2/\Delta_1 + \lambda_2^2/\Delta_2\right)$ and so to test the hypothesis

$H_0 : S_1(t) = S_2(t)$ *vs.* $H_1 : S_1(t) \neq S_2(t)$ for some value of t

$\Leftrightarrow H_0 : \lambda_1 = \lambda_2$ *vs.* $H_1 : \lambda_1 \neq \lambda_2$

We have that under H_0 :

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

3.3.1.1 Example: brain tumour survival times

$n_1=6$ $\Delta_1=5$ $\Im_1=147$

$n_2=6$ $\Delta_2=3$ $\Im_2=193$

So $\hat{\lambda}_1 = 5/147=0.034$ and $\hat{\lambda}_2 = 3/193=0.0155$,

giving W=1.02 which is not significant at 5%,

so no evidence at the 5% level that the survivor functions differ.

3.3.2 Likelihood Ratio Test

Likelihood $L(\lambda_1,\lambda_2) = L(\lambda_1)L(\lambda_2)$,

maximizing $L(\lambda_1,\lambda_2)$ with respect to λ_1 and λ_2 gives

$$L_{max}(\lambda_1,\lambda_2) = L(\hat{\lambda}_1, \hat{\lambda}_2) = \hat{\lambda}_1^{\Delta_1} e^{-\hat{\lambda}_1\Im_1} \hat{\lambda}_2^{\Delta_2} e^{-\hat{\lambda}_2\Im_2}$$

If H_0 is true then the likelihood is $L(\lambda,\lambda)=\lambda^{\Delta_1+\Delta_2} e^{-\lambda(\Im_1+\Im_2)}$

so $\text{Log}_e\{L(\lambda,\lambda)\} = \ell(\lambda,\lambda) = (\Delta_1+\Delta_2)\log_e\lambda - \lambda((\Im_1 + \Im_2)$

$$\Rightarrow \hat{\hat{\lambda}} = \frac{\Delta_1 + \Delta_2}{\Im_1 + \Im_2}$$

$$\text{so } L(\hat{\hat{\lambda}}, \hat{\hat{\lambda}}) = \hat{\hat{\lambda}}^{\Delta_1+\Delta_2} e^{-\hat{\hat{\lambda}}(\Im_1+\Im_2)}$$

and, using the generalized likelihood ratio test we have, under H_0,

$$2\{\ell(\hat{\lambda}_1, \hat{\lambda}_2) - \ell(\hat{\hat{\lambda}}, \hat{\hat{\lambda}})\} \approx \chi_1^2$$

$$\text{i.e. } 2\{\Delta_1 \log_e\left(\frac{\Delta_1}{\Im_1}\right) + \Delta_2 \log_e\left(\frac{\Delta_2}{\Im_2}\right) - (\Delta_1 + \Delta_2)\log_e\left(\frac{\Delta_1 + \Delta_2}{\Im_1 + \Im_2}\right)\} \approx \chi_1^2$$

3.3.2.1 Example: brain tumour survival times

Test statistic = 1.20 < 3.84= $\chi^2_1(5\%)$, i.e. not significant at the 5% level, so again no evidence at the 5% level that the survivor functions differ.

3.4 Computer Implementation

Equivalents of the MLE test can be achieved by estimating parametric regression models with the group indicator as a dummy variable. This is available in **R** and S-PLUS and MINITAB and is described below in §4.1.

3.4.1 Illustration on brain Tumour times in R

Many of the calculation in §3.3.1.1 can be performed easily in R:

```
> library(survival)
> load("braintu.Rdata")
> attach(braintu)
> sum(time[group==1]); sum(time[group==2])
[1] 147
[1] 193
> sum(censor[group==1]); sum(censor[group==2])
[1] 5
[1] 3
```

From these the MLE test statistic can be calculated. A better way is to fit a model of the [censored] survival times on the group indicator using `survreg()` and an exponential distribution:

```
> brain.sv<-Surv(time,censor)
> brain.regexp<-survreg(brain.sv~group, dist="exponential")
> summary(brain.regexp)

Call:
survreg(formula = brain.sv ~ group, dist = "exponential")
            Value Std. Error    z       p
(Intercept) 2.598       1.06 2.44   0.0147
group       0.783       0.73 1.07   0.2836
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
```

Now note that 1/exp(3.381) = 0.034 (i.e. the ML estimate of λ_1) and 1/exp(3.381+0.783) = 0.0155 (i.e. the ML estimate of λ_2)

It can be checked that the standard error of the ML estimate of λ_1 is approximately 0.447/exp(3.381) (using the formula in §2.6.2.2) and that of the ML estimate of λ_2 is approximately $(0.447^2+0.730^2)^{½}$/exp(3.381+0.783) but better is to note that the p-value for testing whether the parameter indicating the group (i.e. whether the groups differ in their survival curves) is 0.284 which is close to the p-value of the MLE test statistic given in §3.3.1.1 of 1.02 which is 0.308 (given by `2*(1-pnorm(1.02)` in R). Further, the chi-squared value of 1.2 given above is precisely the value of the likelihood ratio test statistic.

3.5 Notes

(i) These two tests are asymptotically equivalent.

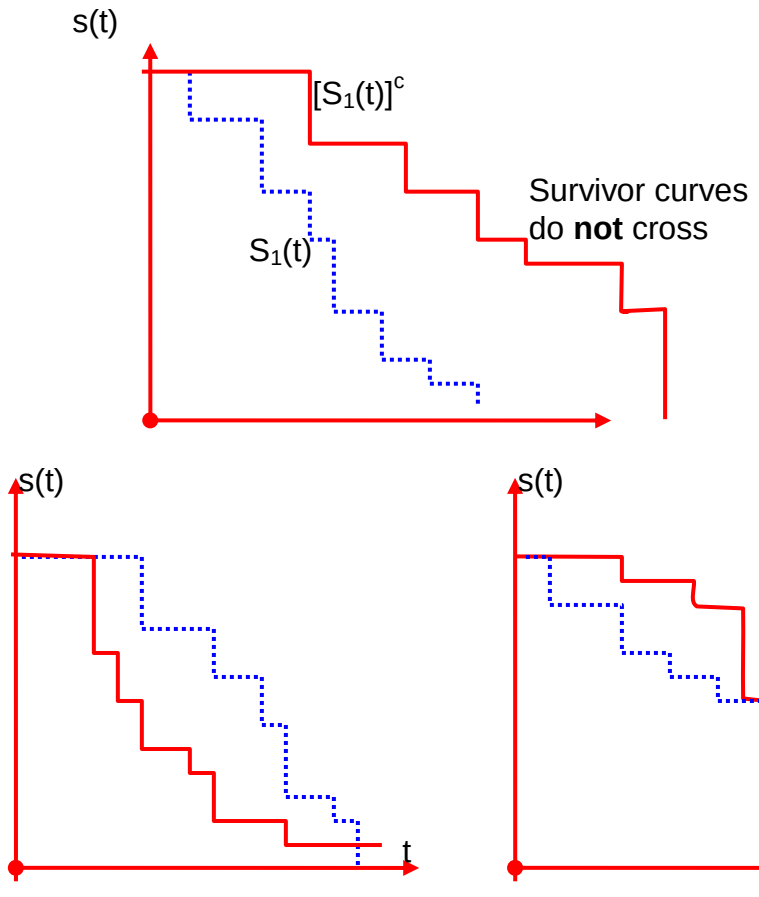
The Likelihood Ratio test is probably better for small samples.

(ii) In large samples little is gained usually in terms of power by using parametric methods instead of the log-rank test.

(iii) The log-rank test (devised by Peto & Peto, 1972, JRSS) assumes that $S_2(t)=[S_1(t)]^c$ or equivalently, $h_2(t)=ch_1(t)$

i.e. **proportional hazards** (see diagrams below)

If the two survivor functions differ but not in this way then the log-rank test could give misleading results.



low power

This is worse — the
log-rank statistic is almost zero

In the case of the third diagram it could be that a modification of the log-rank test is required and a *generalised log-rank test* would be appropriate. This is achieved in **R** by including an extra parameter `rho` (i.e. ρ) in the call to `survidiff()`. This gives weight $[S(t)]^{\rho}$ to each event (i.e. 'death') in the calculation of the log-rank statistic. Thus more weight can be given to differences in the survival curves in the short term by taking a value of $\rho > 0$ (and less weight if $\rho < 0$). This version of the test is a form of the generalized Gehan-Wilcoxon log-rank test. The simple choice of $\rho = 1$ is the Peto & Peto modification of it.

The test would be particularly appropriate when considering a class of model known as *accelerated failure models* and this topic will be returned to in §4.5.

3.6 Summary

- Log-rank test (2 or more groups)
  - Check assumptions with Kaplan-Meier plots first
  - Best when proportional hazards
  - Based on $\Sigma(O–E)^2/E \sim \chi^2_{k-1}$
  - Calculations similar to K-M (already done)
  - Packages may give slightly different answers because of using $\Sigma(O–E)^2/var(E)$
  - Implemented in **R** with function `survdiff()`
  - Generalised (Peto & Peto) log-rank test obtained with `survdiff(.,., rho=1)`
- **MLE Test**
  - 2 groups, single parameter
  - based on two-sample Normal test with s.e.s from general ML theory
- **Likelihood Ratio Test**
  - k groups, any # of parameters
  - difference in sum of maximized likelihoods and pooled likelihood
  - test statistic ~ χ^2_r where r = (k – 1) × # of parameters
- **All of these can be performed in R**

Tasks 3

For the data on the data *leukaemia remission times*
(given on Task Sheet 2)

1) Calculate the log rank statistic for testing for a difference in survival times between the two groups and assess its significance.
2) Assuming that survival times are exponentially distributed, Ex(λ_1) and Ex(λ_2) respectively, estimate λ_1 and λ_2.

3) Assuming that the survival times are exponentially distributed use the estimates from part (ii) to estimate the median survival times of the two groups, providing 95% confidence intervals for each group.

4) Calculate MLE and Likelihood Ratio Test statistics for testing for a difference in survival times between the two groups and assess their significance.

5) Plot the logs of the exponential survivor functions and the Kaplan-Meier survivor functions on the same graph. Comment on the fit of the exponential model to these data.

Comment on the effect of the drug.

4 Regression Models

4.1 Introduction

Each individual may be subject to their

own individual hazard rate $h_i(t)$

e.g. $T_i \sim Ex(\lambda_i)$ with $f_i(t) = \lambda_i e^{-\lambda_i t}$; i=1,2,...,n

Clearly, for the uncensored case $\hat{\lambda}_i = \frac{1}{t_i}$; i=1,2,...n

More useful models are obtained if we can inter-relate the λ_i's,

e.g. incorporate covariates in a regression model.

4.2 Parameteric Regression Models

4.2.1 Exponential Regression Model

Suppose $t_i \sim Ex(\lambda_i)$

We take $E[T_i] = \lambda_i^{-1} = \alpha + \beta x_i$ (i=1,2,...,n)

where the x_i are known values of a covariate.

Then $L(\alpha, \beta) = \prod_{i=1}^{n} \left\{ \frac{1}{\alpha + \beta x_i} e^{-\frac{t_i}{\alpha+\beta x_i}} \right\}$

so $\log_e L = \ell(\alpha, \beta) = -\sum_{i=1}^{n} \log_e(\alpha + \beta x_i) - \sum_{i=1}^{n} \frac{t_i}{\alpha + \beta x_i}$

$$\text{so } \frac{\partial \ell}{\partial \alpha} = -\sum_{i=1}^{n} \frac{1}{\alpha + \beta x_i} + \sum_{i=1}^{n} \frac{t_i}{(\alpha + \beta x_i)^2} = 0$$

$$\text{and } \frac{\partial \ell}{\partial \beta} = -\sum_{i=1}^{n} \frac{x_i}{\alpha + \beta x_i} + \sum_{i=1}^{n} \frac{t_i x_i}{(\alpha + \beta x_i)^2} = 0$$

and we solve these by iterative maximum likelihood. We would have to obtain initial values for α and β and we might do this by plotting the observed survival times t_i against x_i and fitting a line by eye, measuring slope and intercept or obtaining the least squares estimate. These would be very rough estimates but would serve as a starting point for the iterations. This is the method used in the numerical example below and is perhaps the easiest if one is actually doing this 'by hand' as in the example but it is different from the method used in **R** and S-PLUS and other packages which typically model the logarithm of survival time. An illustration of the same example analysed with **R** and S-PLUS is given below.

The general theory of maximum likelihood estimation gives that

asymptotically $$\begin{pmatrix}\hat\alpha\\ \hat\beta\end{pmatrix} \approx N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{bmatrix} -E\left[\frac{\partial^2 \ell}{\partial \alpha^2}\right] & -E\left[\frac{\partial^2 \ell}{\partial \alpha \partial \beta}\right] \\ -E\left[\frac{\partial^2 \ell}{\partial \alpha \partial \beta}\right] & -E\left[\frac{\partial^2 \beta}{\partial \beta^2}\right] \end{bmatrix}^{-1}\right)$$

$$\approx N\left(\begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \begin{bmatrix} \sum \frac{1}{(\alpha+\beta x_i)^2} & \sum \frac{x_i}{(\alpha+\beta x_i)^2} \\ \sum \frac{x_i}{(\alpha+\beta x_i)^2} & \sum \frac{x_i^2}{(\alpha+\beta x_i)^2} \end{bmatrix}^{-1}\right)$$

4.2.1.1 Notes

(i) Estimate λ^{-1} by $\hat\alpha + \hat\beta x$ for given x

(ii) $\text{var}(\hat\alpha + \hat\beta x) = \text{var}(\hat\alpha) + x^2 \text{var}(\hat\beta) + 2x \text{ cov}(\hat\alpha, \hat\beta)$

$\Rightarrow$ e.g. 95% Confidence Interval for λ^{-1}

(using a Normal approximation)

(iii) Strictly we need to impose the condition that $\alpha+\beta x>0$ but this is not a problem in practice if the model is reasonably appropriate.

(iv) Could extend to censored data, bringing in the censored observations to the likelihood in a similar way to that in the simple exponential model (i.e. with the survivor function) but note that this would alter the closed form expressions given for the variances of the estimates given above.

4.2.1.2 Example: myelogenous leukemia

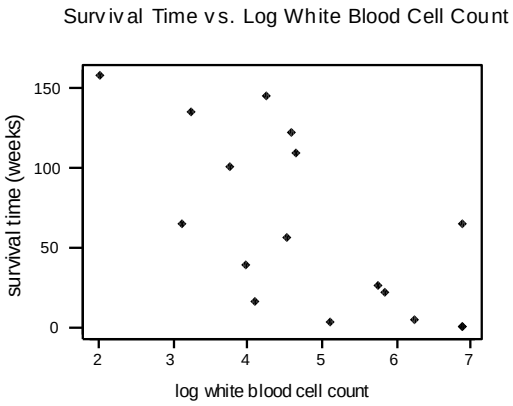
Survival times (from date of diagnosis) of patients with acute myelogenous leukemia.

Covariate: white blood cell count.

because of variability take $\lambda_i^{-1}=\alpha+\beta\log(\text{WBC})$

(AG positive) n=17

Patient	WBC$\times 10^2$	Survival Time (weeks)
1	23	65
2	7.5	156
3	43	100
4	26	134
5	60	16
6	105	108
7	100	121
8	170	4
9	54	39
10	70	143
11	94	56
12	320	26
13	350	22
14	1000	1
15	1000	1
16	520	1
17	1000	5
Median values	100	56



From the graph we can obtain initial estimates of α and β and then iterate.

This gives

$$\hat{\alpha}=240,\hat{\beta}=-44;\hat{\text{var}}(\hat{\alpha})=95.5,\hat{\text{var}}(\hat{\beta})=20.1,\hat{\text{cov}}(\hat{\alpha},\hat{\beta})=-1914$$

and so a 95%CI for $\alpha+\beta x$ ($=\lambda_x^{-1}$) from

$$240-44x-1.96[95.5-3828x+20.1x^2]^{0.5}<=\lambda_x^{-1}<240-44x+1.96[95.5-3828x+20.1x^2]^{0.5}$$

In this example it was fortunate that $\hat{\alpha} + \hat{\beta}x > 0$ for the range of values in the data set. To avoid this problem we could instead model the log of the survival term as a linear function of the covariate and this is the way that such parametric regression models are implemented in computer packages. [NB: here the estimates are MLE estimates obtained by substituting for λ_i in terms of α and β and then, from general asymptotic MLE theory the variances and covariance are obtained from the inverse of the matrix of second derivatives. This is beyond the scope of this course]

4.2.2 Computer Implementation

4.2.2.1 R

Parametric models can be fitted using `survreg()` as described in §2.6.6. As always the first step is to create a survival object with `Surv()` to combine the censoring information with the survival times. This 'object' is then regressed on any appropriate covariates. Estimates of the coefficients of the covariates in the regression, together with their standard errors are provided in the results. 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. Note also that the analysis models the log of the survival times and this needs to be remembered when estimating any quantity from the model such as median survival time for a particular set of covariates.

The procedure is illustrated on the survival times of patients with myelogenous leukemia. In this particular case none of the observations

is censored so in the initial `Surv()` step the censoring variable can be omitted.

```
> library(survival)
Loading required package: splines
> load("wbcleuk.Rdata")
> 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
7        7  100.0      121 4.605170
8        8  170.0        4 5.135798
9        9   54.0       39 3.988984
10      10   70.0      143 4.248495
11      11   94.0       56 4.543295
12      12  320.0       26 5.768321
13      13  350.0       22 5.857933
14      14 1000.0        1 6.907755
15      15 1000.0        1 6.907755
16      16  520.0        1 6.253829
17      17 1000.0        5 6.907755
18       2    7.5      156 2.014903
> attach(wbcleuk)
> survreg(survival~log.wbc.)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regexp<-survreg(wbcleuk.sv~log.wbc.,dist="exponential")
> summary(wbcleuk.regexp)

Call:
survreg(formula = wbcleuk.sv ~ log.wbc., dist = "exponential")
            Value Std. Error     z        p
(Intercept)  7.84      1.052  7.45 9.05e-14
log.wbc.    -0.89      0.220 -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
```

Because the logarithms of survival times are used this models the logarithm of the mean survival time for a given value of log(wbc). Thus the estimated mean survival time for a patient with a wbc of 54 is exp(7.84-0.89log(54)) = 72.95 days. An approximate standard error for this estimate can be calculated from the standard errors for the intercept and coefficient and using the formula for standard errors of a function of an estimate given in §2.6.2.2.

Since this model is an exponential model, the estimated median survival time for a patient with a wbc of 54 is –72.95log(0.5) = 50.57 days. Note that the dataset contains one observed survival time of a patient (number 9) with a wbc of 54 who survived 39 days.

For further illustration given below is the analysis in fitting a Weibull model to the same data. Since the Weibull is the default distribution for `survreg()` it is not necessary to specify that the distribution is Weibull with `dist="weibull"` and it is omitted.

```
> 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= 0.9

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
```

It might be noted that the estimated scale parameter is 0.9, very close to 1.0 which is 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.

The estimated mean survival time for a patient with a wbc of 54 is exp(7.849–0.882log(54) = 76 days, little different from the estimate based on the exponential model. Further investigation (not given here) indicates that the standard error of the Weibull estimate is appreciably larger than the exponential estimate, again illustrating the superiority of the exponential fit.

4.2.2.2 S-PLUS

Parametric models can be fitted from the menus under `Statistics>Survival>Parametric Survival …` and the operation of this follows the familiar pattern of first needing to create a formula to declare which is the survival time, what is the censoring variable and what are the explanatory variables. Note that censored values are handled easily.

Note also that the analysis models the log of the survival time so the estimates from S-PLUS given below differ substantially from those obtained by iteration 'by hand' in fitting the exponential model to the actual survival times. However, it is not difficult to see that the two models are essentially equivalent in that estimates of the actual survival times for a given value of the white blood cell count are very close. The 'long results' output is:–

```
Call:
survReg(formula = Surv(survival) ~ log.wbc., data = wbcleuk,
na.action = na.exclude, dist = "weibull", scale = 0,
     control = list(maxiter = 30, rel.tolerance = 1e-005,
failure = 1))
             Value Std. Error      z         p
(Intercept)  7.849      0.986  7.957 1.76e-015
   log.wbc. -0.882      0.207 -4.265 2.00e-005
 Log(scale) -0.105      0.187 -0.562 5.74e-001

Scale= 0.901

Weibull distribution
Loglik(model)= -84.3   Loglik(intercept only)= -92
     Chisq= 15.4 on 1 degrees of freedom, p= 0.000087
Number of Newton-Raphson Iterations: 4
n= 18
```

Other distributions available are Weibull, extreme, Gaussian (i.e. normal), loggausian, logistic and loglogistic. In practice, the exponential model is rarely used and the Weibull the most common model to try first.

4.2.2.3 MINITAB

Parametric models can be fitted from the menus under `Stat>Reliability/Survival>Regression with Life Data …` with the same set of distributions as in S-PLUS.

4.2.2.4 SPSS

There are currently no facilities for fitting parametric models in SPSS (i.e. up to version 15).

4.2.3 Two-Sample Example

Consider the two group exponential model of §3.3

$$T_1 : f_1(t) = \lambda_1 e^{-\lambda_1 t} \qquad h_1(t)=\lambda_1$$

$$T_2 : f_2(t) = \lambda_2 e^{-\lambda_2 t} \qquad h_2(t)=\lambda_2$$

Let x=0 for group 1

=1 for group 2 — a binary indicator variable

Model: $h(t;x) = \lambda e^{\beta x}$ $= \lambda$ for x=0 (group 1)

or λe^{β} for x=1 (group 2)

i.e. a reparameterization: $\lambda=\lambda_1$; $\beta=\log_e(\lambda_2/\lambda_1)$

The sign of β determines whether $\lambda_2>\lambda_1$ or $\lambda_1>\lambda_2$

($\beta>0$) ($\beta<0$)

4.2.4 Notes

(i) $\log_e h(t;x)=\log_e(\lambda)+\beta x$

So the model is sometimes called

the *log-linear model* for the hazard function

(ii) If we parameterize in this way it ensures that $h(t;x)>0$

(iii) Could extend this to several groups with $h(t;\underline{x})=\lambda e^{\underline{\beta}'\underline{x}}$

(with the use of dummy variables)

$x_1=1$ if group A, $x_1=0$ otherwise; $x_2=1$ if group B, $=0$ otherwise

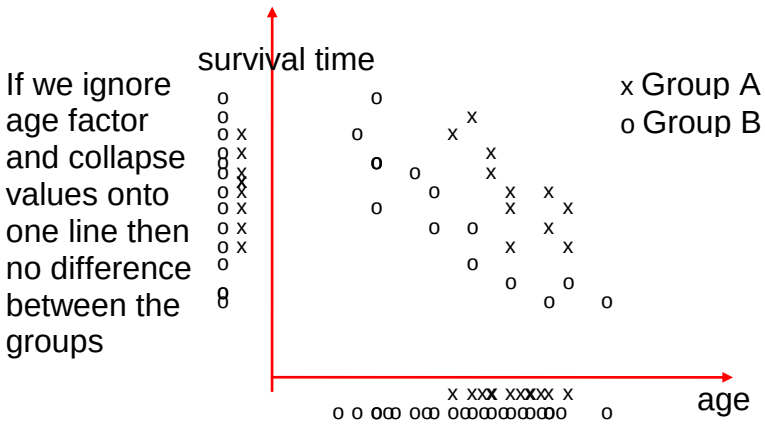
the $\underline{x}=(x_1,x_2)$ and $(1,0)\rightarrow$A ; $(0,1)\rightarrow$B ; $(0,0)\rightarrow$C

4.3 Covariates and Prognostic Factors

In comparisons of treatments randomization in large samples ensures that factors affecting survival time, e.g. stage of disease, age, sex etc., are balanced between the treatment groups. Regression type models can allow for these factors as well within a randomization trial.

