Survival Data Analysis

Outline Solutions to Exercises

Dr Nick Fieller

Department of Probability & Statistics

University of Sheffield

visiting

2012

Notes & Solutions

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	Died	withdrawn						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) 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*

i) 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)

(i) 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	$23.0 \le t$

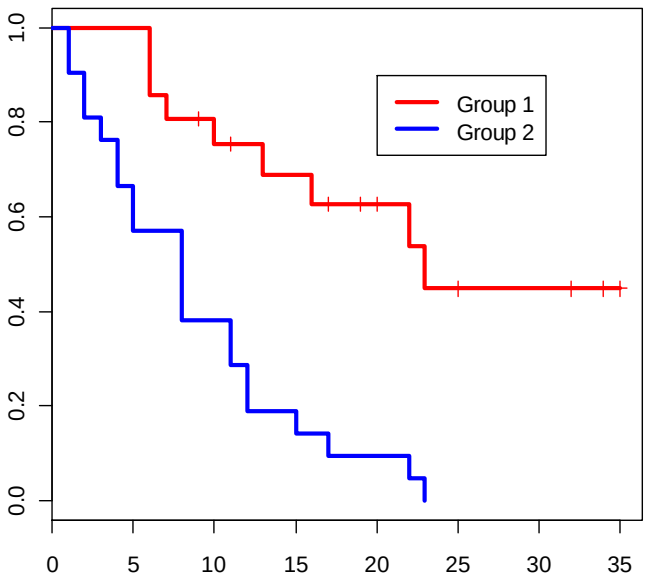
Next the controls:

j	$t_{(j)}$	l_j	r_j	d_j	1	0≤t<1.0
1	1	0	21	2	0.905	1.0≤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"))
```



ii) 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 $\tilde{t}$ such that $\hat{S}(\tilde{t}) = 0.5$ so we have

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

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

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

3) *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.*

   i) Plot the Kaplan-Meier estimates of the survival times of teaspoons

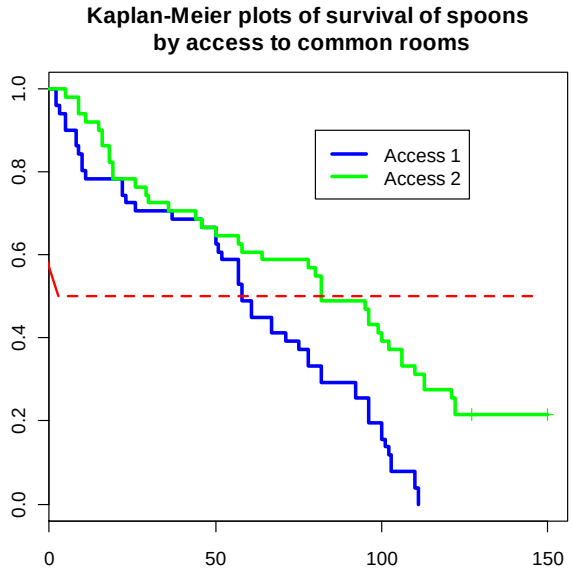
   ii) Estimate the median survival times in the two categories of rooms.

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

4) The data given in file ovarian.Rdata represent survival times in days of 26 patients randomized to one of two forms of chemotherapy (indicated by variable ***treat*** *as 1 or 2) following surgery for ovarian cancer, where status records whether the observation is censored (status = 0) or complete (status =1). Also given are variables* ***age,*** ***rdisease*** *and* ***perf*** *which give information on relevant covariatesfor each subject.*

(Source: Collett, 2003).

i) Compute and plot the Kaplan-Meier product limit estimates of the survivor functions for treatments 1 and 2 provide estimates of the median survival times based upon the Kaplan Meier estimates.

ii) Assess the evidence of a difference between the two treatments provided by a log-rank test.

Solution to be provided later.

5) *For the data on the data leukaemia remission times)*

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

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

iii) *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.*

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

v) *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.*

vi) *Comment on the effect of the drug.*

(iii) 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	1	6	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 S-plus 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

>
```

(iv)

	$\Sigma\delta_i$	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)

(v) 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.

(vi) 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.

(vii) Plots of log[exp{$-\hat{\lambda}_i t$} 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.

(viii) There is strong evidence that the drug prolongs survival.

6) ★*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?*

Solution to follow later

7) ★*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)?*

Solution to follow later

8) *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.

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.

9) *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.

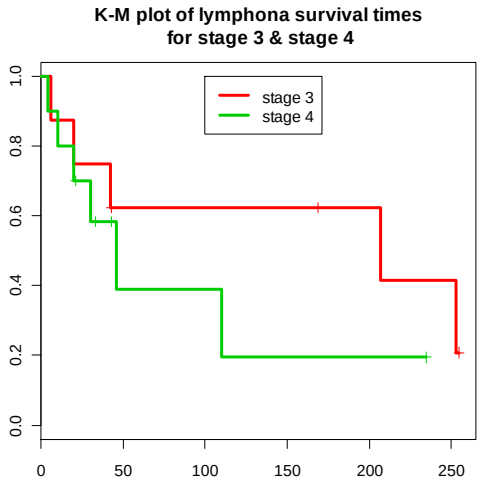
ii) Provide estimates of the two cumulative hazard functions and comment on any differences.