4.3.1 Notes

(i) Resulting procedures can be more sensitive to treatment differences.



(ii) Factors themselves may be of interest → prognostic factors

i.e. for predicting survival times

(iii) Randomization may occasionally 'go wrong'

— regression methods help to correct for this.

(iv) Interactions: treatment effect may be different for 'different' patients (according to the prognostic factors)

e.g. x_1=0 for control group, x_1=1 for treatment group

x_2=0 for stage I, x_2=1 for stage II and x_2=2 for stage III.

Then define $x_3=x_1x_2$ for the interaction

between the factors age and stage of disease

4.3.2 Modelling

(i) look at each factor separately

(ii) try $E[T] = \beta_1 x_1+\beta_2 x_2+...+\beta_p x_p=\underline{\beta}'\underline{x}$

$\rightarrow$ problems since we need to ensure that $\underline{\beta}'\underline{x}>0$ all $\underline{\beta}$ and $\underline{x}$

(iii) try $E[T]=\alpha \exp\{\underline{\beta}'\underline{x}\}$ — log-linear type model

(iv) Try to model through the hazard function

$h(t) \rightarrow h(t;\underline{x}) = h_0(t) \exp\{\underline{\beta}'\underline{x}\}$

where $h_0(t)$ is the 'underlying' hazard rate

(v) β_j reflects the effect of x_j on survival

if $\beta_j>0$: increasing $x_j \Rightarrow$ hazard ↗ $\Rightarrow$ poorer survival prospect

if $\beta_j<0$: increasing $x_j \Rightarrow$ hazard ↘ $\Rightarrow$ better survival prospect

if $\beta_j=0$: increasing $x_j \Rightarrow$ no effect on survival.

4.3.3 Exponential Model

$$h(t;\underline{x}) = \lambda \exp\{\underline{\beta}'\underline{x}\}$$

$$f(t;\underline{x}) = \lambda \exp\{\underline{\beta}'\underline{x}\} \exp\{-\lambda t \exp\{\underline{\beta}'\underline{x}\}\}$$

$$S(t;\underline{x}) = \exp\{-\lambda t \exp\{\underline{\beta}'\underline{x}\}\}$$

Data: $(t_1,\delta_1,\underline{x}_1), (t_2,\delta_2,\underline{x}_2),\ldots, (t_n,\delta_n,\underline{x}_n)$ where $\underline{x}_i=(x_{i1},x_{i2},\ldots,x_{ip})$

Estimate the p+1 parameters $\lambda, \beta_1,\beta_2,\ldots,\beta_p$ by maximum likelihood

$$L(\lambda,\underline{\beta})=\prod_{i=1}^{n}[\lambda \exp\{\underline{\beta}'\underline{x}_i\}]^{\delta_i} \exp\{-\lambda t_i \exp\{\underline{\beta}'\underline{x}_i\}\}$$

$$=\prod_{i=1}^{n}[\lambda\exp\{\underline{\beta}'\underline{x}_i\}\exp\{-\lambda t_i \exp\{\underline{\beta}'\underline{x}_i\}\}]^{\delta_i} [\exp\{-\lambda t_i \exp\{\underline{\beta}'\underline{x}_i\}\}]^{1-\delta_i}$$

$$\ell(\lambda,\underline{\beta}) = \Sigma\delta_i \log_e\lambda + \Sigma\delta_i\, \underline{\beta}'\underline{x}_i - \Sigma\lambda\, t_i \exp\{\underline{\beta}'\underline{x}_i\}$$

$$\frac{\partial\ell}{\partial\lambda} = \frac{\Delta}{\lambda} - \Sigma t_i \exp\{\underline{\beta}'\underline{x}_i\}$$

where $\Delta=\Sigma\delta_i$ the total number of deaths

$$\frac{\partial\ell}{\partial\beta_j} = \Sigma\delta_i x_{ij} - \lambda\Sigma x_{ij}\, t_i \exp\{\underline{\beta}'\underline{x}_i\} \quad (j=1,2,\ldots,p)$$

Setting these two derivatives = 0 and solving iteratively

gives the maximum likelihood estimates of λ and β.

For estimates of variance and standard errors we need

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{\Delta}{\lambda^2}$$

$$\frac{\partial^2 \ell}{\partial \lambda \partial \beta_j} = -\Sigma x_{ij}\, t_i \exp\{\underline{\beta}' \underline{x}_i\} \qquad (j=1,2,...,p)$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\lambda \Sigma x_{ij} x_{ik}\, t_i \exp\{\underline{\beta}' \underline{x}_i\} \quad (j=1,2,...,p;\ k=1,2,...,p)$$

4.3.4 Other Models

Covariates can be incorporated into any model used for survival data by including a multiplicative term $\exp\{\underline{\beta}'\underline{x}\}$ in the expression for the hazard function, e.g. for the Weibull use $h(t;\underline{x})=\lambda\gamma t^{\gamma-1} \exp\{\underline{\beta}'\underline{x}\}$.

Many packages offer a wide choice of distributional models for regression analysis of survival data

4.4 Proportional Hazards Model

This is a semi-parametric model proposed by Cox (1972).

It is convenient and general model for comparing 2 groups of survival times — very widely applied.

$$h(t,\underline{x}) = h_0(t)\exp\{\underline{\beta}'\underline{x}\}$$

4.4.1 Notes

(i) $h_0(t)$ — baseline hazard function, i.e. corresponds to hazard of a patient with $\underline{x}=\underline{0}$

(ii) Dependence of failure time on explanatory variables is precisely modelled; but actual distribution of failure is not parametrically specified, i.e. $h_0(t)$ is not specified.

(iii) Useful in medical situations

— important to know which prognostic variables have an effect and to what extent

— knowing the actual distribution of survival time is not as important.

(iv) **special cases:**

Exponential: $h(t;\underline{x})=\lambda\exp\{\underline{\beta}'\underline{x}\}$ $h_0(t)=\lambda$

Weibull: $h(t;\underline{x}) =\lambda\gamma t^{\gamma-1}\exp\{\underline{\beta}'\underline{x}\}$ $h_0(t)=\lambda\gamma t^{\gamma-1}$

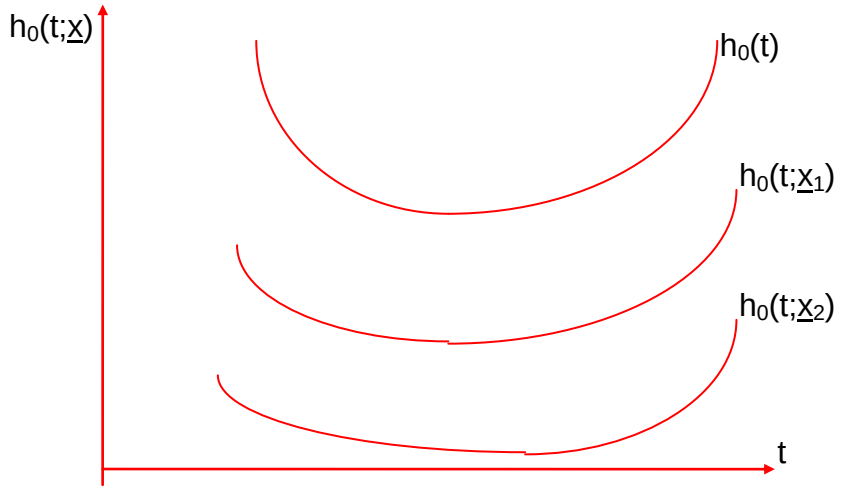
(v) For patients with covariates $\underline{x}_1$ and $\underline{x}_2$ we have

$h(t;\underline{x}_1)/\ h(t;\underline{x}_2) \quad = h_0(t)\ \exp\{\underline{\beta}'\underline{x}_1\}/\ h_0(t)\ \exp\{\underline{\beta}'\underline{x}_2\}$

$= \exp\{\underline{\beta}'(\underline{x}_1 - \underline{x}_2)\}$ ***INDEPENDENT OF t***

i.e. hazard functions for any 2 patients are proportional over time,

i.e. the linear component of the model does not vary with time.



The model assumes patients have the same 'shape' of hazard function — but shifted multiplicatively according to $\underline{x}$.

Note that under this model ***they can never cross*.**

4.4.2 Parameter Estimation

Observations

Survival times	$t_1, t_2, \ldots, t_n$
Censorings	$\delta_1, \delta_2, \ldots \delta_n$
Covariates	$\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n$

$$h(t;\underline{x}_i) = h_0(t) \exp\{\underline{\beta}'\underline{x}_i\}$$

$$S(t;\underline{x}_i) = \exp\{-\int_0^t h_o(u) \exp\{\underline{\beta}'\underline{x}_i\} du\}$$

$$= \exp\{-\exp\{\underline{\beta}'\underline{x}_i\} \int_0^t h_o(u)\, du\}$$

$$= [S_0(t;\underline{x}_i)]^{\exp\{\underline{\beta}'\underline{x}_i\}}$$ — the baseline survivor function

$$f(t;\underline{x}_i) = h_0(t) \exp\{\underline{\beta}'\underline{x}_i\} S(t;\underline{x}_i)$$

$$\text{likelihood} = \prod_{i=1}^{n} [h(t_i;\underline{x}_i)]^{\delta_i} S(t_i;\underline{x}_i)$$

This involves $h_0(t)$, so to proceed further we need to specify a parametric form for $h_0(t)$, e.g. $h_0(t)=\lambda$ say.

Alternatively we can use the ***partial likelihood approach***

4.4.3 Partial Likelihood Approach

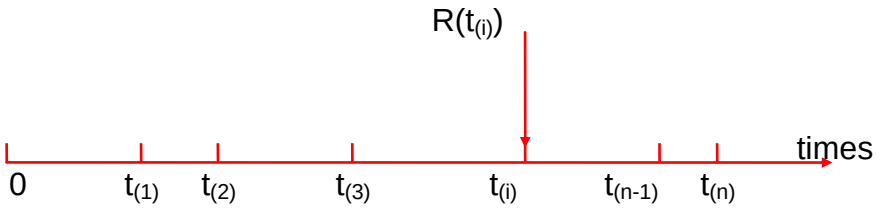
(Called *partial* since it does not make direct use of the actual censored and uncensored survival times)

Suppose that there are *no ties* — only one individual dies at each death time:

Ordered (by time) observations:

Survival times	$t_{(1)},t_{(2)},\ldots,t_{(n)}$
Censorings	$\delta_{(1)},\delta_{(2)},\ldots\delta_{(n)}$
Covariates	$\underline{x}_{(1)},\underline{x}_{(2)},\ldots\underline{x}_{(n)}$

Risk set R(t) at t: set of individuals alive and in the trial just before time t. $R(t_{(i)})$ is the set of individuals who are alive and uncensored at time just before $t_{(i)}$.



Consider time points at which deaths occur:

P[individual (i) dies at $t_{(i)}$ | exactly one patient in the risk set $R(t_{(i)})$ dies at $t_{(i)}$]

$$= \lim_{\delta t \to 0} \left\{ \frac{P[\text{death of (i) in } (t_{(i)}, t_{(i)} + \delta t) \mid R(t_{(i)})}{P[\text{one death in } (t_{(i)}, t_{(i)} + \delta t) \mid R(t_{(i)})} \right\}$$

$$\triangleq \frac{h(t_{(i)}; x_{(i)})\delta t}{\sum_{j \in R(t_{(i)})} h(t_{(i)}; x_j)\delta t}$$

{because .....

P[one death in $[t_{(i)}, t_{(i)}+\delta t]$ | $R(t_{(i)})$]

$$= \sum_{j \in R(t_{(i)})} P[j \text{ dies}]P[\text{others don't die}]$$

$$= \sum_{j \in R(t_{(i)})} h(t_{(i)}; x_j)\delta t \prod_{\substack{k \in R(t_{(i)}) \\ k \neq j}} [1 - h(t_{(i)}; x_k)\delta t$$

$$\triangleq \sum_{j \in R(t_{(i)})} h(t_{(i)}; x_j)\delta t \quad \text{ignoring terms in } (\delta t)^2$$

..... end of explanation}

$$= \frac{h_0(t_{(i)})e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} h_0(t_{(i)})e^{\underline{\beta}'\underline{x}_{(j)}}}$$

$$= \frac{e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} e^{\underline{\beta}'\underline{x}_{(j)}}}$$

NOTE proportional hazards assumption necessary here

Individuals for whom the survival times are 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.

Form the 'likelihood' by taking products over the observed failure times $t_{(i)}$.

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\underline{\beta}'\underline{x}_{(i)}}}{\sum_{j\in R(t_{(i)})} e^{\underline{\beta}'\underline{x}_{(j)}}} \right\}^{\delta_{(i)}}$$

4.4.3.1 Notes

(i) If no censored observations this is a ***conditional*** likelihood, conditional on the observed $t_{(1)},t_{(2)},\ldots,t_{(n)}$.

(ii) With censored observations this is known as a ***partial*** likelihood.

(iii) Use of partial likelihood was justified by Cox (1975) in Biometrika. He showed that the usual likelihood methods apply in this case. So we maximize the (partial) likelihood to estimate β by $\hat{\underline{\beta}}$ and asymptotically

$\hat{\underline{\beta}} \approx N(\underline{\beta}, \left[-\frac{\partial^2 \ell}{\partial\beta_i \partial\beta_j} \right]^{-1}_{\hat{\underline{\beta}}})$ where ℓ is the log likelihood.

(iv)

Consider likelihood contributions at observed failure times. No extra information on $\underline{\beta}$ is obtained from the fact that there are no failures between two specific observed lifetimes.

If we had a parametric form for $h_0(t)$ in our model, there would then be contributions to the inferences about $\underline{\beta}$ from the intervals with no failures. Intervals between successive death times convey no information about the effect of explanatory variables on the hazard of death. This is because the baseline hazard has an arbitrary form and so it is conceivable that $h_0(t)$ and hence h(t) is zero in those time intervals in which there are no deaths. This in turn means that those intervals give no information on the $\underline{\beta}$ parameters.

(v) Ties (Peto's adjustment)

$t_{(1)}<t_{(2)}<...<t_{(k)}$ the k distinct survival times

$d_{(1)}, d_{(2)},.....,d_{(k)}$ numbers of deaths at these times

$D(t_{(1)}), D(t_{(2)}),...D(t_{(k)})$ death set at $t_{(i)}$

Allow each of $d_{(i)}$ deaths at $t_{(i)}$ to contribute a factor (as before) to partial likelihood, each with the same risk set $R(t_{(i)}$

$$\text{'Likelihood'} = \prod_{i=1}^{k} \left\{ \frac{\exp\{ \sum_{j \in D(t_{(i)})} \underline{\beta}' \underline{x}_{(j)}}{\left[\sum_{j \in R(t_{(i)})} e^{\underline{\beta}' \underline{x}_{(j)}} \right]^{d_{(i)}}} \right\}$$

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

Other methods for adjustment for ties exist (e.g. by Cox, Breslow, Efron,.....).

4.4.4 Example: atrial fibrillation

The table below gives details of a proportional hazards model fitted to some data obtained from patients being treated for atrial fibrillation. The purpose of the treatment is to maintain normal heart rhythm and 'survival time' is in terms of time to relapse.

Variable	Coefficient	Standard Error	χ^2 statistic (using lrt)	coeff/s.e.
treatment (0=A, 1=B)	–1.42	0.64	4.89	**–2.22**
age (years)	–0.004	0.034	0.01	–0.12
sex (1=M,0=F)	0.31	0.72	0.18	0.43
volume of heart (mml)	0.0076	0.0036	4.44	**2.11**
Duration of symptoms (months)	–0.004	0.063	0.00	–0.06
digitalisation	–0.59	0.73	0.66	–0.81

Questions: (a) Describe the effects of treatments and additional covariates on time to relapse.

(b) The above analysis did not consider treatment×covariate interactions. How would this be done?

Answers:

(a) Model is

$h(t;\underline{x}) = h_0(t)\exp\{\beta_1 x_1+\beta_2 x_2+...+\beta_6 x_6\}$

where $x_1=0$ for treatment A, and $x_1=1$ for treatment B, i.e.

$h(t;\underline{x}) = h_0(t)\exp\{\beta_2 x_2+...+\beta_6 x_6\}$ for treatment A, and

$h(t;\underline{x}) = h_0(t)\exp\{\beta_1+\beta_2 x_2+...+\beta_6 x_6\}$ for treatment B

and use approximation $\hat{\underline{\beta}} \approx N(\underline{\beta}, \left[-\frac{\partial^2 \ell}{\partial\beta_i \partial\beta_j}\right]^{-1}_{\hat{\underline{\beta}}})$ for calculating standard errors.

Note that $\exp\{\beta_1\}$ has an interpretation of the *hazard ratio* of treatment B to treatment A seen by dividing the last two expressions above.

To see which factors are shewn to affect the hazard we examine the estimated coefficients and their standard errors. We can use the χ^2 statistic or coeff/s.e. for a factor with only 2 levels or a continuous covariate. If there is a factor with k-levels coded as k–1 dummy variables then the χ^2 statistics for each of the dummy variables should be added together and compared with χ^2 on k–1 d.f.

Note that $\chi^2 \approx$ (coeff/s.e.)^2 so if the χ^2 is not given and only the coefficients and their standard errors then it is possible to deduce the χ^2 values and so find the overall χ^2 value for a k-level factor coded as k–1 dummy variables. Note that it is not sensible (i.e. *wrong*) to consider the parameters for each of the levels of a factor in isolation. The k–1 parts of the chi-squared statistic must be combined and an overall assessment made of the factor made.

Treatment: $\hat{\beta}_1/\text{s.e.}(\hat{\beta}_1) = |-2.22| > 1.96$ so we have good evidence of effect of treatment. $\hat{\beta}_1 < 0$ so treatment = 1 decreases hazard, i.e. treatment B is 'better'

Heart volume:

$\hat{\beta}_4/\text{s.e.}(\hat{\beta}_4) = +2.11 > 1.96$ so increased heart volume decreases relapse time.

No evidence that other factors affect relapse time

NB Not shewn that other factors have no effect

It is also useful to calculate confidence intervals for the parameters, not just those where there is evidence that the factor is affecting the survival but also for those where the evidence is not there. This allows assessment of how big the effect could be. For example, the 95% CI for β_3 (M/F) is $0.31 \pm 2 \times 0.72 = (-1.13, 1.75)$, i.e. there could be a large difference between M & F and the data do not exclude this possibility. It may also be useful to calculate a confidence interval for $\exp\{\beta_3\}$ which has the interpretation of the *hazard ratio* of males to females. This would give (0.323, 5.75) so that the hazard for men could be just less than a third that for women or nearly six times as much; the data exclude neither possibility.

(b) Interaction terms would be handled by creating a new variable as the product of the treatment and the covariate values. In this case the treatment is coded as 0 for treatment A and 1 for B, so the value of this interaction term would be 0 for all subjects receiving A and the same as the covariate for those on B. In the example above Treatment is variable x_1 and age is variable x_3 and there are six variables in all. We create a new variable $x_7 = x_1 \times x_3$ and then our model is

$$h(t;\underline{x}) = h_0(t)\exp\{\beta_2 x_2+\beta_3 x_3+...+\beta_6 x_6\} \text{ for treatment A, and}$$

$$h(t;\underline{x}) = h_0(t)\exp\{\beta_1+\beta_2 x_2+(\beta_3+\beta_7)x_3+...+\beta_6 x_6\} \text{ for treatment B}$$

and β_7 reflects the interaction effect, (note that x_7 is identical to x_3 for those on treatment B but 0 for those on A).

Exactly the same method is appropriate for handling interactions between two continuous covariates and between two 2-level factors. Interactions involving a k-level factor can only be handled by converting the factor into k–1 dummy binary variables. In this case the interaction term has k–1 degrees of freedom if it is a k-level factor×covariate interaction or (k–1)(j–1) degrees of freedom for an interaction between a k-level and a j-level factor. This also means that the separate parts of the chi-squared statistic must be combined before assessing significance.

4.4.5 Computer Implementation

4.4.5.1 R

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)
Loading required package: splines
> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regcox<-coxph(wbcleuk.sv~log.wbc.)
> summary(wbcleuk.regcox)
Call:
coxph(formula = wbcleuk.sv ~ log.wbc.)

  n= 18

           coef exp(coef) se(coef)     z Pr(>|z|)
log.wbc. 1.1753    3.2392   0.3244 3.623 0.000292 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         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=8.192e-06
Wald test            = 13.12  on 1 df,   p=0.0002916
Score (logrank) test = 17.39  on 1 df,   p=3.043e-05
```

Primary interest will be in the estimated coefficients and their standard errors to perform partial z-tests and estimate hazard ratios.

4.4.5.2 S-PLUS

Cox proportional hazards models can be fitted from the menus under `Statistics>Survival>Cox Proportional Hazards …` and the operation of this follows the familiar pattern of first needing to create a formula to declare which is the survival time, what is the censoring variable and what are the explanatory variables. Note that censored values are handled easily. The 'long results' output for the very simple example of survival times with acute myelogenous leukaemia which has no censoring and only one explanatory variable (log white blood cell count):–

```
     *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(survival) ~ log.wbc., data = wbcleuk,
na.action = na.exclude, method = "efron", robust = F)

  n= 18

         coef exp(coef) se(coef)    z       p
log.wbc. 1.18      3.24    0.324 3.62 0.00029

         exp(coef) exp(-coef) lower .95 upper .95
log.wbc.      3.24      0.309      1.72      6.12

Rsquare= 0.669   (max possible= 0.982 )
Likelihood ratio test= 19.9  on 1 df,   p=8.19e-006
Wald test            = 13.1  on 1 df,   p=0.000292
Score (logrank) test = 17.4  on 1 df,   p=0.0000304
```

It is possible to produce a Kaplan-Meier plot of the survival time calculated for a subject with mean values of all the covariates.

4.4.5.3 MINITAB

There are currently no facilities in MINITAB for Cox regression (i.e. up to version 15)

4.4.5.4 SPSS

Cox proportional hazards regression is available through the menus:–

`Analyze>Survival>Cox Regression ...`

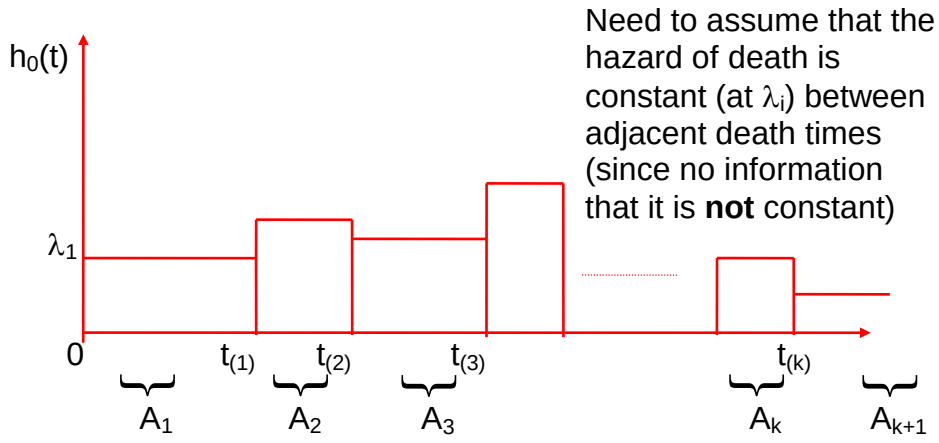
Need to specify value indicating uncensored values and the graphical output indicates the censored values as with S-PLUS.

4.4.6 Estimation of h(t)

We can obtain $\hat{\underline{\beta}}$, can we also estimate $h_0(t)$ to get $\hat{h}_0(t)$ and then $\hat{S}_0(t)$ and thus $\hat{S}(t,\underline{x}) = \hat{S}_0(t)\exp(\hat{\underline{\beta}}'\underline{x})$

There several methods available for the estimation of the baseline hazard function, e.g. Breslow (1974, Biometrics).

Let $t_{(1)}<t_{(2)}<...<t_{(k)}$ be the k distinct death times.



The full likelihood is

$$L = \prod_{i=1}^{n} \{h_0(t_i) e^{\underline{\beta}' \underline{x}_i}\}^{\delta_i} \exp\{-\int_0^{t_i} h_0(u) e^{\underline{\beta}' \underline{x}_i} du\}$$

and since $h_0(t)=\lambda_j$ if $t \in A_j=(t_{(j-1)}, t_{jj})]$ we have

$$\int_0^{t_i} h_0(u) = \int_0^{t_{(1)}} \lambda_1 du + \int_{t_{(1)}}^{t_{(2)}} \lambda_2 du + ... + \int_{t_{(j-1)}}^{t_i} \lambda_j du$$

so

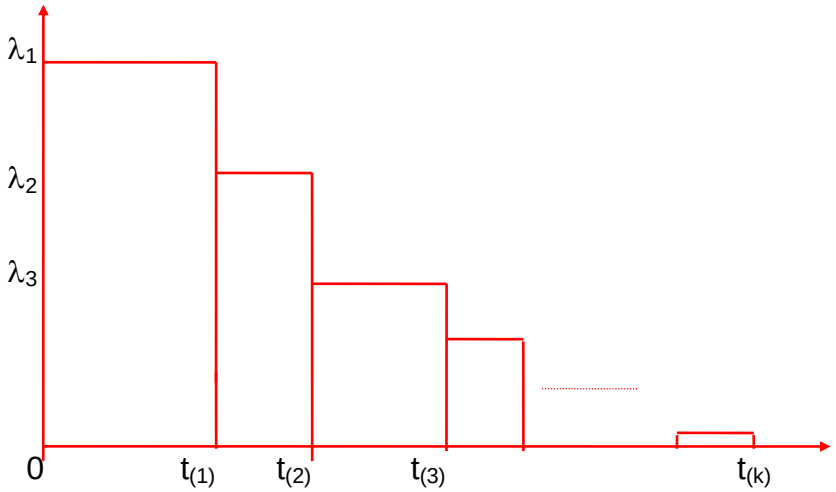
$$L = \prod_{j=1}^{k+1} \prod_{i \in A_j} \{\lambda_j e^{\underline{\beta}' \underline{x}_i}\}^{\delta_i} \exp\left[-e^{\underline{\beta}' \underline{x}_i} [\lambda_1 t_{(1)} + \lambda_2 (t_{(2)} - t_{(1)}) + ... + \lambda_j (t_i - t_{(j-1)})\right]$$

replace $\underline{\beta}$ by $\hat{\underline{\beta}}$,, then find estimates $\hat{\lambda}_1, \hat{\lambda}_2, ...$ by MLE.

$$\Rightarrow \hat{\lambda}_j = \frac{d_j}{\sum_{i \in R(t_{(j)}} e^{\hat{\underline{\beta}}' \underline{x}_i} (t_{(j)} - t_{(j-1)}) + \sum_{i \in A_j} e^{\hat{\underline{\beta}}' \underline{x}_i} (t_i - t_{(j-1)})},$$

where d_j = number of deaths at $t_{(j)}$ and noting that the term in the denominator summing over $i \in A_j$ does not appear if all lifetimes are observed.

If the estimates of λ_j suggest a pattern, e.g. roughly constant or exponentially decreasing or..... then this might suggest a parametric form for $h_0(t)$.



4.4.7 Model checking

4.4.7.1 log–log plots

e.g. two treatments, $x_1=0$ and $x_1=1$;

$h(t;x)=h_0(t)e^{\beta x_1}$, so $h_1(t)=h_0(t)$ and $h_2(t)=h_0(t)e^{\beta}$

$$-\log_e S_1(t)=H_1(t)=\int_0^t h_1(u)du$$

$$-\log_e S_2(t)=H_2(t)=\int_0^t h_2(u)du$$

so $\log_e[-\log_e S_2(t)]=\beta+\log_e[-\log_e S_1(t)]$

so if we plot $\log_e[-\log_e\{\hat{S}_j(t)\}]$ *vs.* t for both groups we should get parallel curves a distance β apart — if the curves cross then a proportional hazards model is not appropriate.

If this diagnostic test 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 (see details below in §4.4.7.3 and Collett (2003) §11.1.1) and the other option is to consider a parametric model which does not have the proportional hazards property.

4.4.7.2 Residuals

Various types of residuals can be defined for survival regression models. Full discussions are given in Collett (2003) §4 and §7 for Cox and parametric regression respectively and Everitt & Rabe-Heskith (2001) §17.5. Of particular note are *Schoenfeld Residual*s which can be obtained with the **R** function `cox.zph(.)`. Schoefeld residuals are defined only for non-censored observations and there is a separate set for each of the covariates. They are defined as 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:

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

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 $R(t_i)$ 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 shew no dependence. Deviation from independence will indicate inadequacy of the model.

4.4.7.3 Implementation in R

The non-obvious step for log-log plots needed is that to produce separate plots of the estimated survivor function for different levels of a factor the factor needs to be used in the model as a stratum indicator using the function **`strata(.)`**, i.e. the proportional hazards model is fitted [almost] separately within each level of the factor. This step cannot be by-passed. For example, for the lymphoma data where there are two levels of the variable `stage` (which must be converted to a factor) the following:–

```
> library(survival)
Loading required package: splines
> load("lymphoma.Rdata")
> attach(lymphoma)
> stage<-factor(stage)
> lymph.cox<-coxph(Surv(time,censor)~strata(stage))
> lymph.cox
Call:coxph(formula=Surv(time,censor) ~ strata(stage))

Null model
  log likelihood= -17.77164
  n= 18
 > plot(survfit(lymph.cox))
```

–:will produce two separate K-M plots.

To produce a plot of the log survivor function the vertical scale needs to be changed appropriately:–

```
>plot(survfit(lymph.cox),fun="cloglog",lty=2:3,col=4:5)
```

where the line styles and colours have been chosen with the parameters `lty` and `col` (this also uses a log scale for the horizontal axis which is useful since it tends to 'straighten' the plot and so makes it easier to judge parallelism). To make the plot more visible it is usually best to adjust the thickness of the lines with the `lwd` parameter, perhaps `lwd=3`. To find out more about plotting parameters type `help(par)`.

4.4.7.4 S-PLUS implementation

In S-PLUS log-log plots can be drawn using the same commands as in **R** given in §4.4.7.3.

4.4.8★ Time-dependent covariates

It is possible to extend the model to 'time-dependent covariates'

e.g. $h(t;x)=h_0e^{\beta x+\gamma xt}$

which is equivalent to allowing β to change with time. This may be appropriate if x=0 means receiving treatment, x=1 means receiving no treatment (so without treatment patient gets worse and worse).

More generally, we have if $\underline{x}=\underline{x}(t)$ then $h(t;\underline{x})=h(t;\underline{x}(t))=h_0(t)e^{\underline{\beta}'\underline{x}(t)}$ and this gives

$$S(t) = \exp\left\{-\int_0^t e^{\underline{\beta}'\underline{x}(u)} h_0(u)du\right\}$$

so the survivor function depends not only on the baseline hazard function $h_0(t)$ but also on the values of the time-dependent covariates over the interval (0,t).

One application is to transplant studies. Define

z(t)=0 before transplant

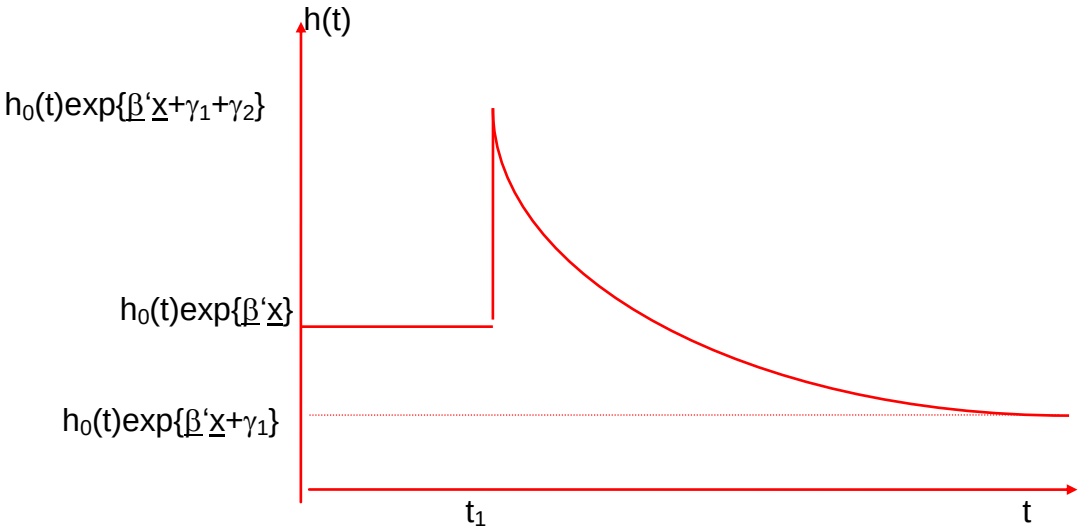
z(t)=1 after transplant

and then $h(t)=h_0(t)e^{\underline{\beta}'\underline{x}+\gamma z(t)}$ where $\underline{x}$ are non-time-dependent covariates. Then the effect of the transplant is measured by γ.

A more sophisticated model is given by Cox & Oakes (1984) with

$$h(t;\underline{x}) = h_0(t)\exp\{\underline{\beta}'\underline{x} + z(t)[\gamma_1 + \gamma_2 e^{\gamma_3(t-t_1)}]\}$$

with z(t) = 0 for $t < t_1$, 1 for $t > t_1$



In this model the effect of the transplant is to increase the hazard to $h_0(t)\exp\{\underline{\beta}'\underline{x}+\gamma_1+\gamma_2\}$ (assuming $\gamma_1+\gamma_2>0$) from which it decreases exponentially to $h_0(t)\exp\{\underline{\beta}'\underline{x}+\gamma_1\}$ which is less than the initial hazard if $\gamma_1<0$.

Tasks 4

1) The **R** function `survreg()` for fitting parametric regression models allows a choice of distributions with the parameter `dist`. These include `"weibull"`, `"exponential"`, `"gaussian"`, `"logistic"`, `"lognormal"` and `"loglogistic"`. Which of these distributions will give proportional hazards models if all parameters are to be estimated?

2) Which if the choices for the parameter `dist` will give proportional hazards models if one or more of the parameters are fixed (i.e. specified as having a fixed numerical value and are not estimated)?

3) The table below gives details of a proportional hazards model fitted to some data obtained from patients being treated for kidney failure where 'survival time' is in terms of time to relapse.

Variable	Coefficient	Standard Error	χ^2 statistic (using L.R.T)
Treatment 0 = Treat A 1 = Treat B	-1.63	0.75	4.71
Age (years)	-0.003	0.024	0.01
Sex 0 = female 1 = male	0.67	0.32	3.91
Obesity 0 = no 1 = yes	0.0092	0.0045	4.44
Duration of symptoms prior to treatment (months)	-0.003	0.075	0.01

Describe the effects of treatment and additional covariates on time to relapse, giving point and interval estimates of hazards ratios where appropriate.

Exercises 2

1) The table below gives some details of fitting a proportional hazards regression model to times to recurrence of a certain disease. The data were obtained during a randomised clinical trial of a new treatment. The factors investigated were treatment (coded by $x_1 = 0$ for placebo, $x_1 = 1$ for treatment), stage of disease (coded by $x_2 = 0$ for stage I, $x_2 = 1$ for stage II, $x_2 = 2$ for stage III) and the interaction

between treatment and stage of disease (coded by x_3 where $x_3 = x_1 \times x_2$

	variable	coefficient	standard error
Treatment	x_1	–0.18	0.10
Stage	x_2	+0.32	0.21
Interaction	x_3	–0.66	0.11

i) Specify the form of the proportional hazards model used for this analysis in terms of the baseline hazard function $h_0(t)$ and the covariates.

ii) Describe in detail the effects of these factors on the time to recurrence of the disease.

iii) Show diagrammatically the form of the relationship between the survivor functions and the stage of the disease for the two different treatment groups.

2) The data in the file `AHprostate.Rdata` have been adapted from Andrews and Herzberg (1985) and give results of a trial on treatments for prostate cancer. Various covariates were recorded. The variables in the data file are given in the table below.

Variable	Description	Levels
Stage	Stage of Disease	3=No evidence of distant metastasis, 4=evidence of distant metastasis
RX	Treatment Groups	1=Control Arm, 2=Experimental Arm
Dtime	Complete months to follow up	
Status	Survival Status	
AgeYrs	Age of Patient	89 denotes 89 or more
Wt	Weight Index	Weight (Kg) – Height (cm) + 200
PF	Performance Rating	
HX	History of cardiovascular disease	0=no, 1=yes
SBP	Systolic blood pressure	
DBP	Diastolic blood pressure	
EKG	Electrocardiogram	
HG	Serum Haemoglobin,g/100ml	
SZ	Size of primary tumour	cm2
SG	Combined index of tumour	stage and histologic grade
AP	Serum prostatic acid phosphatase	
BM	Bone metastases	0=no, 1=yes

Investigate whether there is evidence that the treatment improves survival time to follow up, making due allowance for any other prognostic factors on the subjects.

3) The data file prostatic.Rdata contains data on a double blind randomised controlled clinical trial to compare treatments for prostatic cancer. The data are extracted from Collett (2003) who gives the original reference. The data file contains records for each patient of the treatment received (coded as 0 or 1 for placebo and 1.0 mg of diethylstilbestrol respectively, treatments being administered daily by mouth), survival time from entry to trial, with a status variable indicating whether or not the observation was censored (value 0) or complete (value 1), age at entry to the trial, serum haemoglobin level in gm/100ml, size of primary tumour in cm^2 and the value of a combined index of tumour stage and grade (the Gleason Index), larger values indicating a more advanced stage of tumour.

   i) Construct Kaplan-Meier plots of the survival times for the two treatment groups.

   ii) Making allowance for the values of the various covariates, assess whether the data provide evidence that the two treatment groups experience different survival prospects.

   iii) Construct a log–log plot for treatment, averaging over other covariates.

   iv) ★Choosing any parametric regression (see Survival tasks 4) model which does ***not*** have the proportional hazards property, fit the model and assess whether this alters your conclusions reached in part ii).

   v) ★ Choosing a parametric AFT model, estimate the parameters and compare your conclusions with those from parts ii) and iv).

[credit will not be lost if parts iii) – v) are not submitted, they are for 'interest' and as an aid to those continuing to MAS6062]

4.5★ Accelerated Failure Time Regression Models

4.5.1★ Introduction

In situations where the proportional hazard assumption is violated (as illustrated in the notes in §3.5) the alternatives are to use a parametric model (which does not have the proportional hazards property of course) or to consider a class of models termed *accelerated failure time (**AFT**) models*. These are models where the survivor function for a subject with covariate **x** takes the form $S(t;\mathbf{x}) = S_0(t.e^{\beta'\mathbf{x}})$ where $S_0(.)$ is some baseline parametric survivor function. Of course for certain choices of distribution this model has the proportional hazards property, for example the exponential where $S_0(t) = \exp\{\lambda_0 t\}$ and the Weibull. For this reason an accelerated failure time model is often taken to refer to a Weibull model.

The name comes from *accelerated-life* testing, typically of electronic components, where the components under test would be subject to enhanced stress, e.g. higher voltages than those under which the components would be used. The objective would be to complete the whole experiment within a shorter time. For example old-fashioned electric light bulbs might be expected to have a mean lifetime of 1000 hours under normal operating conditions but increasing voltage and operating temperature might accelerate the time to failure so that all those under test would fail within a week and thus allow an analysis to be completed without the complication of large numbers of censored observations. In these situations it was plausible to consider that for the stressed components time itself was progressing faster and so in terms of a model that the effect of a covariate (e.g. voltage) could be regarded as multiplying the timescale by some numerical factor.

In recent years these models have been considered for use in medical situations, especially now that software for fitting them to data including censored observations has become widely available. The **R** package `eha` which has a full set of facilities dates from 2003 and the current version 1.2.18 from February 2010.

Unfortunately some texts within the medical area advocate use of AFT models when 'it is desired to speed up the time to the event' — such as time to relief of symptoms and other 'positive' events whereas proportional hazards models should be used for situations where it is required to slow down the time to a 'negative' event, such as death. This is clearly incorrect**:** – use of a particular statistical model does not influence the time of occurrence of events. The model has to be chosen so that it fits the available data (not vice versa). However, it is true that experience has shown that for experiments on a short time scale (days or a very few weeks) AFT models are often found to work well whilst for long term studies (years) proportional hazards models may be found to be preferable. There is no guarantee that this will hold for any data set that is encountered and there is no substitute for investigating models and checking validity with appropriate diagnostics (log–log plots, residual etc). In part this 'folk-lore' belief and sweeping generalisation could be a reflection that in longer term studies the numbers of actual events observed can be relatively small and so it is difficult to discover that proportional hazards models do not fit the data but in short term studies where events are relatively common it is easier to detect deviations from a model.

4.5.2★ Implementation in R

The functions for fitting AFT models are in library `eha`. The particular function for fitting a regression is `aftreg()` and it may be noted that he library also contains routines for fitting proportional hazards parametric and non-parametric regression models (`aftreg.fit()` and `coxreg()` respectively. The second of these is an alternative to `coxph()` but care must be used when comparing them since the parameterisation is different. The other functions in the library (especially for diagnostics) can be found by the command `library(help=eha)` and so are not described here.

4.5.3★ Example

This is the same example as analysed with `coxph()` in §4.4.5.1.

First, a model with the default Weibull distribution is fitted and then one using a Gompertz distribution. You are invited to compare the results from fitting these two models and that given in §4.4.5.1 using appropriate diagnostics. Not surprisingly, the results from the non-parametric proportional hazards model and from the Weibull accelerated failure time model are close (though parameterized differently) since the Weibll model belongs to both families.

```
> library(eha)
Loading required package: survival
Loading required package: splines

> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regcox<-coxph(wbcleuk.sv~log.wbc.)
> summary(wbcleuk.regcox)
> load("wbcleuk.Rdata")
> attach(wbcleuk)
> wbcleuk.sv<-Surv(survival)
> wbcleuk.regaftW<-aftreg(wbcleuk.sv~log.wbc.)
```

```
> summary(wbcleuk.regaftW)
Call:
aftreg(formula = wbcleuk.sv ~ log.wbc.)

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
log.wbc.            3.589     0.882     2.415     0.207     0.000

log(scale)                    7.440  1702.521     1.119     0.000
log(shape)                    0.105     1.110     0.187     0.574

Events                    18
Total time at risk          1154
Max. log. likelihood      -84.309
LR test statistic         15.4
Degrees of freedom        1
Overall p-value           8.6824e-05
>
> wbcleuk.regaftG<-aftreg(wbcleuk.sv~log.wbc., dist="gompertz")
> summary(wbcleuk.regaftG)
Call:
aftreg(formula = wbcleuk.sv ~ log.wbc., dist = "gompertz")

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
log.wbc.            3.589     0.872     2.393     0.240     0.000

log(scale)                    8.392  4412.587     1.131     0.000

 Shape is fixed at  1

Events                    18
Total time at risk          1154
Max. log. likelihood      -84.892
LR test statistic         16.5
Degrees of freedom        1
Overall p-value           4.97646e-05
>
```

4.6 Summary & Conclusions

Regression models

- allow individual hazards linked by covariates
- Investigate prognostic factors
  - Factors of interest
  - Allow for covariates
  - More precise analysis
  - Could model mean of survival distribution
    - e.g. exponential regression
  - Better to model hazard function
    - Model $h(t,\underline{x})=h_0(t)\exp\{\underline{b}\underline{x}\}$ ensures $h(t,\underline{x})>0$
  - Estimation by [numerical] MLE or partial MLE
  - Can estimate survival times for given covariates
  - Parametric regression models are available in **R**, S-PLUS and MINITAB but not SPSS.
    - In these cases the logarithm of the mean is modelled as a linear function of the covariates which ensures that the estimate mean is positive.
  - Semi-parametric proportional hazards Cox regression models are available in **R**, S-PLUS and SPSS but not MINITAB.

Proportional Hazards Models

- Only models dependence on covariates
- No statements about survival times
- Only effect of covariates on hazards
- Estimation by maximum partial likelihood
- Check proportional hazards by log-log plots
- May suggest parametric modelNo allowance for individual variability
- i.e. no term in σ_i^2 for i^{th} individual
- Frailty models do allow for this (some facility in **R** and S+)

Accelerated failure time models

- Accelerated failure time models may provide an alternative in cases where the proportional hazards assumption is untenable. Such models are available in **R**.

Further family of models

- A further family of model is the family of proportional odds models where the odds ratio of surviving beyond a time t for two individuals with different covariates is independent of t,

$$\text{i.e.} \quad \frac{S(t;x)}{1-S(t;x)} = \frac{S(t)}{1-S(t)}\exp\{\beta' x\}$$

  This class of models has been studied by some authors but facilities for fitting them to censored data are currently not available in **R**.

Tasks 5

1) The data given below represent survival times for lymphoma patients according to the stage of tumour (where * denotes a censored value):

Stage 3	6	20	42	43*	169*	207	253	255*		
Stage 4	4	10	20	21*	30	33*	43*	46	110	235*

i) Compute the Kaplan-Meier product limit estimates of the survivor functions for stage 3 and stage 4 separately.

ii) Provide estimates of the two cumulative hazard functions and comment on any differences.

iii) By using the log-rank test, compare the survival distributions for the two stages.

2) ⋆ In an accelerated-life survival model the survivor function for an individual with covariate x satisfies

$$S(t:x) = S_0(te^{\beta x}),$$

where $S_0(t)$ is some baseline survivor function.

a) ⋆ Show that the corresponding hazard function satisfies

$$h(t:x) = e^{\beta x}h_0(te^{\beta x})$$

where $h_0(t)$ is the baseline hazard function for $S_0(t)$.

b) ⋆ In a trial where n independent patients, with covariate values $x_1,x_2,\ldots,x_n$ enter at the same time, suppose that all death times are observed and that $S_0(t) = e^{-\lambda t}$ $(t>0)$. Show that the survival time T has an exponential form, and is of proportional hazard form.

c) ***** Show that the distribution of the time to the first death in the trial is exponential with mean

$$[\lambda\Sigma\exp(\beta x_i)]^{-1}$$

d) ***** Show that the probability that the j^{th} patient is the first to die is given by $\exp(\beta x_j)/\Sigma\exp(\beta x_i)$

4) ⋆ The **R** function `aftreg()` in library `eha` fits parametric accelerated failure time models. The parameter `dist` offers a choice of parametric distributions between `"weibull"`, `"gompertz"`, `"ev"`, `"loglogistic"` and `"lognormal"`.

a) How can this be used to fit an exponential distribution?

b) Which of these distributions also a proportional hazards model?

5★ Competing Risks

5.1★ Introduction

The survival models considered up to now consider situations where there is a single event of interest and the analysis is concerned with investigating the properties of the single distribution of the failure time distribution or the dependence of the time to this event upon various covariates. This section provides a brief introduction to more complex situations where individuals are exposed to risks of different types of events and the time measured is that to the first of these to occur. One example might be when there are different 'causes' of failure, e.g. death following a liver transplant might be from rejection of the organ or from an infection. The dependence of the time to death from different causes on covariates such as tissue type, blood group and ethnicity might be different for the different causes.

At first sight it is tempting when considering one particular cause (e.g. rejection) to regard observations where the cause of failure is not of interest (e.g. from infection) as censored, together with those subjects who are still alive at the end of the study. However, care must be taken in this. While it is possible to use some tools from survival analysis described in earlier chapters, such as log rank tests and Cox proportional hazards regression extreme care must be used with others (e.g. Kaplan-Meier estimates). The key problem is essentially that in the presence of competing risks the probability of failure from the r^{th} cause by time t does not tend to 1 as t increases in the presence of competing risks other than the r^{th}. Also note that if a failure occurs then we observe which cause is responsible but if an observation is censored then we do not. The essential point is that the censoring distribution is not independent of the time to failure from a competing event. In single event survival analysis an assumption of Kaplan-Meier estimation is that

individuals who are censored could in principle fail at some later time but if an individual fails from an event not of interest but is treated as if it were censored from the point of view of the event of interest it certainly cannot fail at some future time from the event of interest. This means that Kaplan-Meier estimation over-estimates the [probability of failure and this bias is greater if the competing events have greater hazards than the one of interest.

5.2* Basic terminology

Let T be the time to failure from any of K different causes, r =1, …, K, and R the actual cause (so completed observations of failure times consist of pairs (T,R) and we may also have uncompleted times, i.e. censored observations after times c_i. Let t_j, j = 1,…,N be the ordered distinct times of failure, and d_{rj} be the number of failures from cause r at time t_j, then $d_j = \Sigma_r d_{rj}$ is the total number of failures at time t_j. Note that $d_{rj} \neq 0$ for at least one r and that typically $d_{rj} \neq 0$ for only one r. $d = \Sigma_j d_j$ is the total number of failures. Let n_j be the number of individuals at risk at time t_j (note that these are at risk of failure from any of the causes).

The *cumulative incidence function* for cause r is

$$I_r = P[T \leq t, R = r], \text{ for } r = 1,\ldots,r.$$

The *cause specific hazard function* is $\lambda_r(t)$ given by

$$\lambda_r(t) = \lim_{\delta t \to 0}\left(\frac{P\left[T \leq t + \delta t, R = r \mid T > t\right]}{\delta t}\right)$$

The *cumulative cause specific hazard function* is given by

$$\Lambda_r(t) = \int_0^t \lambda_r(u)du \text{ and define}$$

$$S_r(t) = \exp(-\Lambda_r(t))$$

the *r^{th} competing event survival distribution.* Note that

$$S_r(t) = \int_t^\infty \lambda_r(u) S_r(u) du$$

(c.f. §2.1.3) and that

$$S(t) = \prod_{r=1}^{K} S_r(t) = \exp\left(-\sum_{r=1}^{K} \Lambda_r(t)\right)$$

is the probability of not having failed from any cause at time t. Note that

$$I_r(t) = \int_0^t \lambda_r(u) S(u) du$$

and that $I_r(\infty) = P[R = r]$ so it is not a proper distribution function; it is sometimes referred to as a 'subdistribution function'. Note that the integral above involves S(t) and not $S_r(t)$ because $S_r(t)$ is the probability of failing from cause r after time t but the probability of failing from cause r after t requires not failing from *any cause* before time t. Consequently $I_r(t) \neq 1 - S_r(t) = F_r(t)$. Generally it is $I_r(t)$ that is of interest rather than $S_r(t)$ and this is a key reason that naïve Kaplan-Meier estimates have to be used with care.

Clearly the overall hazard function of the distribution of T is $\lambda(t) = \Sigma\lambda_r(t)$. The conditional probability that failure is from cause r given that failure is before time t is

$$\phi_r^I(t) = P[R = r \mid T \leq t] = \frac{I_r(t)}{\sum_{j=1}^{K} I_j(t)}$$

and the conditional probability of failure from cause r within a short interval given survival up until time t is

$$\phi_r^\lambda(t) = \lim_{\delta t \to 0} P[R = r \mid t < T \leq t + \delta t]$$

$$= \lim_{\delta t \to 0}\left(\frac{P\left[t < T \leq t + \delta t, R = r \mid T > t\right]}{P[T \leq t + \delta t \mid t > t]}\right) = \frac{\lambda_r(t)}{\sum_{j=1}^{K} \lambda_j(t)}$$

5.3★ Estimation of hazard and survivor function

S(t) can be estimated by the usual overall Kaplan-Meier estimate

$$\hat{S}(t) = \prod_1^N (1 - \frac{d_j}{n_j})$$

The naïve Kaplan-Meier estimates

$$\hat{S}_r(t) = \prod_1^N (1 - \frac{d_{rj}}{n_j})$$

are generally not of interest since the comments above shew $1 - \hat{S}_r(t)$ is a biased estimate of the probability of failing from cause r by time t. Of more interest is an estimate of $I_r(t)$. We can estimate $\lambda_r(t)$ by

$$\hat{\lambda}_r(t_j) = d_{rj} \big/ n_j$$

and if $p_r(t_j)$ is the unconditional probability of failing at t_j then

$$\hat{p}_r(t_j) = \hat{\lambda}_r(t_j)\hat{S}(t_{j-1}), \text{ with } \hat{S}(t_0) = 1.$$

and the cumulative incidence function for cause r is estimated by

$$\hat{I}_r(t) = \sum_{j;\, t_j \le t} \hat{p}_r(t_j)$$

Standard packages which handle survival data can be used to calculate the naïve Kaplan-Meier estimates but more specialist ones specifically providing competing risks facilities are required or estimation of $I_r(t)$ (see below).

5.4★ Analysis of effects of covariates

Note first that the cause specific hazards are directly estimable standard methods which model the effects of covariates on hazards are readily adapted for use in the presence of competing hazards, provided some care in the interpretation is taken. In the case of a small number (i.e. one or two) factors with two or three levels then it is possible to modify the standard log-rank test to test for equality of groups. For regression on continuous covariates the dependence of the individual hazard rates on covariates can be modelled with Cox proportional hazards techniques. Provided the data are arranged in an appropriate augmented form (see below) it is possible to allow for the effects of covariates to be different or identical for different hazards and to test for equality of effects. The usual arrangement of a data file is that each individual contributes one row, with variables indicating time of failure (or censoring), a cause indicating the cause or censored and the various covariates. Thus the cause variable will have K+1 distinct values. For basic analyses e.g. naïve Kaplan-Meier estimation) the cause variable will be the censoring indicator, defining the value for the event of current interest as the code to indicate the event has occurred (see below). The appropriate augmented data file if there are K competing risks is obtained by repeating each individual's data in K rows but adding a censoring indicator, status, to which has values 0 for all causes other than the one causing failure when it has value1. Censored failure times have 0 for status in all rows. Further details are beyond the scope of this brief introduction.

5.5⋆ Implementation in R

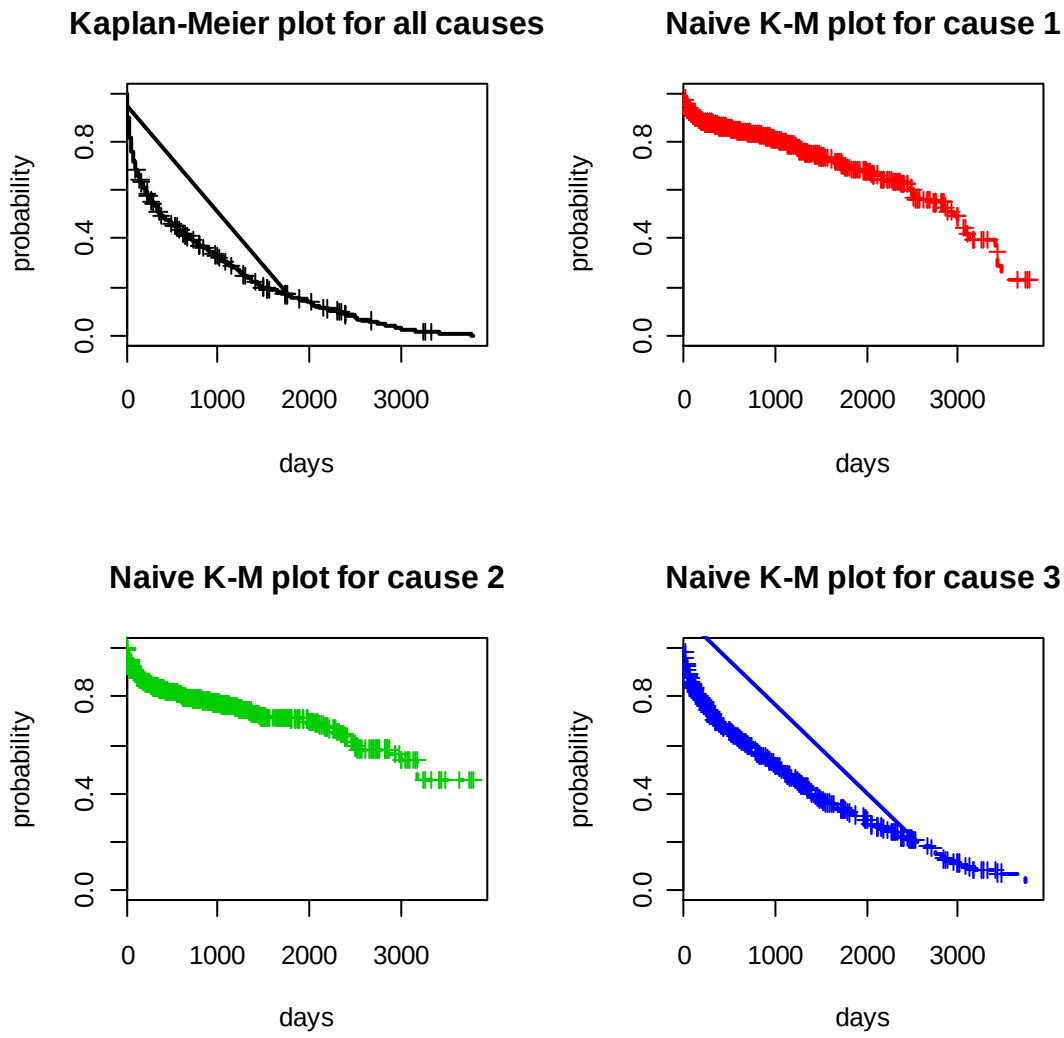
There are a several specialist libraries which provide facilities for handling problems of competing risks. Here we illustrate the library **`cmprsk`**, this has to be downloaded from the CRAN website (or a local mirror site). It automatically loads the **`survival`** library which is bundled with the standard installation of **R** so **`survival`** does not need to be opened first.

5.5.1 Example on organ transplants

These data are semi-artificial. For the purposes of illustration they have been adapted from a much larger data set from a ten year study on survival rates following an organ transplant. The data are provided by NHS Blood and Transplant (www.nhsbt.nhs.uk); codes have been changed and covariates removed. The data set used has 1086 observations of survival times in days (variable **`days`**) following an organ transplant with three different possible causes of fatality (coded as 1, 2 and 3). The 144 censored observations are indicated by a value 0 for variable **`cause`**.

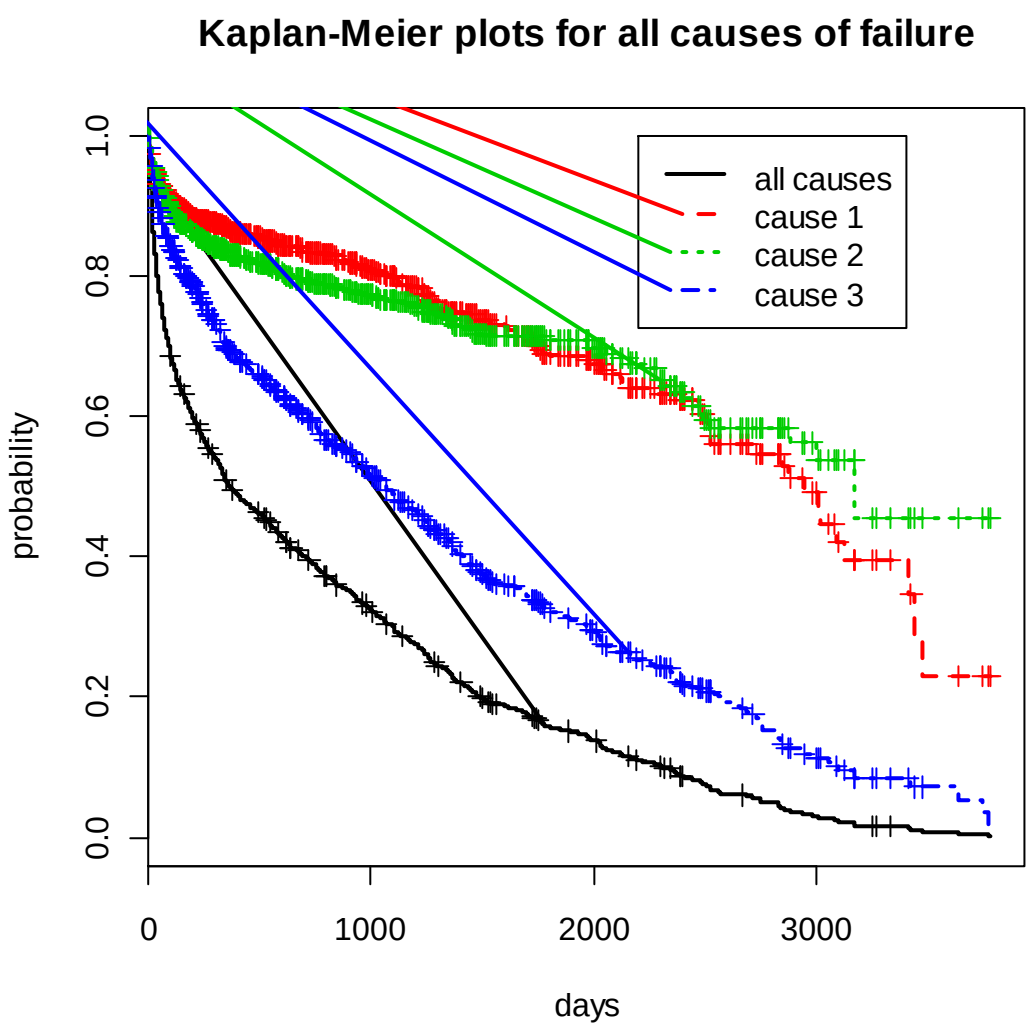
```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

. . . . . .
. . . . . .
. . . . . .

[Previously saved workspace restored]

> library(cmprsk)
Loading required package: survival
Loading required package: splines
> attach(organ)
> organ[1:5,]
  cause days
1     0 1754
2     0   70
3     0   29
4     0 2671
5     0  522
```

```
> par(mfrow=c(2,2))
> plot(survfit(Surv(days,cause!=0)~1),col=1,lty=1,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Kaplan-Meier plot for all causes")
>
> plot(survfit(Surv(days,cause==1)~1),col=2,lty=2,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 1")
>
> plot(survfit(Surv(days,cause==2)~1),col=3,lty=3,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 2")
>
>plot(survfit(Surv(days,cause==3)~1),col=4,lty=4,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Naive K-M plot for cause 3")
>
```



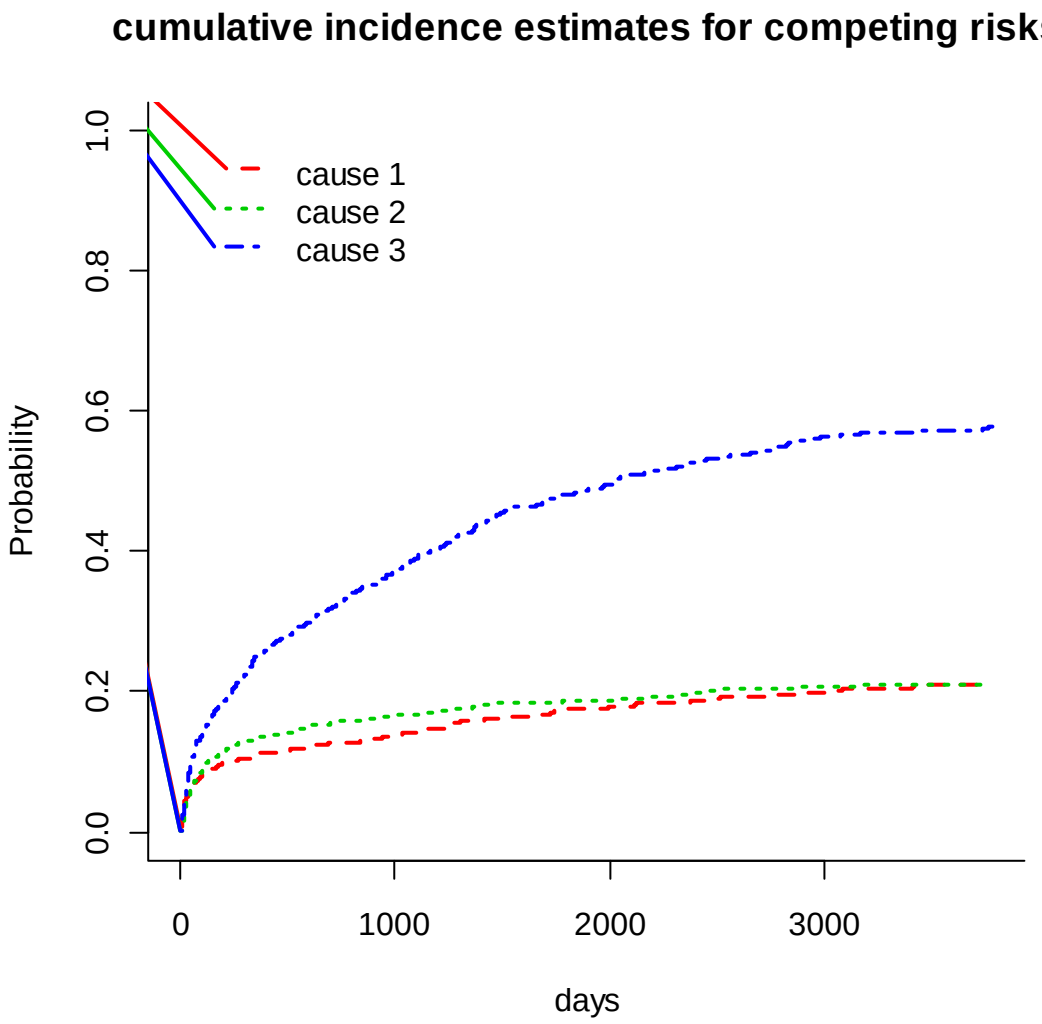
It is clear from the plots above that there is a difference in the survival patterns for the three causes but it is dangerous to draw very specific conclusions on the nature of the patterns since these naïve estimates are biased, possibly to different degrees.

```
>
> par(mfrow=c(1,1))
>
> plot(survfit(Surv(days,cause!=0)~1),col=1,lty=1,lwd=2,conf.int=0,
+ ylab="probability",xlab="days")
> title("Kaplan-Meier plots for all causes of failure")
>
> lines(survfit(Surv(days,cause==1)~1),col=2,lty=2,lwd=2,conf.int=0)
> lines(survfit(Surv(days,cause==2)~1),col=3,lty=3,lwd=2,conf.int=0)
> lines(survfit(Surv(days,cause==3)~1),col=4,lty=4,lwd=2,conf.int=0)
> leg.txt=c("all causes","cause 1","cause 2", "cause 3")
> legend(2200,1.0,leg.txt,col=1:4,lty=1:4,lwd=2)
>
```



Now we look at the cumulative incidence functions

```
> organ.cmprsk<-cuminc(days, cause)
> plot(organ.cmprsk,
+ main="cumulative incidence estimates for competing risks",
+ xlab="days", lty=2:4,lwd=2,col=2:4,
+ curvlab=c("cause 1", "cause 2", "cause 3"))
>
> curvlab=c("cause 1", "cause 2", "cause 3"))
```



It is clear here that there is a substantial difference between cause 3 and the first two which are similar. Further, it can be seen how different these are from the naïve Kaplan-Meier estimates (which would be reflections about a vertical axis of the plots above).

Exercises 3

1) The table below gives the survival times in weeks of 28 patients with kidney tumours who were randomized to receive either radiation therapy alone or radiation plus chemotherapy.
Radiation alone: 6* 6 9* 10* 11 11* 19 19* 20* 25 37 38 39*
Radiation + chemo: 1 2 2 5 5 8 10 12 15* 21 22 22 27 30 40*
(* indicates a censored observation)

(The data are in file `kidneytumour.Rdata`)

i) Calculate the Kaplan-Meier estimates of the two survivor functions for the two treatments

ii) Without making any assumptions on the form of the survival distributions estimate the median survival times of subjects receiving the two treatments.

iii) Assuming that the survival times are exponentially distributed Ex(λ_1) and Ex(λ_2) respectively, estimate λ_1 and λ_2 and calculate approximate 95% confidence intervals for them.

iv) Using a log-rank test assess the effect of chemotherapy on the survival times of these patients.

v) Using a likelihood ratio test (assuming the exponential model is valid) assess the effect of chemotherapy on the survival times of these patients

vi) If the subjects had not been followed up for more than 20 weeks how would your conclusions based on the two tests have been altered?

2) The data (from Collett, 2003) given below represent survival times in days of 26 patients randomized to one of two forms of chemotherapy following surgery for ovarian cancer, where status

records whether the observation is censored (status = 0) or complete (status =1), (The data are given in file `ovarian.Rdata`):

Treatment A		Treatment B	
time	status	time	status
59	1	353	1
115	1	365	1
156	1	377	0
268	1	421	0
329	1	464	1
431	1	475	1
448	0	563	1
477	0	744	0
638	1	769	0
803	0	770	0
855	0	1129	0
1040	0	1206	0
1106	0	1227	0
Totals	***6725 7***	***8863 5***	

i) Compute the Kaplan-Meier product limit estimates of the survivor functions for treatments A and B and provide estimates of the median survival times based upon the Kaplan Meier estimates.

ii) Using a log-rank test, assess whether the two treatment differ in effectiveness.

iii) As well as the data in the table above the datafile contains values of three covariates. Using an appropriate regression model, assess whether the treatments differ in effectiveness after making allowance for any effects attributable to the covariates.

3) The table below gives the life-times (in hours) of valves used in kidney dialysis machines. Two types of valves (thirty of each) were tested on an accelerated rig which operated a maximum of sixty

valves under increased pressure and flow. A * indicates that the valve was still functioning at the end of the 36 hours of the test.

Failure times of dialysis valves									
Valve Type A									
6*	6*	6*	6*	8	9	11	12	12*	12*
13	15	16	16	16	17	18	20	20	20
26	30*	30*	30*	32	35	36*	36*	36*	36*
Valve Type B									
1	2	3	3	4	5	5	6*	6*	7
8	10	15	20	25	28	29	30	30*	30*
30*	30*	30*	30*	30*	30*	30*	31	36*	36*

i) Calculate and plot the Kaplan-Meier estimates of the survivor functions for the two types of valve using firstly a natural and secondly a logarithmic vertical scale (datafile `dialysis.Rdata`).

ii) Using a log rank test assess whether the survival patterns of the two valve types differ.

iii) Assuming that the survival times are exponentially distributed with rates λ_j, j=1, 2 estimate λ_1 and λ_2 and hence the median survival times and provide standard errors for the median survival times for each type.

iv) Using your results from part (iii) test whether the two types of valves have equal median survival times.

v) Is there any reason to doubt the validity of exponential models for either of these two survival distributions? Justify your answer.

vi) Is there any reason to doubt the validity and suitability of the log rank test for these data? Justify your answer.

vii) Using a Weibull regression model assess whether there is a difference in survival patterns between the two types of valve.

4) In a study to determine the efficacy of two new strains (A and B) of macrophages fifty-four bacterial colonies were randomized to one of three groups to receive inoculation by either a control strain or one of the two new strains. The table below gives the times to extinction in hours of the colonies (datafile `macrophages.Rdata`).

Control,treatment=1	Strain A treatment=2	Strain B treatment=3
3	5	8
5	6	11
6	6	13
6	7	14*
6	7	15*
9	8	18
12*	9	18
12	13	20
23	13*	20
26	15	20
27	17*	21
38	17	22
40	21	24*
48	23	25*
59	23	25
71*	23	27*
75*	42*	28
76*		29
80*		

Table 1: Times to extinction of bacterial colonies (* censored)

i) Calculate and plot the Kaplan-Meier estimates of the survivor functions for the two types of valve using firstly a natural and secondly a logarithmic vertical scale.

ii) Using the Kaplan-Meier calculations estimate the median survival times of the three treatment groups.

iii) Do the two graphical displays (on natural and logarithmic vertical scales) suggest that exponential models are appropriate or inappropriate for the three groups? Justify your answer.

iv) Using a log rank test assess whether the survival patterns of the three treatment groups differ.

v) Is there any reason to doubt the validity and/or suitability of the log rank test for these data?

vi) Using appropriate regression models (either or both of semi-parametric and parametric) investigate whether there is evidence of a difference in effectiveness of the two strains of macrophage and whether these are different from the control treatment.

5) The data for this project are contained in the file `hips.Rdata`. The data have been adapted from a dataset provided by Dr J.M. Wilkinson and arise from a study of factors influencing implant failure due to osteolysis after total hip anthroplasty (THA). Between February 2000 and February 2001 about 180 subjects with failed implants and about 300 with [apparently] intact implants were recruited. Some of the 300 were discovered on examination to have failed implants so the study consisted of 214 with failed and 267 intact implants. Various factors and covariates were recorded on most of the subjects as listed below.

i) Particular interest is in whether there are genetic as well as environmental factors affecting implant failure after THA. Particular genes of interest are TNF-238 and TNF-308 and the presence or absence of the ‘a’ allele of each of these was recorded by DNA examination.

code	**variable**
studyno	study number
group	0=intact, 1=failed
gender	0=female, 1=male
tnf_308a	0=absence of TNF-308a allele, 1=presence
tnf_238a	0=absence of TNF-238a allele, 1=presence
ht	height
wt	weight
bmi	body mass index
smoke	0=non-smoker, 1=smoker
psoriasi	0=no psoriasis, 1=psoriasis
interpcu	measure on implant wear
idate	date of implant
iyear	year of implant
ageatind	age at implant
age55	0=under 55, 1 = over 55
isurgeon	surgeon
iimplant	type of implant
idatloos	date of examination
lysisfre	length of osteolysis-free survival of implant
irevdate	date of revision surgery
revfrees	length of revision-free survival of implant

Notes & Solutions for Tasks 1

1) *Derive a clinical life table for [at least the first five years of] the survival data of patients with angina pectoris given in Example 1 in the notes and reproduced below.*

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*	

Note number died in first interval is 2418 – 1962 = 456 and in second interval it is 1962 – 39 – 1697 = 226 &c.

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 hazard function
	Died	withdrawn						
x to x+1	d_x	w_x	n_x	n_x'	q_x	p_x	l_x	h_x
0 – 1	456	0	2418	2418	0.1886	0.8114	100	0.208
1 – 2	226	39	1962	1942.5	0.1163	0.8837	81.1	0.123
2 – 3	152	22	1697	1686	0.0902	0.9098	71.7	0.095
3 – 4	171	23	11523	1511.5	0.1131	0.8869	65.2	0.12
4 – 5								
5 – 6								
6 – 7								
7 – 8								
8 – 9								
9 – 10								
10 –11								
11–12								
12–13								
13–14	9	27	95	81.5	0.1104	0.8896	18.4	0.117
14–15	6	23	59	47.5	0.1263	0.8737	16.4	0.135
15+	30	0					14.3	

2) *Show that the probability that an individual lives longer than t_1+t_2 years given he has attained t_1 years is equal to the unconditional probability that he survives at least t_2 years if and only if the survival distribution is of exponential form.*

[Note the well-known the result that:—

if $g(t_1+t_2)=g(t_1)+g(t_2)$ for all t_1 and t_2, then $g(t)=\alpha t$ for some real α.]

If survival distribution is E(λ) then $S(t)=e^{-\lambda t}$

so $P[T \geq t_1 + t_2 \mid T \geq t_1] = P[T \geq t_1 + t_2]/ P[T \geq t_1]$

$= \exp\{-\lambda(t_1 + t_2)\}/ \exp\{-\lambda t_1\} = \exp\{-\lambda t_2\} = S(t_2) = P[T \geq t_2]$

If $P[T \geq t_1 + t_2 \mid T \geq t_1] = P[T \geq t_2]$

then $P[T \geq t_1 + t_2]/ P[T \geq t_1] = P[T \geq t_2]$ so $S(t_1 + t_2) = S(t_1)S(t_2)$

so $\log[S(t_1 + t_2)] = \log[S(t_1)] + \log[S(t_2)]$

so $\log[S(t)] = \alpha t$ for some α so $S(t) = e^{\alpha t}$ and since T is +ve and $0 \leq S(t) \leq 1$ we must have $\alpha<0$ so $S(t) = e^{-\lambda t}$ for $\lambda>0$.

Notes & Solutions for Tasks 2

1) The data below give the times of remission (in weeks) of two groups of leukaemia patients randomized to a treatment or a control group. [indicates a censored value]*

1	*drug-6-MP*	*6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13,*
		16, 17, 19*, 20*, 22, 23, 25*, 32*,*
		32, 34*, 35**
2	*control*	*1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8,*
		11, 11, 12, 12, 15, 17, 22, 23

1) Obtain (by hand and by computer package) and plot the Kaplan-Meier survivor functions for the data (obtaining separate functions for control and drug patients).

(The data are given in file leukaemia remission times on the relevant web pages)

a) First, the treated:

j	$\mathbf{t_{(j)}}$	$\mathbf{l_j}$	$\mathbf{r_j}$	$\mathbf{d_j}$	1	$0 \le t < 6.0$
1	6	0	21	3	0.857	$6.0 \le t < 7.0$
2	7	1	17	1	0.807	$7.0 \le t < 10.0$
3	10	1	15	1	0.753	$10.0 \le t < 13.0$
4	13	2	12	1	0.690	$13.0 \le t < 16.0$
5	16	0	11	1	0.627	$16.0 \le t < 22.0$
6	22	3	7	1	0.538	$22.0 \le t < 23.0$
7	23	0	6	1	0.448	$23.0 \le t$

Next the controls:

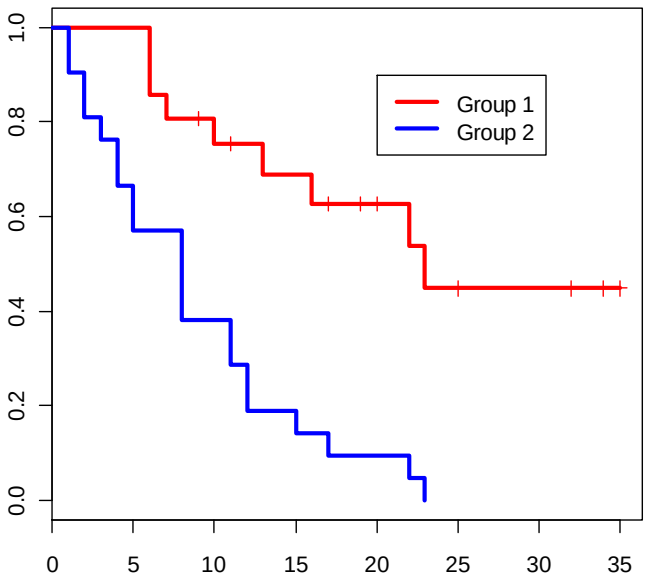
j	$\mathbf{t_{(j)}}$	$\mathbf{l_j}$	$\mathbf{r_j}$	$\mathbf{d_j}$	1	$0 \le t < 1.0$
1	1	0	21	2	0.905	$1.0 \le t < 2.0$

2	2	0	19	2	0.810	2.0≤t<3.0
3	3	0	17	1	0.762	3.0≤t<4.0
4	4	0	16	2	0.666	4.0≤t<5.0
5	5	0	14	2	0.571	5.0≤t<8.0
6	8	0	12	4	0.381	8.0≤t<11.0
7	11	0	8	2	0.286	11.0≤t<12.0
8	12	0	6	2	0.190	12.0≤t<15.0
9	15	0	4	1	0.143	15.0≤t<17.0
10	17	0	3	1	0.095	17.0≤t<22.0
11	22	0	2	1	0.048	22.0≤t<23.0
12	23	0	1	1	0	23.0≤t

The transcript from **R** is given below.

```
> library(survival)
Loading required package: splines
> attach(leukrem)
> leuk.sv<-Surv(time,censor)
> leuk.fit<-survfit(leuk.sv~group)
> summary(leuk.fit)
Call: survfit(formula = leuk.sv ~ group)

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                group=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
```

```
   12       6       2   0.1905  0.0857      0.07887         0.460
   15       4       1   0.1429  0.0764      0.05011         0.407
   17       3       1   0.0952  0.0641      0.02549         0.356
   22       2       1   0.0476  0.0465      0.00703         0.322
   23       1       1   0.0000     NaN           NA            NA
> plot(leuk.fit, col=c("red","blue"), lwd=3)
> legtext=c("Group 1","Group 2")
> legend(20,0.9, legtext,lwd=3,col=c("red","blue"))
```



2) Estimate the median survival times for the two groups.

Note that **R** does not interpolate to estimate the median (though in passing note that SPSS does do this (not shewn here)). For group one we have that $\hat{S}(22) = 0.538$ and $\hat{S}(23) = 0.448$ and we want $\tilde{\tilde{t}}$ such that $\hat{S}(\tilde{\tilde{t}}) = 0.5$ so we have

$\tilde{\tilde{t}} = 22 + (0.538 - 0.5)*(23 - 22)/(0.538 - 0.448) = 22.42$

For group two we have $\hat{S}(5) = 0.571$ and $\hat{S}(8) = 0.381$ so

$\tilde{\tilde{t}} = 5 + (0.571 - 0.5)*(8 - 5)/(0.571 - 0.381) = 6.12$

2) *In an Institute for Medical Research and Public Health in Australia a study was reported in 2005 in which the survival of teaspoons was investigated. 102 teaspoons were purchased and discreetly numbered, 16 of these were of higher quality that the other 86. Equal numbers of teaspoons of each type were placed in eight tearooms around the institute, with equal numbers in communal rooms and programme-linked rooms. Audits were taken at various times during the following five months and the day on which a teaspoon went missing was recorded. The data are given in the dataset spoons.Rdata, with variables indicating day of disappearance, category of tearoom (1 for communal room) and type of teaspoon.*

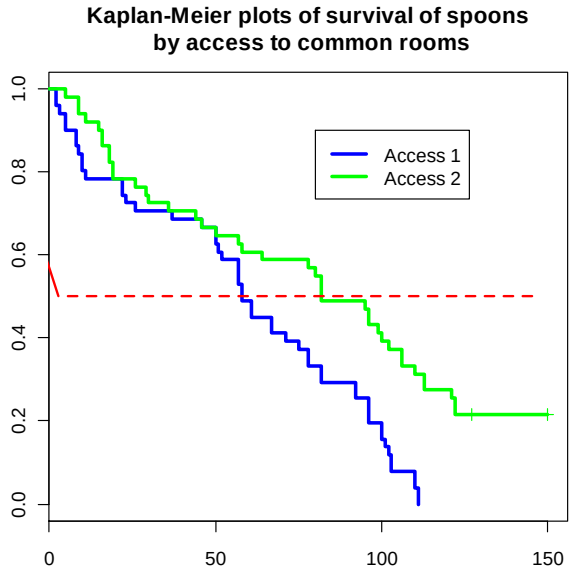
   1) Plot the Kaplan-Meier estimates of the survival times of teaspoons and estimate the median survival times in the two categories of rooms.

The script file for performing the calculations, **spoons.R**, is

```
library(survival)
ls()
attach(spoons)
spoons[1:5,]
spoons.sv<-Surv(Day,complete)
spoons.fit<-survfit(spoons.sv~Access)

plot(spoons.fit, lwd=3, col=c("blue", "green"),
main="Kaplan-Meier plots of survival of spoons\n by access to
common rooms" )
legtext=c("Access 1","Access 2")
legend(80,0.9, legtext,lwd=3,col=c("blue", "green"))
lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")

summary(spoons.fit)
```

A complete record of the **R** session running this is

```
>
> library(survival)
Loading required package: splines
> ls()
[1] "spoons"
> attach(spoons)
> spoons[1:5,]
  Day Access Value complete
1   2      1     1        1
2   2      1     1        1
3   3      1     1        1
4   5      1     2        1
5   5      1     1        1
> spoons.sv<-Surv(Day,complete)
> spoons.fit<-survfit(spoons.sv~Access)
> plot(spoons.fit)
> plot(spoons.fit, lwd=3, col=c("blue", "green"),
+ main="Kaplan-Meier plots of survival of spoons\n by access
to common rooms" )
> legtext=c("Access 1","Access 2")
> legend(80,0.9, legtext,lwd=3,col=c("blue", "green"))
> lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")
```



Note the commands used to produce the title and legend and most importantly the dashed line drawn at an estimated survival probability of 0.5. This allows a preliminary initial guess at the medians for the two groups as about 55 and 80 respectively.

```
> summary(spoons.fit)
Call: survfit(formula = spoons.sv ~ Access)

                Access=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     51       2   0.9608  0.0272       0.9090        1.000
    3     49       1   0.9412  0.0329       0.8788        1.000
    5     48       2   0.9020  0.0416       0.8239        0.987
    8     46       2   0.8627  0.0482       0.7733        0.963
    9     44       1   0.8431  0.0509       0.7490        0.949
   10     43       2   0.8039  0.0556       0.7020        0.921
   11     41       1   0.7843  0.0576       0.6792        0.906
   22     40       2   0.7451  0.0610       0.6346        0.875
   23     38       1   0.7255  0.0625       0.6128        0.859
   26     37       1   0.7059  0.0638       0.5913        0.843
   37     36       1   0.6863  0.0650       0.5700        0.826
   46     35       1   0.6667  0.0660       0.5491        0.809
   50     34       2   0.6275  0.0677       0.5079        0.775
   51     32       1   0.6078  0.0684       0.4876        0.758
   52     31       1   0.5882  0.0689       0.4675        0.740
   57     30       3   0.5294  0.0699       0.4087        0.686
   58     27       2   0.4902  0.0700       0.3705        0.649
   61     25       2   0.4510  0.0697       0.3332        0.610
   67     23       2   0.4118  0.0689       0.2966        0.572
   71     21       1   0.3922  0.0684       0.2787        0.552
   75     20       1   0.3725  0.0677       0.2609        0.532
   78     19       2   0.3333  0.0660       0.2261        0.491
   82     17       2   0.2941  0.0638       0.1923        0.450
   92     15       2   0.2549  0.0610       0.1594        0.408
   96     13       3   0.1961  0.0556       0.1125        0.342
  100     10       2   0.1569  0.0509       0.0830        0.296
  101      8       1   0.1373  0.0482       0.0690        0.273
  102      7       1   0.1176  0.0451       0.0555        0.249
  103      6       2   0.0784  0.0376       0.0306        0.201
  110      4       2   0.0392  0.0272       0.0101        0.153
  111      2       2   0.0000     NaN           NA           NA

                Access=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     51       1    0.980  0.0194        0.943        1.000
    9     50       2    0.941  0.0329        0.879        1.000
   11     48       1    0.922  0.0376        0.851        0.998
   15     47       1    0.902  0.0416        0.824        0.987
   16     46       2    0.863  0.0482        0.773        0.963
   18     44       2    0.824  0.0534        0.725        0.935
   19     42       2    0.784  0.0576        0.679        0.906
   26     40       1    0.765  0.0594        0.657        0.890
   29     39       1    0.745  0.0610        0.635        0.875
   30     38       1    0.725  0.0625        0.613        0.859
   36     37       1    0.706  0.0638        0.591        0.843
   44     36       1    0.686  0.0650        0.570        0.826
   46     35       1    0.667  0.0660        0.549        0.809
   50     34       1    0.647  0.0669        0.528        0.792
   57     33       1    0.627  0.0677        0.508        0.775
   58     32       1    0.608  0.0684        0.488        0.758
   64     31       1    0.588  0.0689        0.468        0.740
```

```
 78     30       1    0.569  0.0694        0.448        0.722
 80     29       1    0.549  0.0697        0.428        0.704
 82     28       3    0.490  0.0700        0.371        0.649
 95     25       1    0.471  0.0699        0.352        0.630
 96     24       2    0.431  0.0694        0.315        0.591
 99     22       1    0.412  0.0689        0.297        0.572
100     21       1    0.392  0.0684        0.279        0.552
102     20       1    0.373  0.0677        0.261        0.532
106     19       2    0.333  0.0660        0.226        0.491
110     17       1    0.314  0.0650        0.209        0.471
113     16       2    0.275  0.0625        0.176        0.429
121     14       1    0.255  0.0610        0.159        0.408
122     13       2    0.216  0.0576        0.128        0.364
```

>

To estimate the medians, look at the K-M estimates above: For Access group 1 we see the median must be between 57 and 58 when the survival probabilities are 0.53 and 0.49. A common sense estimate is that the median is late in the afternoon of the 58th day. An over-precise estimate is obtained by

57 + (0.5294 – 0.5)★(58 – 57)/(0.5284 – 0.4902) = 57.77 days

(using **R**).

For Access group 2 it is about 81 or

80 + (0.549 – 0.5)★(82 – 80)/(0.549 – 0.490) = 81.66 days

Notes & Solutions for Tasks 3

For the data on the data leukaemia remission times

(given on Task Sheet 2 & on the appropriate Web pages)

1) Calculate the log rank statistic for testing for a difference in survival times between the two groups and assess its significance.

2) Assuming that survival times are exponentially distributed, Ex(λ_1) and Ex(λ_2) respectively, estimate λ_1 and λ_2.

3) Assuming that the survival times are exponentially distributed use the estimates from part (ii) to estimate the median survival times of the two groups, providing 95% confidence intervals for each group.

4) Calculate MLE and Likelihood Ratio Test statistics for testing for a difference in survival times between the two groups and assess their significance.

5) Plot the logs of the exponential survivor functions and the Kaplan-Meier survivor functions on the same graph. Comment on the fit of the exponential model to these data.

6) Comment on the effect of the drug.

1) Log rank test:

		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	1	21	21	42	0	2	2	1	1
2	2	21	19	40	0	2	2	1.05	0.95
3	3	21	17	38	0	1	1	0.553	0.447
17	23	6	1	7	1	1	2	1.714	0.286
					O_1=9	O_2=21		E_1=19.25	E_2=10.75

Log rank statistic is 15.28 on 1 d.f. and so the data give very good evidence that the survival patterns of treated and control groups

are different with the treated group having better longer remission times.

N.B. If you do this in R or Minitab then the value of the log rank statistic will be slightly different (16.79) since these packages make a variance adjustment for tied event times:

```
> library(survival)
Loading required package: splines
> attach(leukrem)
> leuk.sv<-Surv(time,censor)
> survdiff(leuk.sv~group)
Call:
survdiff(formula = leuk.sv ~ group)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 21        9     19.3      5.46      16.8
group=2 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
>
```

2)

	$\Sigma\delta_i$	Σt_i	$\hat{\lambda}$	s.e.($\hat{\lambda}$)	95%CI for λ
Drug	9	359	0.0251	0.0084	(0.0091, 0.0421)
Control	21	182	0.1154	0.0252	(0.0660, 0.1648)

3) Estimate of median is $-\hat{\lambda}^{-1}\log(0.5)$ which has (using formula on P37) standard error $-\log(0.5)\hat{\lambda}^{-1}/(\Sigma\delta_i)^{½}$. 95% confidence intervals obtained from estimate +/– 2×st.err, (or use 1.96 not 2 for spurious exactness). This gives 27.62 with 95% interval (27.62 +/– 18.41) = (9.21, 46.03) for group 1 and 6.01, (6.01 +/– 2.62)= (3.39, 8.63) for the control group.

4) MLE Test statistic is –3.404 (observation of N(0,1) under H_0) and LRT statistic is 16.49 (observation of χ_1^2 under H_0) which yields the same conclusion as log rank test.

5) Plots of log[exp{$-\hat{\lambda}$it} and the log{K-M estimates} on same graph (not shewn here) look close for each group, so the plots suggest that the exponential model is suitable. There looks to be a difference between the control and treated groups.

6) There is strong evidence that the drug prolongs survival.

Notes & Solutions for Tasks 4

1) The table below gives details of a proportional hazards model fitted to some data obtained from patients being treated for kidney failure where 'survival time' is in terms of time to relapse.

Variable	***Coefficient***	***Standard Error***	***χ^2 statistic (using L.R.T)***	p-value
Treatment *0 =Treatment A* *1 =Treatment B*	*-1.63*	*0.75*	*4.71*	< 0.05
Age (years)	*-0.003*	*0.024*	*0.01*	>>0.10
Sex *0 = female* *1 = male*	*0.67*	*0.32*	*3.91*	< 0.05
Obesity *0 = no* *1 = yes*	*0.0092*	*0.0045*	*4.44*	< 0.05
Duration of symptoms prior to treatment (months)	*-0.003*	*0.075*	*0.01*	>>0.10

Describe the effects of treatment and additional covariates on time to relapse.

1) It is clear that there is little evidence that either the subject's age or the duration of symptoms affect the relapse time. There is good evidence that (a) Treatment B gives a longer time to relapse, (b) females have a longer relapse time than males and (c) obese subjects have shorter relapse times than non-obese ones.

The ratio of hazards for subjects on treatment B to treatment A (with otherwise common values of covariates) is $e^{-1.63} = 0.196$ with 95% CI (0.044, 0.878). Similarly for males relative to females the corresponding figures are 1.954 with 95%CI (1.03, 3.706) and for obese to non-obese 1.009 and 95%CI (1.0002, 1.0184) (i.e.

actually very little effect). For an increase of one year in age, the 95%CI for the proportional change in hazard is (0.950, 1.04) and for an increase of one month in duration of symptoms it is (0.985, 1.012).

In summary, the most important effects on relapse time are the treatment, treatment B reducing the hazard of relapse to about a fifth of that on treatment A, and the sex of the subject, with males having a hazard of about twice that of comparable females.

Notes & Solutions for Tasks 5

*1) The data given below represent survival times for lymphoma patients according to the stage of tumour (where * denotes a censored value):*

Stage 3	*6*	*20*	*42*	*43**	*169**	*207*	*253*	*255**		
Stage 4	*4*	*10*	*20*	*21**	*30*	*33**	*43**	*46*	*110*	*235**

i) Compute the Kaplan-Meier product limit estimates of the survivor functions for stage 3 and stage 4 separately.

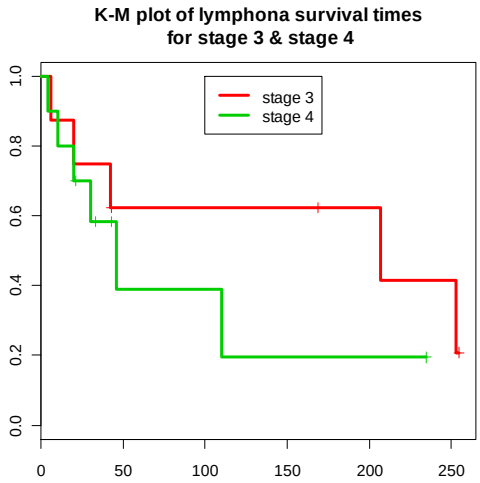
ii) Provide estimates of the two cumulative hazard functions and comment on any differences.

By using the log-rank test, compare the survival distributions for the two stages.

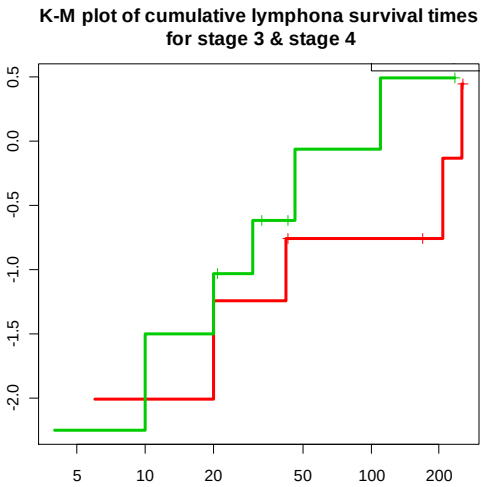
First the S-Plus version, first using the menu in `Statistics>Survival>Nonparametric Survival…)`. Note the production of the Survival plot on a logarithmic vertical scale by clicking the appropriate box on the dialogue box of the plot menu.

```
> library(survival)
Loading required package: splines
> attach(lymphoma)
> lymph.sv<-Surv(time,censor)
> lymphsurv<-survfit(lymph.sv~stage)
>
> summary(lymphsurv)
Call: survfit(formula = lymph.sv ~ stage)

                stage=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6      8       1    0.875   0.117       0.6734            1
   20      7       1    0.750   0.153       0.5027            1
   42      6       1    0.625   0.171       0.3654            1
  207      3       1    0.417   0.205       0.1590            1
  253      2       1    0.208   0.179       0.0385            1
```

```
                stage=4
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    4     10       1    0.900  0.0949       0.7320            1
   10      9       1    0.800  0.1265       0.5868            1
   20      8       1    0.700  0.1449       0.4665            1
   30      6       1    0.583  0.1610       0.3396            1
   46      3       1    0.389  0.1916       0.1480            1
  110      2       1    0.194  0.1676       0.0359            1
> plot(lymphsurv,lwd=3,col=c(2,3),
+ main="K-M plot of lymphona survival times\n for stage 3 & stage 4")
> legend(100,1,c("stage 3", "stage 4"), lwd=3, col=c(2,3))
>
```




```
> plot(lymphsurv,lwd=3,col=c(2,3),fun="cloglog",
+ main="K-M plot of cumulative lymphona survival times\n for stage 3
& stage 4")
> legend(100,1,c("stage 3", "stage 4"), lwd=3, col=c(2,3))
>
```




```
>
```

For the log rank test has to be obtained from the command line:

```
> survdiff(lymph.sv~stage)
Call:
survdiff(formula = lymph.sv ~ stage)

         N Observed Expected (O-E)^2/E (O-E)^2/V
stage=3  8        5     6.37     0.296     0.804
stage=4 10        6     4.63     0.408     0.804

 Chisq= 0.8  on 1 degrees of freedom, p= 0.37
>
```

2) *In an accelerated-life survival model the survivor function for an individual with covariate x satisfies*

$$S(t:x) = S_0(te^{\beta x}),$$

where $S_0(t)$ is some baseline survivor function.

i) Show that the corresponding hazard function satisfies

$$h(t:x) = e^{\beta x}h_0(te^{\beta x})$$

where $h_0(t)$ is the baseline hazard function for $S_0(t)$.

$h(t;x) = -S'(t;x)/S(t;x) = -e^{\beta x}S_0'(te^{\beta x})/ S_0(te^{\beta x}) = e^{\beta x}\{S_0'(t;x)/ S_0(t;x) = e^{\beta x}h_0(te^{\beta x})$.

ii) In a trial where n independent patients, with covariate values $x_1,x_2,...,x_n$ enter at the same time, suppose that all death times are observed and that $S_0(t) = e^{-\lambda t}$ (t>0). Show that the survival time T has an exponential form, and is of proportional hazard form.

$S_0(t) = e^{-\lambda t}$ so $h_0(t) = -(-\lambda e^{-\lambda t})/e^{-\lambda t} = \lambda$. So T_i has survivor function $\exp\{-\lambda t \exp(\beta x_i)\}$ and hazard function $\lambda\exp(\beta x_i)$ so is of proportional hazards form

iii) ***** *Show that the distribution of the time to the first death in the trial is exponential with mean*

$$[\lambda\Sigma\exp(\beta x_i)]^{-1}$$

Let $M=\min\{T_1,T_2,\ldots,T_n\}$ then $S_M(t) = P[T_1>t, T_2>t,\ldots,T_n>t]$

$$= \prod_{i=1}^{n} P[T_i > t] = \prod_{i=1}^{n} S(t;x_i) = \prod_{i=1}^{n} \exp\{-\lambda t e^{\beta x_i}\} = \exp\{-\lambda t \Sigma e^{\beta x_i}\}$$

and so M is exponential with mean $[\lambda\Sigma e^{\beta x_i}]^{-1}$.

iv) ***** *Show that the probability that the j^{th} patient is the first to die is given by $\exp(\beta x_j)/\Sigma\exp(\beta x_i)$*

j^{th} is the first if $T_1>T_j$, $T_2>T_j$,… ,$_{\text{not j,}}$…,$T_n > T_j$ so this has probability $P[T_1>T_j, T_2>T_j,\ldots,_{\text{not j,}},\ldots,T_n > T_j]$ `

$$= \int_{t_j=0}^{\infty} P[T_1 > T_j,\ldots,_{\text{not j}},\ldots T_n > T_j \mid T_j = t_j] f(t_j;x_j)dt_j$$

$$= \int_{t_j=0}^{\infty} P[T_1 > T_j,\ldots,_{\text{not j}},\ldots T_n > T_j] f(t_j;x_j)dt_j$$

$$= \int_{t_j=0}^{\infty} \prod_{\substack{i=1 \\ i\neq j}}^{n} \exp\{-\lambda t_j e^{\beta x_i}\} \lambda e^{\beta x_j} \exp\{-\lambda t_j e^{\beta x_j}\} dt_j = \int_{t_j=0}^{\infty} \lambda e^{\beta x_j} \exp\{-\lambda t_j \sum_{i=1}^{n} e^{\beta x_i}\} dt_j$$

$$= \frac{e^{\beta x_j}}{\sum_{i=1}^{n} e^{\beta x_i}}$$

Notes & Solutions for Exercises 1

1) Returning to the Australian study on survival of spoons,

i) Is there evidence that the disappearance of spoons is dependent upon either the category of tearoom or the value of the spoon?

ii) What is the average rate of loss of teaspoons?

iii) If the Institute where the study was conducted has 150 employees, how many teaspoons should be purchased annually to provide one spoon for every two people?

(N.B. You should appreciate that the data given here are those observed at the Australian institution so you are advised to evaluate your answer to this question using common sense: the answer should be within the petty cash budget of the tea-room).

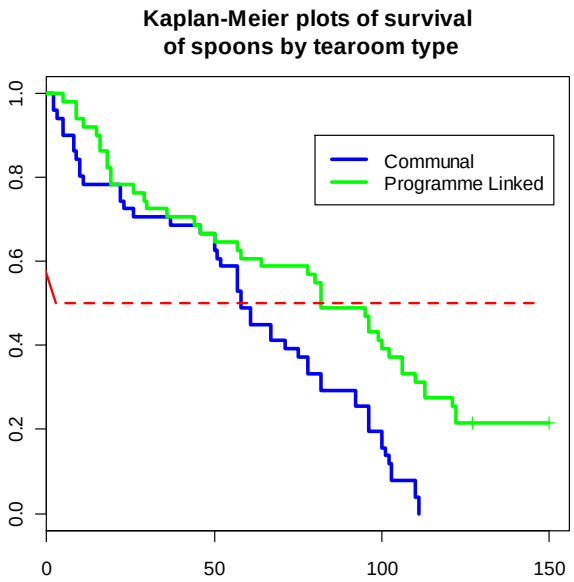
Source: ***The case of the disappearing teaspoons: longitudinal cohort study of the displacement of teaspoons in an Australian research institute,*** by Megan S C Lim, Margaret E Hellard, Campbell K Aitken. (2005). http://www.bmj.com/content/331/7531/1498.pdf%2Bhtml BMJ VOLUME 331 24-31 DECEMBER 2005, p1498-1500

Script file: #Q4

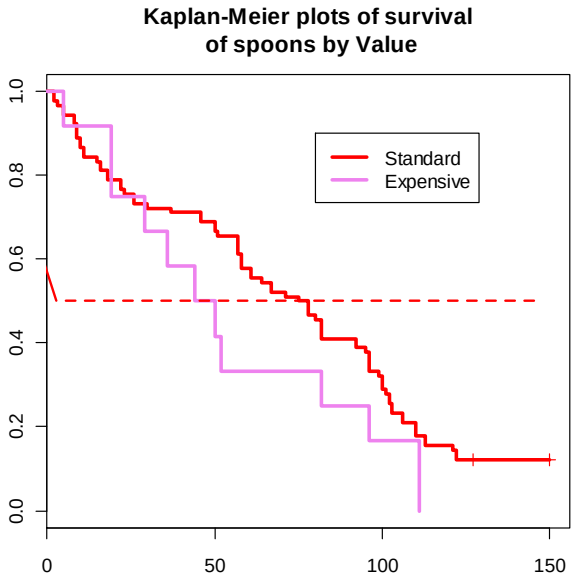
```
detach(cirrhosis)
library(survival)
load("spoons.Rdata")
attach(spoons)
spoons[1:5,]
spoons.sv<-Surv(Day,complete)
spoonsAcc.fit<-survfit(spoons.sv~Access)
plot(spoonsAcc.fit, lwd=3, col=c("blue", "green"),
main="Kaplan-Meier plots of survival\n of spoons by
tearoom type" )
legtext=c("Communal","Programme Linked")
legend(80,0.9, legtext,lwd=3,col=c("blue", "green"))
lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")

spoonsVal.fit<-survfit(spoons.sv~Value)

plot(spoonsVal.fit, lwd=3, col=c("red", "violet"),
main="Kaplan-Meier plots of survival\n of spoons by
Value" )
legtext=c("Standard","Expensive")
legend(80,0.9, legtext,lwd=3,col=c("red", "violet"))
lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")


survdiff(spoons.sv~Access)
survdiff(spoons.sv~Value)
missing<-sum(complete)
missprog<-sum(complete*(Access-1))
misscomm<-missing-missprog
missprog;misscomm
max(Day[Access*complete==1])
max(Day[(Access-1)*complete==1])
```

A record of the session follows:

```
> library(survival)
Loading required package: splines
> load("spoons.Rdata")
> attach(spoons)
> spoons[1:5,]
  Day Access Value complete
1   2      1     1        1
2   2      1     1        1
3   3      1     1        1
4   5      1     2        1
5   5      1     1        1
> spoons.sv<-Surv(Day,complete)
> spoonsAcc.fit<-survfit(spoons.sv~Access)
> plot(spoonsAcc.fit, lwd=3, col=c("blue", "green"),
+ main="Kaplan-Meier plots of survival\n of spoons by
tearoom type" )
> legtext=c("Communal","Programme Linked")
> legend(80,0.9, legtext,lwd=3,col=c("blue", "green"))
> lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")
```




```
> plot(spoonsVal.fit, lwd=3, col=c("red", "violet"),
+ main="Kaplan-Meier plots of survival\n of spoons by
Value" )
> legtext=c("Standard","Expensive")
> legend(80,0.9, legtext,lwd=3,col=c("red", "violet"))
> lines(c(0,150),c(0.5,0.5),col="red",lwd=2,lty="dashed")
>
```



We can see from the K-M plots that the loss of spoons from the Communal Rooms is greater than that from the Programme Linked and that more expensive spoons tend to disappear at a faster rate. To assess how strong the evidence it we do

```
> survdiff(spoons.sv~Access)
Call:
survdiff(formula = spoons.sv ~ Access)

          N Observed Expected (O-E)^2/E (O-E)^2/V
Access=1 51       51     35.1      7.18      13.1
Access=2 51       40     55.9      4.51      13.1

 Chisq= 13.1  on 1 degrees of freedom, p= 3e-04
>
> survdiff(spoons.sv~Value)
Call:
survdiff(formula = spoons.sv ~ Value)
         N Observed Expected (O-E)^2/E (O-E)^2/V
Value=1 90       79    82.89     0.182      2.14
Value=2 12       12     8.11     1.864      2.14
 Chisq= 2.1  on 1 degrees of freedom, p= 0.144
>
```

indicating that there is very strong evidence ($p < .001$) that the loss from the communal tearooms is at a faster rate than in Programme Linked ones, but little evidence that the rate depends upon value of spoon.

For part (iii) we need to appreciate that spoons go missing because people take them and not because they wear out or evaporate away to nothing or surreptitiously migrating to a spoonoid planet (see Lim *et al,* 2005 & Adams, 1979). Next, note that all of the marked spoons had disappeared from the communal rooms by day 122 so presumably more might have disappeared from these rooms had there been more marked ones so the estimate of the rate of spoon loss should be based just on the state after 122 days. This also assumes that the proportion of marked spoons is constant over the period, rather than at each day each surviving spoon has an equal chance of being removed, i.e. the proportion of marked spoons is small. The rate at which spoons go missing from the institution is roughly 91/122 per day which suggests that in a year the loss will be about 273, so need an initial stock of about 350, topped up to this level each year (about 270 approximately) thereafter.

Notes & Solutions for Exercises 2

1) The table below gives some details of fitting a proportional hazards regression model to times to recurrence of a certain disease. The data were obtained during a randomised clinical trial of a new treatment. The factors investigated were treatment (coded by $x_1 = 0$ for placebo, $x_1 = 1$ for treatment), stage of disease (coded by $x_2 = 0$ for stage I, $x_2 = 1$ for stage II, $x_2 = 2$ for stage III) and the interaction between treatment and stage of disease (coded by x_3 where $x_3 = x_1 \times x_2$

	variable	*coefficient*	*standard error*
Treatment	x_1	*–0.18*	*0.10*
Stage	x_2	*+0.32*	*0.21*
Interaction	x_3	*–0.66*	*0.11*

i) Specify the form of the proportional hazards model used for this analysis in terms of the baseline hazard function $h_0(t)$ and the covariates.

The form of the model is $h(t)=h_0(t)\exp\{\beta_1x_1+\beta_2x_2+\beta_3x_3\}$ where $h_0(t)$ is the baseline hazard function and x_i (i=1,2,3) are as defined above. For subjects on placebo this becomes

$h(t)=h_0(t)$ for stage I

$h(t)=h_0(t)\exp\{\beta_2\}$ for stage II

$h(t)=h_0(t)\exp\{2\beta_2\}$ for stage III

and for those receiving treatment it is

$h(t)=h_0(t)\exp\{\beta_1\}$ for stage I

$h(t)=h_0(t)\exp\{\beta_1+\beta_2+\beta_3\}$ for stage II

$h(t)=h_0(t)\exp\{\beta_1+2\beta_2+2\beta_3\}$ for stage III

ii) Describe in detail the effects of these factors on the time to recurrence of the disease.

estimated values of $h(t)/h_0(t)$ (i.e. hazard ratio relative to those on placebo at stage I), with approximate 95%CIs, are

placebo, stage II: 1.38, (0.90, 2.10)

placebo:stage III: 1.90, (0.82, 4.39)

treatment, stage I: 0.84, (0.68, 1.22)

treatment, stage II: 0.59, (0.36, 0.99)

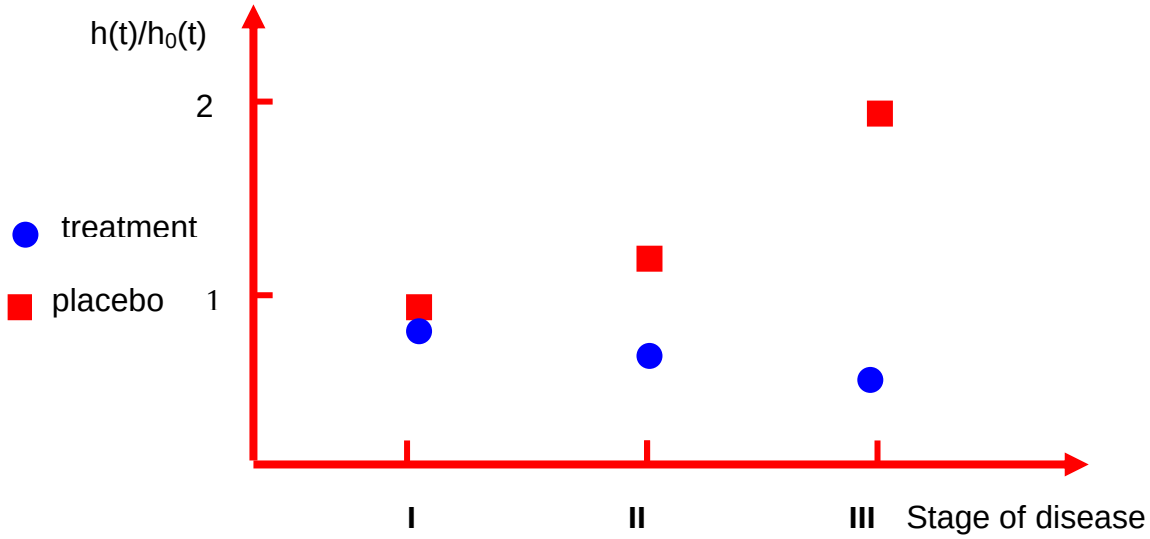
treatment, stage III: 0.42, (0.16, 1.11)

(assuming estimates of the β_i are independent). Note, CIs calculated as estimate+/– $2\times$s.e. and e.g. s.e.$(\beta_1+\beta_2)=(0.10^2+0.21^2)^{½}$ etc. and to get CI of e.g. exp{β_1} take exp{CI for β_1}.

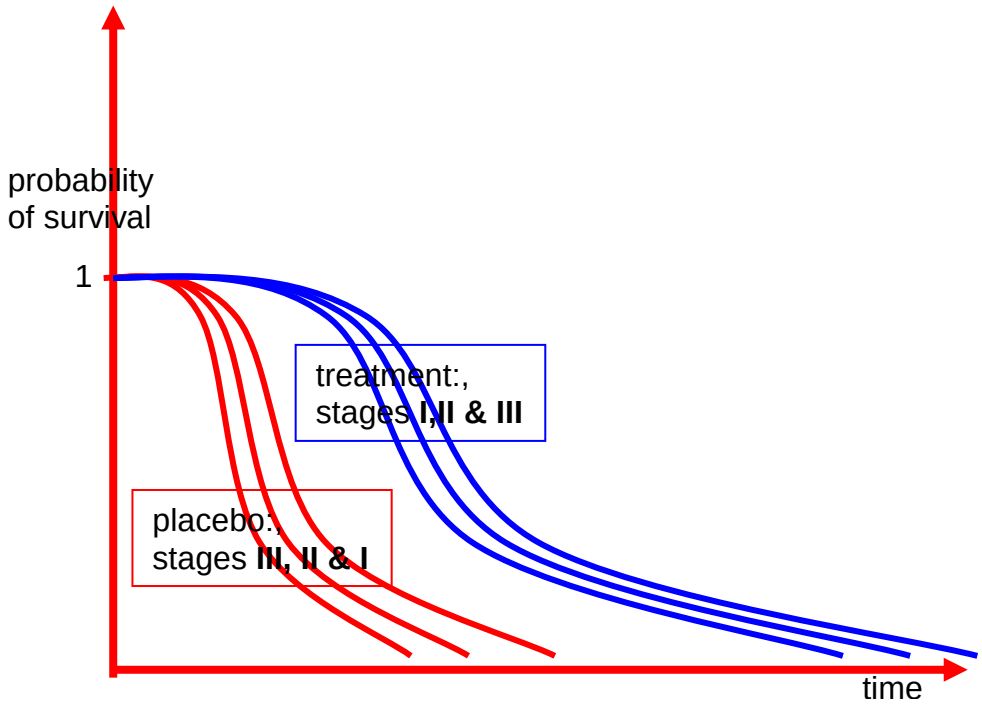
Thus, on untreated patients the hazards increase with stage of disease and so their survival prospects decrease with stage of disease. For patients on the treatment the effect of stage of disease is negated and indeed perhaps slightly reversed, though looking att he CIs the evidence for an actual improvement in survival with stage of disease is weak.

iii) Show diagrammatically the form of the relationship between the survivor functions and the stage of the disease for the two different treatment groups.

First, a diagram of the hazard ratios:



This allows a sketch of the relative positions of the survival times



2) *The data in the file AHprostate.Rdata have been adapted from Andrews and Herzberg (1985) and give results of a trial on treatments for prostate cancer. Various covariates were recorded. The variables in the data file are given in the table below.*

Variable	Description	Levels
Stage	Stage of Disease	3=No evidence of distant metastasis, 4=evidence of distant metastasis
RX	Treatment Groups	1=Control Arm, 2=Experimental Arm
Dtime	Complete months to follow up	
Status	Survival Status	
AgeYrs	Age of Patient	89 denotes 89 or more
Wt	Weight Index	Weight (Kg) – Height (cm) + 200
PF	Performance Rating	
HX	History of cardiovascular disease	0=no, 1=yes
SBP	Systolic blood pressure	
DBP	Diastolic blood pressure	
EKG	Electrocardiogram	
HG	Serum Haemoglobin,g/100ml	
SZ	Size of primary tumour	cm2
SG	Combined index of tumour	stage and histologic grade
AP	Serum prostatic acid phosphatase	
BM	Bone metastases	0=no, 1=yes

Investigate whether there is evidence that the treatment improves survival time to follow up, making due allowance for any other prognostic factors on the subjects.

Below is an extensively annotated **R** session, noting particularly some ways of using **R** to achieve quick recoding of variables and extraction of the names of levels of a factor to use in the legend of a plot.

```
>
> ls()
[1] "AHprostate"
> names(AHprostate)
 [1] "stage"    "rx"         "dtime"    "status"  "age"       "wt"
"pf"      "hx"
 [9] "sbp"      "dbp"      "ekg"      "hg"        "sz"        "sg"
"ap"      "bm"
```

```
> AHprostate[1:10,]
   stage             rx dtime                   status age  wt
1      3 0.2 mg estrogen    72                    alive  75  76
2      3 0.2 mg estrogen     1          dead - other ca  54 116
3      3 5.0 mg estrogen    40   dead - cerebrovascular  69 102
4      3 0.2 mg estrogen    20   dead - cerebrovascular  75  94
5      3         placebo    65                    alive  67  99
6      3 0.2 mg estrogen    24      dead - prostatic ca  71  98
7      3         placebo    46 dead - heart or vascular  75 100
8      3         placebo    62                    alive  73 114
9      3 1.0 mg estrogen    61                    alive  60 110
10     3 1.0 mg estrogen    60                    alive  78 107
                    pf hx sbp dbp                          ekg       hg sz sg
1       normal activity  0  15   9                 heart strain 13.79883  2  8
2       normal activity  0  13   7 heart block or conduction def 14.59961 42 NA
3       normal activity  1  14   8                 heart strain 13.39844  3  9
4  in bed < 50% daytime  1  14   7                       benign 17.59766  4  8
5       normal activity  0  17  10                       normal 13.39844 34  8
6       normal activity  0  19  10                       normal 15.09961 10 11
7       normal activity  0  14  10                       benign 13.00000 13  9
8       normal activity  1  17  11                 heart strain 12.59961  3  9
9       normal activity  0  12   8                       normal 14.59961  4 10
10      normal activity  1  13   8                       old MI 13.00000 21  6
          ap bm
1  0.2999878  0
2  0.6999512  0
3  0.2999878  0
4  0.8999023  0
5  0.5000000  0
6  0.5999756  0
7  0.7999268  0
8  0.5999756  0
9  0.6999512  0
10 0.3999634  0
```

Note that the censoring variable `status` has one category for 'alive' and several categories for 'dead'. For a simple survival analysis using the function `Surv(.)` these need to be coded as 0 or 1 (or, a little unusually, as `FALSE` and `TRUE`). The function `unclass(.)` will assign numeric values to the factor levels starting at 1 taking the levels in alphabetic order. Then the statement `unclass(status) > 1` is a logical assertion which is either true (if status is one of the dead categories for that case) or false (if status for that case is alive).

It is also worth noting that there appear to be several different treatment levels (variable `rx`) which was not obvious from the description in the question. To check just how many levels and what they are do:

```
> summary(status)
                        alive       dead - cerebrovascular
                          141                           30
    dead - heart or vascular              dead - other ca
                           91                           23
dead - other specific non-ca          dead - prostatic ca
                           27                          127
    dead - pulmonary embolus   dead - respiratory disease
                           13                           16
        dead - unknown cause    dead - unspecified non-ca
                            5                            7
> summary(rx)
0.2 mg estrogen   1.0 mg estrogen   5.0 mg estrogen placebo
            118                 118                 120
124
>
```

This shews that there is indeed only one 'alive' category and the others all begin with the letter `d` later in the alphabet than `a`. There are four treatment groups, three in increasing doses of estrogen and one placebo. Note that **R** will take these in alpha-numeric order so that groups 1-3 will be the three treated groups (lowest first) and the fourth is the placebo group. The fact that there are nearly equal numbers in the four groups suggest strongly that the allocation to treatment groups was completely at random and so one would expect the covariates to be balanced between treatment groups, i.e. it is likely that an analysis allowing for the effects of the covariates will produce much the same results from a simple analysis ignoring them.

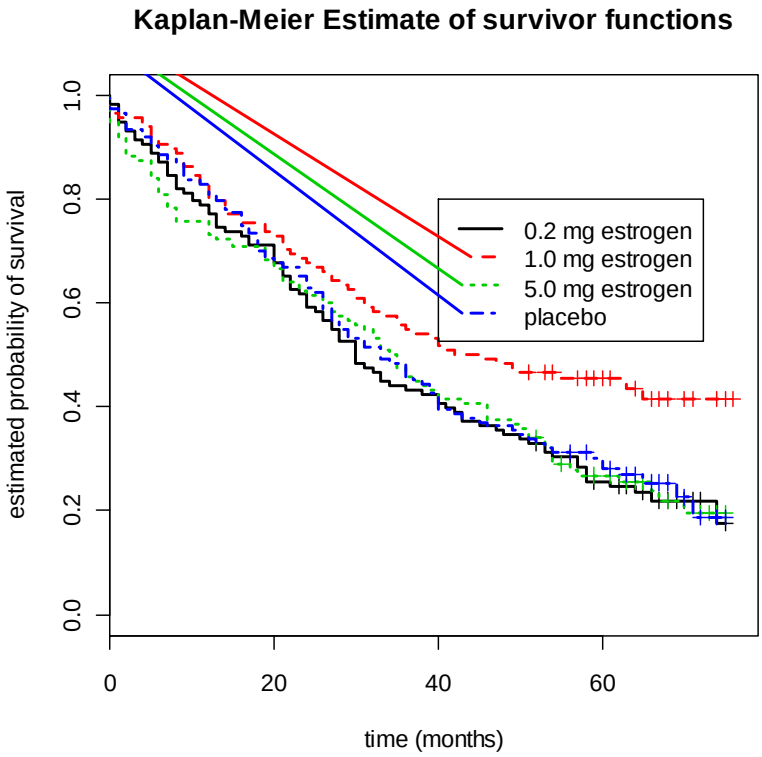
Now begin the main analysis, calling the survival library to start.

```
> library(survival)
Loading required package: splines
> attach(AHprostate)
> dead<-unclass(status) > 1
> #
> # note way of recoding status to logical variable
> # indicating dead (TRUE) or alive (FALSE) : Check it's ok
> #
> dead[1:10]
 [1] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
FALSE
> status[1:10]
```

```
 [1] alive                    dead - other ca          dead -
cerebrovascular
 [4] dead - cerebrovascular   alive                    dead -
prostatic ca
 [7] dead - heart or vascular alive                   alive
[10] alive
10 Levels: alive dead - cerebrovascular ... dead - unspecified
non-ca
>
> # Also note that a Logical variable is ok as a status
indicator
> # instead of the more usual 0 (alive) 1 (dead) binary
indicator
> #
> #
```

First, an exploratory analysis to look at the differences between treatments without any allowance for covariates.

```
> d.surv<-Surv(dtime, dead)
> dfit.raw<-survfit(d.surv~rx)
> plot(dfit.raw, col=c(1:4), lty=c(1:4),lwd=2,
+      main="Kaplan-Meier      Estimate      of      survivor
functions",xlab="time (months)",
+ ylab="estimated probability of survival")
> legend(40,0.8,levels(rx),col=c(1:4),lty=c(1:4),lwd=2)
> #
> # note use of levels() to get the required text for the
legend
> #
> # NOW CLICK Recording in drop down menu from History
> # to allow scrolling between graphs
> #
```



This shews (slightly surprisingly) that if there are any differences between the treatment groups then they rest almost entirely in the 1.0 estrogen group having better survival prospects (note the K-M curve is above those of the other groups) and that there is little difference between the placebo and low and high doses of estrogen. A quick check with a log-rank test gives:

```
> survdiff(d.surv~rx)
Call:
survdiff(formula = d.surv ~ rx)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
rx=0.2 mg estrogen 118       91     80.4     1.392      1.86
rx=1.0 mg estrogen 118       66     90.7     6.729      9.38
rx=5.0 mg estrogen 120       90     81.3     0.926      1.24
rx=placebo         124       92     86.6     0.343      0.47

 Chisq= 9.6  on 3 degrees of freedom, p= 0.0225
```

shewing there to be an overall difference between the groups with the largest contribution to the chi-squared statistics coming from the 1.0 estrogen group.

Now move to the analysis bringing in the covariates. The way to handle this is to use a regression model, first fit all the covariates and then fit the additional factor of treatment. Equivalently, looking at a model including all the covariates and the treatment factor and then comparing parameter estimates with standard errors (for a continuous or binary factor) or a chi-squared value (for multi-level factors) gives a test of whether the covariate concerned plays a role in survival prospects ***given all the other terms are already included*.**

```
)
> d.ph1<-
coxph(d.surv~stage+rx+age+age+wt+pf+hx+sbp+dbp+ekg+hg+sz+sg+ap
+bm)
> summary(d.ph1)
Call:
coxph(formula = d.surv ~ stage + rx + age + age + wt + pf + hx
+
    sbp + dbp + ekg + hg + sz + sg + ap + bm)

  n= 462, number of events= 329
   (18 observations deleted due to missingness)

                                      coef exp(coef) se(coef)     z Pr(>|z|)
stage                              -0.11983   0.88707  0.17767 -0.67    0.500
rx1.0 mg estrogen                  -0.32483   0.72265  0.17353 -1.87    0.061
rx5.0 mg estrogen                   0.02239   1.02264  0.15824  0.14    0.887
rxplacebo                           0.00594   1.00596  0.15620  0.04    0.970
age                                 0.01905   1.01923  0.00901  2.11    0.034
wt                                 -0.01125   0.98881  0.00477 -2.36    0.018
pfin bed < 50% daytime             -1.32166   0.26669  0.84580 -1.56    0.118
pfin bed > 50% daytime             -1.14808   0.31724  0.84264 -1.36    0.173
pfnormal activity                  -1.62114   0.19767  0.82029 -1.98    0.048
hx                                  0.51401   1.67199  0.12224  4.21  2.6e-05
sbp                                -0.02806   0.97233  0.02972 -0.94    0.345
dbp                                 0.05180   1.05317  0.04875  1.06    0.288
ekgbenign                          -0.42886   0.65125  0.49375 -0.87    0.385
ekgheart block or conduction def   -0.56507   0.56832  0.50269 -1.12    0.261
ekgheart strain                    -0.06081   0.94100  0.43804 -0.14    0.890
ekgnormal                          -0.54088   0.58224  0.43899 -1.23    0.218
ekgold MI                          -0.48629   0.61490  0.45573 -1.07    0.286
ekgrecent MI                        0.28346   1.32772  1.11232  0.25    0.799
ekgrhythmic disturb & electrolyte ch -0.16633   0.84676  0.45999 -0.36    0.718
hg                                 -0.07286   0.92973  0.03320 -2.19    0.028
sz                                  0.01868   1.01886  0.00461  4.05  5.1e-05
sg                                  0.08470   1.08839  0.04223  2.01    0.045
ap                                 -0.00152   0.99848  0.00101 -1.51    0.132
bm                                  0.25079   1.28504  0.18658  1.34    0.179

stage
```

```
rx1.0 mg estrogen                    .
rx5.0 mg estrogen
rxplacebo
age                                  *
wt                                   *
pfin bed < 50% daytime
pfin bed > 50% daytime
pfnormal activity                    *
hx                                   ***
sbp
dbp
ekgbenign
ekgheart block or conduction def
ekgheart strain
ekgnormal
ekgold MI
ekgrecent MI
ekgrhythmic disturb & electrolyte ch
hg                                   *
sz                                   ***
sg                                   *
ap
bm
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.654  (se = 0.017 )
Rsquare= 0.206   (max possible= 1 )
Likelihood ratio test= 107  on 24 df,   p=2e-12
Wald test            = 106  on 24 df,   p=2.53e-12
Score (logrank) test = 111  on 24 df,   p=4.54e-13
```

(Some output deleted)

Note this indicates marginal significance for the coefficient of the 1.0 estrogen group (taking 0.2 estrogen as the reference group). To calculate the chi-squared value we need

```
> 1.87**2+0.14**2+0.04**2
[1] 3.5181 "
> 1-pchisq(3.5181,3)
[1] 0.31842
```

indicating a p-value of 32%, i.e. little to no evidence at all. **NB:** this is actually an approximation to the likelihood ratio statistic and in fact with so many terms in the model it turns out to be a rather poor approximation (see below).

It is also possible to use the function `anova(.)` on the results of the Cox proportional hazards regression `d.ph1` but extreme care must be used (the first analysis below is **incorrect**):

```
anova(d.ph1)
Analysis of Deviance Table
 Cox model: response is d.surv
Terms added sequentially (first to last)
```

```
      loglik Chisq Df Pr(>|Chi|)
NULL   -1838
stage  -1834  8.77  1    0.00307 **
rx     -1829 10.23  3    0.01673 *
age    -1823 10.38  1    0.00127 **
wt     -1820  6.12  1    0.01334 *
pf     -1814 13.12  3    0.00439 **
hx     -1807 13.36  1    0.00026 ***
sbp    -1807  0.26  1    0.60874
dbp    -1807  0.03  1    0.86904
ekg    -1801 11.58  7    0.11532
hg     -1798  5.83  1    0.01573 *
sz     -1789 18.84  1    1.4e-05 ***
sg     -1787  4.34  1    0.03716 *
ap     -1786  2.20  1    0.13842
bm     -1785  1.79  1    0.18143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This appears to shew that the effect of the factor `rx` is significant with a p-value of 0.01673. **However**, note the warning given at the top that it is assumed that terms are added sequentially, so the p-values indicate the strength of the evidence for the effect of each term ***assuming all the preceding terms are already in the model***. To overcome this and answer the question asked, the factor `rx` has to be the last to be included:

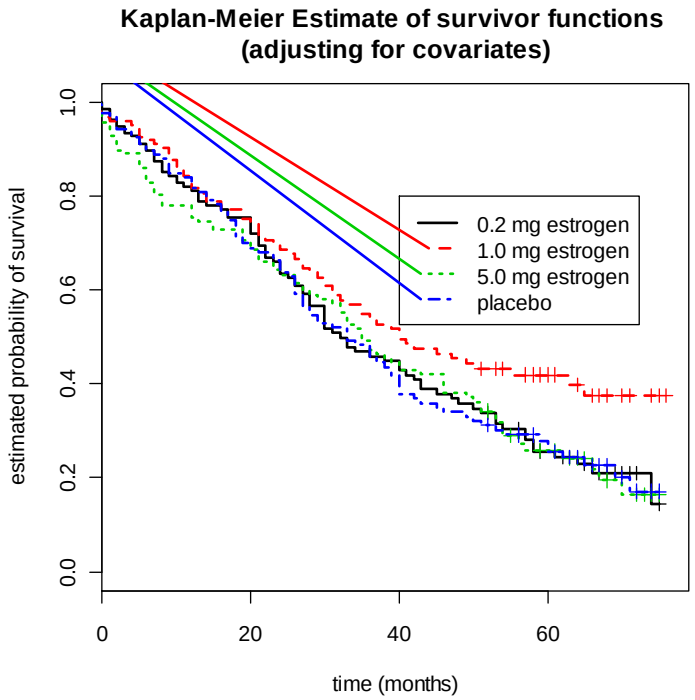
```
> d.ph<-
coxph(d.surv~stage+age+age+wt+pf+hx+sbp+dbp+ekg+hg+sz+sg+ap+bm
+rx)
```

```
> anova(d.ph)
Analysis of Deviance Table
 Cox model: response is d.surv
Terms added sequentially (first to last)
      loglik Chisq Df Pr(>|Chi|)
NULL   -1838
stage  -1834  8.77  1     0.0031 **
age    -1828 10.72  1     0.0011 **
wt     -1825  6.07  1     0.0137 *
pf     -1818 15.12  3     0.0017 **
hx     -1810 15.75  1    7.2e-05 ***
sbp    -1810  0.26  1     0.6118
dbp    -1810  0.02  1     0.8964
ekg    -1804 12.29  7     0.0913 .
hg     -1801  5.55  1     0.0185 *
sz     -1792 18.24  1    1.9e-05 ***
sg     -1790  4.23  1     0.0396 *
ap     -1788  3.07  1     0.0798 .
bm     -1787  1.15  1     0.2825
rx     -1785  5.61  3     0.1324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

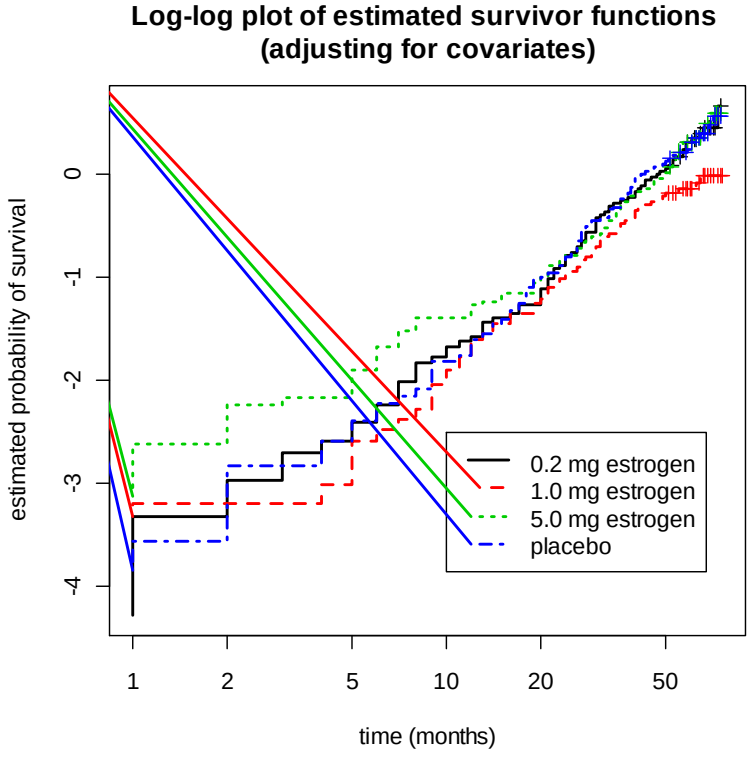
which gives a p-value for the factor `rx` of 0.1324 (on 3df) revealing how poor is the approximation used earlier.

To investigate the appropriateness of the proportional hazards assumption we do the following:

```
> #
> # fit the treatment rx as a stratum variable
> #
>
> d.ph2<-
coxph(d.surv~stage+strata(rx)+age+age+wt+pf+hx+sbp+dbp+ekg+hg+
sz+sg+ap+bm)
>
> plot(survfit(d.ph2), col=c(1:4), lty=c(1:4),lwd=2,
+ main="Kaplan-Meier Estimate of survivor functions
+ (adjusting for covariates)",
+ xlab="time (months)",
+ ylab="estimated probability of survival")
> legend(40,0.8,levels(rx),col=c(1:4),lty=c(1:4),lwd=2)
```




```
>         plot(survfit(d.ph2),         fun="cloglog",col=c(1:4),
lty=c(1:4),lwd=2,
+ main="Log-log plot of estimated survivor functions
+ (adjusting for covariates)",
+ xlab="time (months)",
+ ylab="estimated probability of survival")
> legend(10,-2.5,levels(rx),col=c(1:4),lty=c(1:4),lwd=2)
```

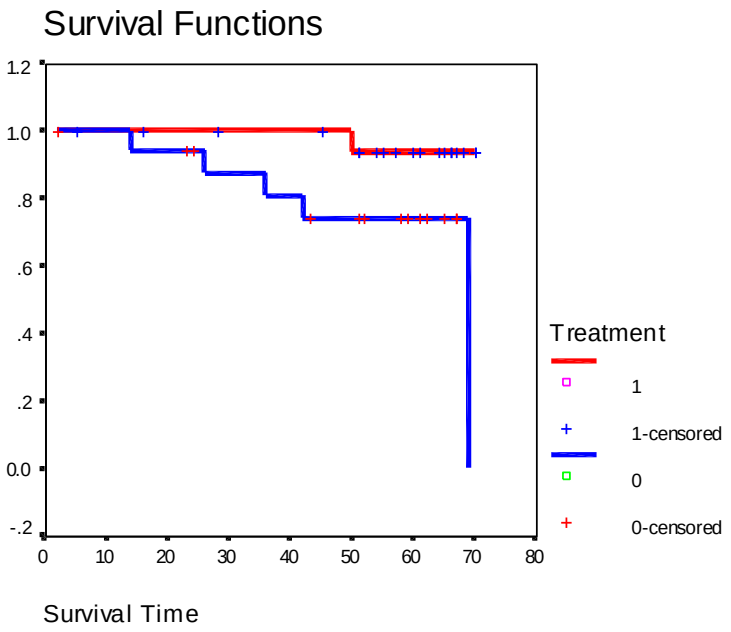


This does not indicate any substantial deviation from the proportional hazards assumption.

3) *The data file Prostatic given in SPSS, S-PLUS and Minitab formats contains data on a double blind randomised controlled clinical trial to compare treatments for prostatic cancer. The data are extracted from Collett (2003) who gives the original reference. The data file contains records for each patient of the treatment received (coded as 0 or 1 for placebo and 1.0 mg of diethylstilbestrol respectively, treatments being administered daily by mouth), survival time from entry to trial, with a status variable indicating whether or not the observation was censored (value 0) or complete (value 1), age at entry to the trial, serum haemoglobin level in gm/100ml, size of primary tumour in cm^2 and the value of a combined index of tumour stage and grade (the Gleason Index), larger values indicating a more advanced stage of tumour.*

   i) *Construct Kaplan-Meier plots of the survival times for the two treatment groups.*

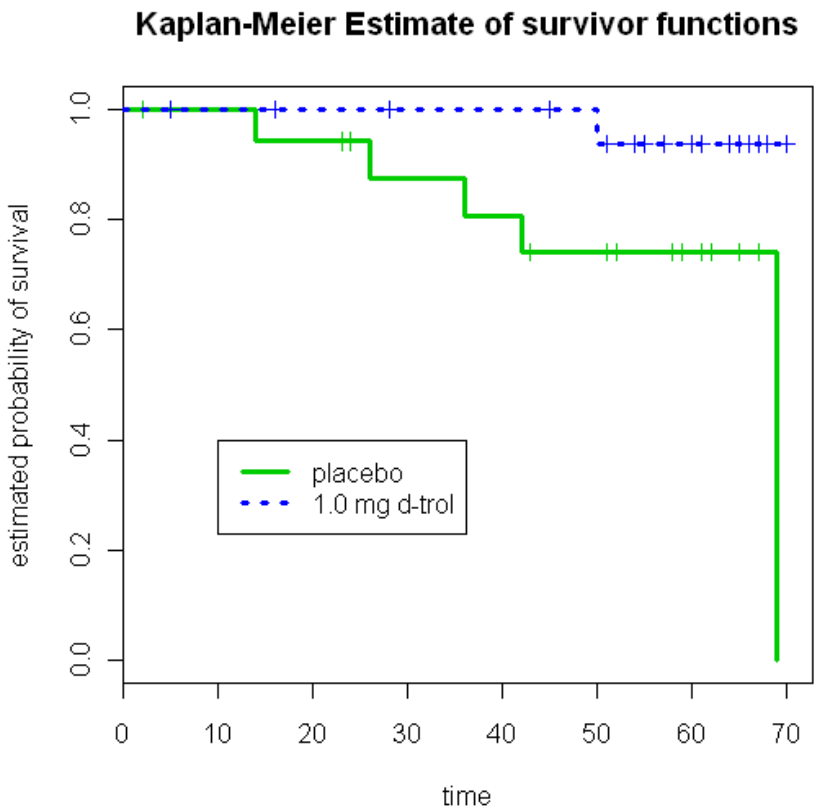
   Plot produced by SPSS, Analyze>Survival>Kaplan-Meier, (and then edited to make the lines thicker and more distinct colours — double click on picture in SPSS to call up chart editor to do this).



Note the large number of censored cases with only 1 death on treatment and that treatment has higher survival prospects than placebo.

In **R** we have

```
> attach(prostatic)
> prostatic[1:5,]
  Treatment Survival.Time Status Age Serum.Haem. Tumour.Size Gleason.Index
1         0            65      0  67        13.4          34             8
2         1            61      0  60        14.6           4            10
3         1            60      0  77        15.6           3             8
4         0            58      0  64        16.2           6             9
5         1            51      0  65        14.1          21             9
>
> library(survival)
Loading required package: splines

> prostatic.sv<-Surv(Survival.Time,Status)
> prostaticfit <- survfit(Surv(Survival.Time,Status)~ Treatment)
>
> plot(prostaticfit, lty=c(1,3), lwd=3, col=3:4,
+ main="Kaplan-Meier Estimate of survivor functions",xlab="time",
+ ylab="estimated probability of survival")
> legtext<-c("placebo", "1.0 mg d-trol")
> legend(10,0.4,legtext,lty=c(1,3), lwd=3, col=3:4)
>
```



ii) Making allowance for the values of the various covariates, assess whether the data provide evidence that the two treatment groups experience different survival prospects.

Performing a Cox Regression in SPSS with Analyze>Survival>Cox Regression gives the following. Here Treatment has been declared as categorical but then the reference category has been changed from 'last' to 'first' (& then click on 'change') to ensure that it keeps the effective coding of Treatment as 0 for placebo and 1 for treatment instead of swapping them around, so coefficient < 0 indicates enhanced survival. In S-PLUS this is not necessary and apart from this the parameter estimates and standard errors etc are essentially identical. Investigation of treating 'Gleason' as categorical reveals this is not sensible since there is evidence of gross over-fitting (large estimates with enormous standard errors).

Variables	B	SE	Sig.	Exp(B)
TREATMEN	-1.182	1.210	.329	.307
AGE	.044	.072	.541	1.045
SERUM_HA	-.022	.453	.961	.978
TUMOUR_S	.094	.052	.071	1.099
GLEASON	.723	.350	.039	2.061

The conclusion to be drawn is that although the hazard ratio of those on treatment to placebo is estimated as about 0.3 there is little evidence that this is not due to the differing values of the covariates in the two treatment groups, notably Gleason index (treated on a linear scale) and tumour size.

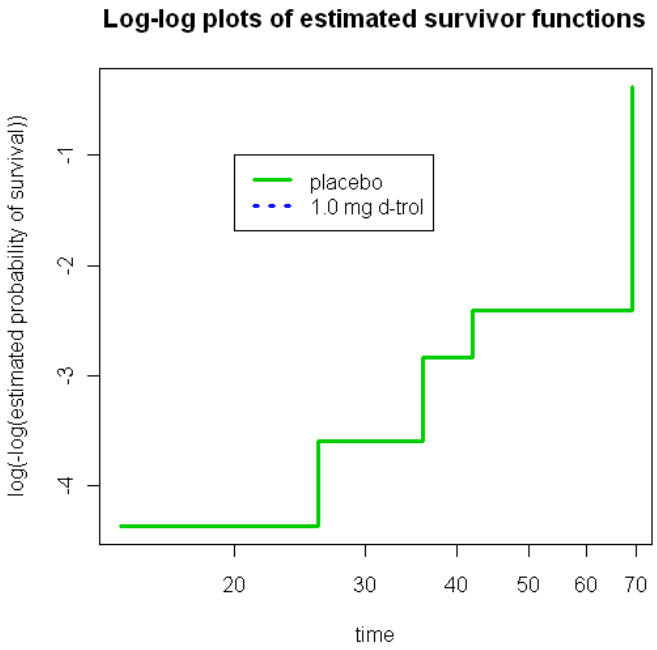
In **R** we have

```
> prostatic.ph<-coxph(prostatic.sv ~ Treatment + Age + Serum.Haem. +
Tumour.Size + Gleason.Index)
> options(digits=2)
> summary(prostatic.ph)
Call:
coxph(formula = prostatic.sv ~ Treatment + Age + Serum.Haem. +
    Tumour.Size + Gleason.Index)

  n= 38

                coef exp(coef) se(coef)     z Pr(>|z|)
Treatment     -1.1821    0.3066   1.2103 -0.98    0.329
Age            0.0440    1.0450   0.0720  0.61    0.541
Serum.Haem.   -0.0221    0.9781   0.4527 -0.05    0.961
Tumour.Size    0.0940    1.0985   0.0521  1.80    0.071 .
Gleason.Index  0.7234    2.0615   0.3500  2.07    0.039 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
Treatment         0.307      3.261    0.0286      3.29
Age               1.045      0.957    0.9074      1.20
Serum.Haem.       0.978      1.022    0.4027      2.38
Tumour.Size       1.099      0.910    0.9919      1.22
Gleason.Index     2.061      0.485    1.0382      4.09

Rsquare= 0.311   (max possible= 0.616 )
Likelihood ratio test= 14.2  on 5 df,   p=0.0145
Wald test            = 10.1  on 5 df,   p=0.0735
Score (logrank) test = 15  on 5 df,   p=0.0104
```

iii) Construct a log–log plot for treatment, averaging over other covariates.

In **R**

```
> prostatic.ph2<-coxph(prostatic.sv ~ strata(Treatment) + Age +
Serum.Haem. + Tumour.Size + Gleason.Index)
>  plot(survfit(prostatic.ph2),fun="cloglog",  lty=c(1,3),  lwd=3,
col=3:4,
+ main="Log-log plots of estimated survivor functions",xlab="time",
+ ylab="log(-log(estimated probability of survival))")
> legtext<-c("placebo", "1.0 mg d-trol")
> legend(20,-1,legtext,lty=c(1,3), lwd=3, col=3:4)
>
```



Note that the estimated survivor function for those on medication takes only two values (since there is just one event) and since log(–log(1.0)) = log(0) = ∞ the plot of that function is suppressed.

iv) ✭*Choosing any parametric regression (see Survival tasks 4) model which does* ***not*** *have the proportional hazards property, fit the model and assess whether this alters your conclusions reached in part ii).*

Choosing the lognormal distribution (which does not have the proportional hazards property and using `survreg()` gives

```
> prostatic.ln<-survreg(prostatic.sv ~ Treatment + Age +
+Serum.Haem. + Tumour.Size + Gleason.Index, dist="lognormal")
> summary(prostatic.ln)
                Value Std. Error      z      p
(Intercept)   10.1405     4.2999  2.358 0.0184
Treatment      0.7338     0.4593  1.598 0.1101
Age           -0.0226     0.0349 -0.646 0.5186
Serum.Haem.   -0.0449     0.1426 -0.315 0.7531
Tumour.Size   -0.0288     0.0203 -1.422 0.1552
Gleason.Index -0.3285     0.1753 -1.875 0.0609
Log(scale)    -0.4799     0.3123 -1.536 0.1244

Scale= 0.619
```

(omitting some details).

At first sight these results may appear to be substantially different from those using the Cox model (e.g. signs of coefficients are reversed) but this is because parametric models consider survival times rather than hazard rates.

v) ✭ *Choosing a parametric AFT model, estimate the parameters and compare your conclusions with those from parts ii) and iv).*

Accelerated failure time models are available in the library `eha` which must be downloaded and opened and regression function `aftreg()`.

```
> library(eha)
>  prostatic.gp<-aftreg(prostatic.sv  ~  Treatment  +  Age  +
Serum.Haem. +
+ Tumour.Size + Gleason.Index, dist="gompertz")
>
> summary(prostatic.gp)
Call:
aftreg(formula = prostatic.sv ~ Treatment + Age + Serum.Haem.
+
    Tumour.Size + Gleason.Index, dist = "gompertz")

Covariate          W.mean      Coef Exp(Coef)  se(Coef)    Wald p
Treatment           0.566    -0.956     0.384     1.127     0.396
Age                68.043    -0.005     0.995     0.043     0.903
Serum.Haem.        14.086     0.132     1.141     0.369     0.720
Tumour.Size        10.210     0.064     1.066     0.034     0.061
Gleason.Index       9.037     0.486     1.625     0.240     0.043

log(scale)                   12.732 338331.237     6.359     0.045

 Shape is fixed at  1

Events                    6
Total time at risk         1890
Max. log. likelihood      -33.237
LR test statistic         14.2
Degrees of freedom        5
Overall p-value           0.0144738
>
```

[credit will not be lost if parts iii) – v) are not submitted, they are for 'interest' and as an aid to those continuing to MAS6062]

Notes & Solutions for Exercises 3

Numerical solutions to hand calculations

1 (kidney tumour data)

vii) **R**

viii) Medians from K-M estimates: 26.85 and 11.005.

ix) Estimates and CIs for λ_1 and λ_2: 0.024 (.0044,0.0436) and 0.0586 (0.0261, 0.0911)

x) Log-rank test statistic is 3.53.

xi) LRT statistic is 3.56

xii) LRT statistic becomes 2.66

2 (Ovarian cancer)

iv) medians: 504.5 and outside range so no estimate possible.

v) **R**

vi) **R**

3 (Dialysis valves)

viii) **R**

ix) **R**

x) Estimates of λ_1 and λ_2 are 0.0290 and 0.0293, estimates of medians are 23.9 hours and 23.6 hours with s.e.s of 5.79 and 5.74 hours.

xi) z-statistic is 0,037.

xii) Yes. Survivor plots on log scale should be straight lines starting at (0,0) but here they cross.

xiii) Crossing survival curves indicates low power not invalidity of log-rank test.

xiv) **R**.

4 (macrophages)

vii) **R**.

viii) Medians: 26.29, 13.89 and 20.2.

ix) **R**.

x) **R**.

xi) Crossing survival curves indicates low power not invalidity

xii) **R**.

5 Hip transplants.

ii) **R**

APPENDIX 0: Maximum Likelihood Estimation

A0.0 Estimation

Suppose $x_1,\ldots,x_n$ are n independent observations of a random variable X which has density function $f(.;\theta)$ depending on an unknown parameter θ. There are various methods of estimating θ from the observations $x_1,\ldots,x_n$ — such as the method of least squares, the method of moments, the method of minimum chi-squared, ….. etc. 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 is essentially 'the probability of observing the data $x_1,\ldots,x_n$' (this probability 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***.

A0.1 Definition

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

$$L(\theta;x_1,\ldots,x_n)=f(x_1;\theta)f(x_2;\theta)\ldots.f(x_n;\theta) \text{ if X is continuous}$$

or $$L(\theta;x_1,\ldots,x_n)=P[X=x_1;\theta]P[X=x_2;\theta]\ldots.P[X=x_n;\theta] \text{ if X is discrete}$$

(i.e. it is the product of the values of the density function or probability function evaluated at each of the observations — it is the 'probability' of observing the data just obtained).

A0.2 Examples

(all with data $x_1,\ldots,x_n$)

(i) $X\sim N(\mu,1)$: $f(x;\mu)= (2\pi)^{-½}\exp\{-½(x-\mu)^2\}$

$L(\mu;x_1,\ldots,x_n)= (2\pi)^{-½n}\exp\{-½\Sigma(x_i-\mu)^2\}$

(ii) $X\sim Ex(\lambda)$: $f(x;\lambda)=\lambda e^{-\lambda x}$

$L(\lambda;x_1,\ldots,x_n)= \lambda^n\exp\{-\lambda\Sigma x_i\}$

(iii) $X\sim Bin(m,p)$: $P[X=x]={}^mC_x p^x(1-p)^{m-x}$

$$L(p; x_1,\ldots,x_n)=\prod_{i=1}\binom{m}{x_i}p^{\sum x_i}(1-p)^{\sum(m-x_i)}$$

(iv) $X\sim N(\mu,\sigma^2)$: $f(x;\mu,\sigma)= (2\pi)^{-½}\sigma^{-1}\exp\{-½(x-\mu)^2/\sigma^2\}$

$L(\mu,\sigma;x_1,\ldots,x_n)= (2\pi)^{-½n}\sigma^{-n}\exp\{-½\Sigma(x_i-\mu)^2/\sigma^2\}$

(note that in this example the parameter $\theta=(\mu,\sigma)$ has two components)

(v) $X\sim Po(\lambda)$: $P[X=x]= \lambda^x e^{-\lambda}/x!$

$L(\lambda;x_1,\ldots,x_n)=\lambda^{\Sigma x}e^{-n\lambda}/\Pi x_i!$

To obtain the maximum likelihood estimates of the parameters in these cases we maximize the likelihoods w.r.t the unknown parameters. 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).

(i) $L(\mu;x_1,\ldots,x_n)= (2\pi)^{-½n}\exp\{-½\Sigma(x_i-\mu)^2\}$

$\log(L(\mu))=\ell(\mu)=-½n\log(2\pi)-½\Sigma(x_i-\mu)^2$

$\partial\ell/\partial\mu=\Sigma(x_i-\mu)$ and setting this to zero gives $\Sigma x_i=n\mu$ so $\hat{\mu}=\bar{x}$

(the hat ^ on the parameter indicates
that it is the estimate of the parameter)

(ii) $L(\lambda;x_1,\ldots,x_n)= \lambda^n\exp\{-\lambda\Sigma x_i\}\log(L(\lambda))=\ell(\lambda)=n\log(\lambda)-\lambda\Sigma x_i$

$\partial\ell/\partial\lambda=n/\lambda-\Sigma x_i$ and so $\hat{\lambda} = \frac{1}{\bar{x}}$

(iii) $L(p; x_1,\ldots,x_n)=\prod_{i=1}\binom{m}{x_i} p^{\Sigma x_i}(1-p)^{\Sigma(m-x_i)}$

$\log(L(p))=\ell(p)= \Sigma x_i\log(p) + \Sigma(m-x_i)\log(1-p) + K$

(K a constant not involving p)

$\partial\ell/\partial p=\Sigma x_i/p -\Sigma(m-x_i)/(1-p)$ and so $\hat{p} = \bar{x}/m$

(iv) $L(\mu,\sigma;x_1,\ldots,x_n)= (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\{-\frac{1}{2}\Sigma(x_i-\mu)^2/\sigma^2\}$

$\log(L(\mu,\sigma))=\ell(\mu,\sigma)=-\frac{1}{2}n\log(2\pi)-n\log(\sigma)-\frac{1}{2}\Sigma(x_i-\mu)^2/\sigma^2$

so $\partial\ell/\partial\mu=\Sigma(x_i-\mu)/\sigma^2$ and $\partial\ell/\partial\sigma=-n/\sigma+\Sigma(x_i-\mu)^2/\sigma^3$

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

(v) $L(\lambda;x_1,\ldots,x_n)=\lambda^{\Sigma x}e^{-n\lambda}/\Pi x_i!$

$\log(L(\lambda))=\ell(\lambda)=\Sigma x_i\log(\lambda) - n\lambda + K$

$\partial\ell/\partial\lambda=\Sigma x_i/\lambda -n$ so $\hat{\lambda} = \bar{x}$.

A0.3 Further 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.

To obtain the variance of the mle we need to calculate the expected value of the second derivative of the log-likelihood $E[(\partial^2\ell/\partial\theta^2)]$ and then the variance is minus the reciprocal of this, i.e.

$$\text{var}(\hat{\theta}) = -\{E[(\partial^2\ell/\partial\theta^2)]\}^{-1}$$

(note: if θ is a vector parameter of dimension p then we can interpret $\partial^2\ell/\partial\theta^2$ as a p×p matrix in which case we need to take the inverse of the matrix of expected values to get the variance-covariance matrix. To get just the variances of individual mles we can work with them singly and this is ok if we only want individual confidence intervals.

A0.4 Examples:

(continuing the examples above)

(i) $\partial\ell/\partial\mu=\Sigma(x_i-\mu)$, so $\partial^2\ell/\partial\mu^2=-n$ and thus $E[\partial^2\ell/\partial\mu^2]=-n$ and thus

$$\operatorname{var}\hat{\mu}=n^{-1}$$

(ii) $\partial\ell/\partial\lambda=n/\lambda-\Sigma x_i$, so $\partial^2\ell/\partial\lambda^2=-n/\lambda^2$, so $E[\partial^2\ell/\partial\lambda^2]=-n/\lambda^2$ and thus $\operatorname{var}(\hat{\lambda}) = \lambda^2/n$. In this case we would substitute the mle $\hat{\lambda}$ for λ to get the standard error of $\hat{\lambda}$ as $\hat{\lambda}/n^{½}$

(iii) $\partial\ell/\partial p=\Sigma x_i/p \ -\Sigma(m-x_i)/(1-p)$ so $\partial^2\ell/\partial p^2=-\Sigma x_i/p^2 \ +\Sigma(m-x_i)/(1-p)^2$ and thus

$$E[\partial^2\ell/\partial p^2]=-\Sigma mp/p^2 +\Sigma(m-mp)/(1-p)^2$$

(noting that $E[x_i]=mp$ for each i

$$=-nm/p+nm/(1-p)=-nm/p(1-p)$$

and so $\operatorname{var}(\hat{p})=p(1-p)/nm$

(iv) $\partial\ell/\partial\mu=\Sigma(x_i-\mu)/\sigma^2$ and $\partial\ell/\partial\sigma=-n/\sigma+\Sigma(x_i-\mu)^2/\sigma^3$ so

$\partial^2\ell/\partial\mu^2=-n/\sigma^2$ and $\partial^2\ell/\partial\sigma^2=n/\sigma^2-3.\Sigma(x_i-\mu)^2/\sigma^4$

and $\partial^2\ell/\partial\mu\partial\sigma=-\Sigma(x_i-\mu)/\sigma^3$

Now $E[x_i-\mu]=0$ and $E[(x_i-\mu)^2]=\sigma^2$ so $E[\partial^2\ell/\partial\mu^2]=-n/\sigma^2$,

$E[\partial^2\ell/\partial\sigma^2]=-2n/\sigma^2$ and $E[\partial^2\ell/\partial\mu\partial\sigma]=0$ and thus we have

$\text{var}(\hat{\mu})=\sigma^2/n$, $\text{var}(\hat{\sigma})=\sigma^2/2n$ and $\text{cov}(\hat{\mu},\hat{\sigma})=0$.

(v) $\partial\ell/\partial\lambda=\Sigma x_i/\lambda$ so $\partial^2\ell/\partial\lambda^2=-\Sigma x_i/\lambda^2$ and we have that $E[x_i]=\lambda$ so $E[\partial^2\ell/\partial\lambda^2]=-n/\lambda$ and thus $\text{var}(\hat{\lambda})=\lambda/n$

Again, in examples (iii)–(v) we would substitute the mles for the unknown parameters in the expressions for the variances to get standard errors (taking square roots) and thus obtain an approximate 95% confidence interval as **mle** $\pm$ **2** $\times$ **s.e.(mle)** ,

i.e. an approximate 95% confidence interval for θ **is** $\hat{\theta}\pm 2\times$**s.e.(**$\hat{\theta}$**)**

A0.5 [Generalized] Likelihood Ratio Tests

A useful procedure for constructing hypothesis tests is an adaptation of the simple likelihood ratio test — recall that the Neyman-Pearson lemma shews that the most powerful test of a given size of one simple hypothesis against another is based on the likelihood ratio. (A simple hypothesis is one that involves no unknown parameters — the likelihood is fully specified under the hypothesis). The generalization is that [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.

In fact, it is more usual to consider the [natural] logarithm of this ratio (or equivalently the difference in maximized log-likelihoods) since there are theoretical results that allow the significance level of this statistic to be calculated.

Specifically, if we have data $x_1,\ldots,x_n$ from a random variable X whose distribution depends on a parameter θ and if we are testing a hypothesis H_0 against an alternative H_A then ***the likelihood ratio statistic*** is $\lambda=2\{\ell(\hat{\theta}_A)-\ell(\hat{\theta}_o)\}$ where $\hat{\theta}_A$ and $\hat{\theta}_o$ are the estimates of θ under the hypotheses H_A and H_0 respectively. H_0 is rejected in favour of H_A if λ is sufficiently large. It can be shewn that for large sample sizes λ is approximately distributed as χ^2 on r degrees of freedom, where r is the difference in numbers of parameters estimated under H_A and H_0. Note that $\ell(\hat{\theta}_A)$ and $\ell(\hat{\theta}_o)$ are the actual maximum values of the log-likelihoods under H_A and H_0. Sometimes we cannot obtain mles

explicitly (or algebraically) but we can obtain the maximum values of the log-likelihoods numerically using some general optimization program.

A0.5.1 Examples

(all with data $x_1,\ldots,x_n$)

(i) $X\sim N(\mu,1)$; to test H_0: $\mu=0$ *vs.* H_A: $\mu \neq 0$

Now $L(\mu)=(2\pi)^{-\frac{1}{2}n}\exp\{-\frac{1}{2}\Sigma(x_i-\mu)^2\}$

Under H_0 , $\mu=0$, so under H_0 the maximum (in fact the only) value of $L(\mu)$ is $(2\pi)^{-\frac{1}{2}n}\exp\{-\frac{1}{2}\Sigma x_i^2\}$,

i.e. $\hat{\mu}_0=0$ and $\ell(\hat{\mu}_0)=-\frac{1}{2}n\log(2\pi)-\frac{1}{2}\Sigma x_i^2$

Under H_A we just have the ordinary likelihood and the mle of μ is $\hat{\mu}_A=\bar{x}$ giving $\ell(\hat{\mu}_A)=-\frac{1}{2}n\log(2\pi)-\frac{1}{2}\Sigma(x_i-\bar{x})^2$, this gives the likelihood ratio statistic as $\lambda=-2\{\ell(\hat{\mu}_A)-\ell(\hat{\mu}_0)\}=\Sigma x_i^2-\Sigma(x_i-\bar{x})^2=n\bar{x}^2$ and we reject H_0 if this is large when compared with χ_1^2.

(ii) $X\sim Ex(\lambda)$; to test H_0: $\lambda=\lambda_0$ *vs.* H_A: $\lambda \neq \lambda_0$

$L(\lambda)=\lambda^n\exp\{-\lambda\Sigma x_i\}$.

Under H_0 $\lambda=\lambda_0$ so $\hat{\lambda}_0=\lambda_0$ and $\ell(\hat{\lambda}_0)=n\log(\lambda_0)-\lambda_0\Sigma x_i\}$.

Under H_A we have $\hat{\lambda}_A=\frac{1}{\bar{x}}$ so $\ell(\hat{\lambda}_A)=n\log(\bar{x})-n$

and the lrt statistic is $2\{n\log(\bar{x})-n-n\log(\lambda_0)+\lambda_0\Sigma x_i\}$ which would be referred to a χ_1^2 distribution.

(iii) $X\sim N(\mu,\sigma^2)$; to test H_0: $\mu=0$ *vs.* $H_A\neq 0$ with σ^2 unknown.

Here we need to estimate σ under both H_0 (i.e. assuming $\mu=0$) and under H_A (not assuming $\mu=0$) and use these estimates in maximizing the likelihoods.

We have $\ell(\mu,\sigma)=-\frac{1}{2}n\log(2\pi)-n\log(\sigma)-\frac{1}{2}\Sigma(x_i-\mu)^2/\sigma^2$ so under H_0 we have $\hat{\mu}_0=0$ and $\sigma_0^2=\frac{1}{n}\Sigma x_1^{\,2}$

and then $\ell(\hat{\mu}_0,\sigma_0^2)=-\frac{1}{2}n\log(2\pi)-\frac{1}{2}n\log(\frac{1}{n}\Sigma x_i^2)-\frac{1}{2}n$.

Under H_A we have $\hat{\mu}_A=\bar{x}$ and $\hat{\sigma}_A^2=\frac{1}{n}\Sigma(x_i-\bar{x})^2$ giving

$\ell(\hat{\mu}_A,\hat{\sigma}_A^2)=-\frac{1}{2}n\log(2\pi)-\frac{1}{2}n\log(\frac{1}{n}\Sigma(x_i-\bar{x})^2)-\frac{1}{2}n$ and thus the lrt statistic is $\lambda=\{n\log(\Sigma x_1^{\,2)}-n\log(\Sigma(x_i-\bar{x})^2))$ which would be referred to a χ_1^2 distribution. (Note that the $\frac{1}{n}$ terms in the logs are $-n\log(n)$ and so cancel each other). It can be shewn that this statistic is a monotonic function of (and therefore equivalent to) the usual t-statistic for testing $\mu=0$ when σ is unknown.

Further examples to try are:

(i) $x_i\sim N(\mu,\sigma^2)$, H_0: $\sigma^2=\sigma_0^{\,2}$, μ unknown, H_A: $\sigma^2\neq\sigma_0^{\,2}$.

(ii) $x_i\sim N(\mu,\sigma^2)$, H_0: $\sigma^2=\sigma_0^{\,2}$, μ known, H_A: $\sigma^2\neq\sigma_0^{\,2}$.

(iii) $x_i\sim Bin(m,p)$, m known, H_0: $p=p_0$, H_A: $p\neq p_0$,

(iv) $x_i\sim Po(\lambda)$, H_0: $\lambda=\lambda_0$, H_A: $\lambda\neq\lambda_0$.