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

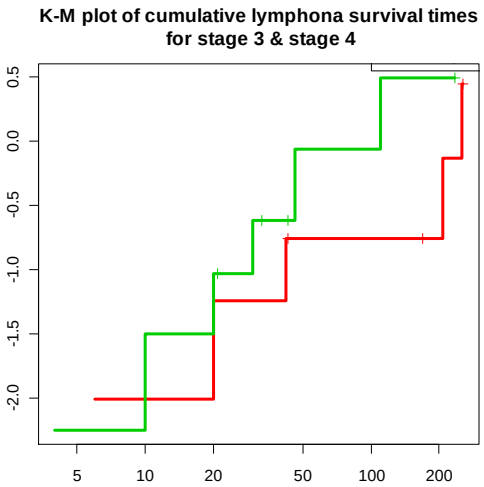
First the S-Plus version, first using the menu in `Statistics>Survival>Nonparametric Survival…)`. Note the production of the Survival plot on a logarithmic vertical scale by clicking the appropriate box on the dialogue box of the plot menu.

```
> 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
>
```

10) ★ *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?

Solution to follow later

11) 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/327/7429/1459.full.pdf BMJ VOLUME 331 24-31 DECEMBER 200, p1498-1500

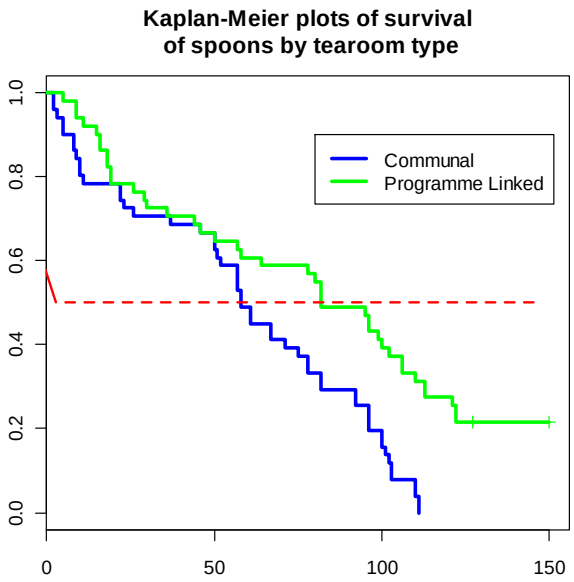
Script file: #Q4

```
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)
```

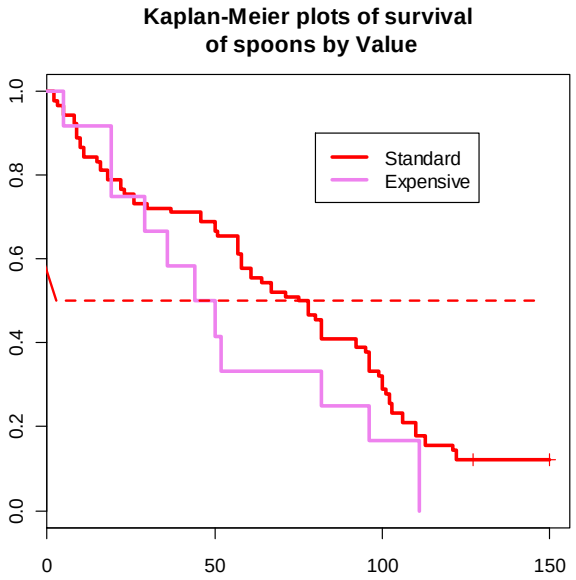
```
missing<-sum(complete)
missprog<-sum(complete*(Access-1))
misscomm<-missing-missprog
missprog;misscomm
max(Day[Access*complete==1])
max(Day[(Access-1)*complete==1])
```

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) (how many are needed at the start of the year to provide 75 by the end of the year) we really need to know the number of people in the institution where the study took place (which actually was 140 in the original publication). If we assume that there were 150 in the study (perhaps the best guess without any other information), then we can proceed. Next 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.

12) 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)

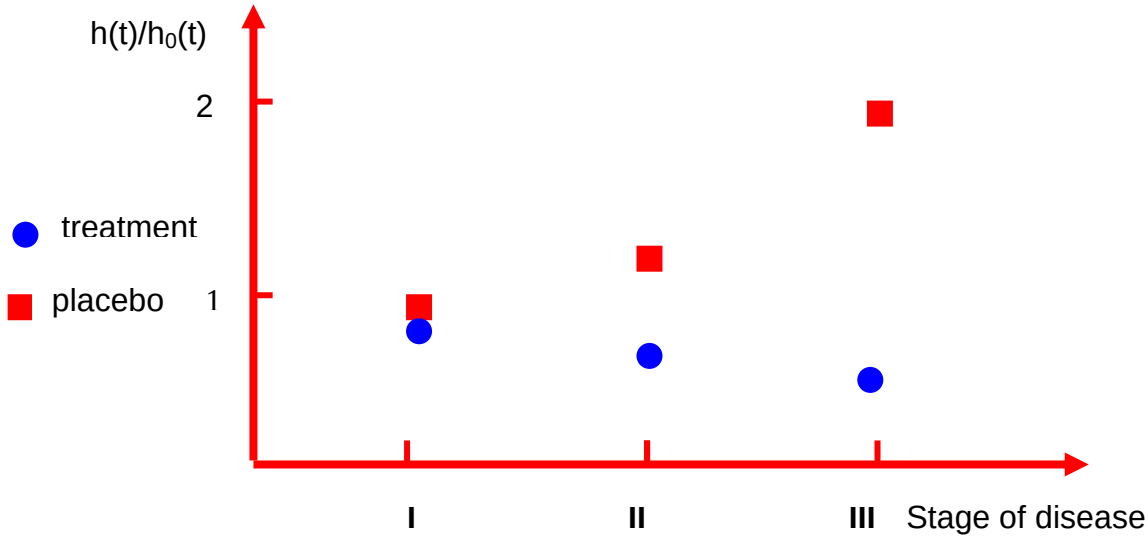
treatment, stage III: 0.42, (0.16, 1.11)

(assuming estimates of the β_i are independent). Note, CIs calculated as estimate+/– 2×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. $\exp\{\beta_1\}$ take exp{CI for β_1}.

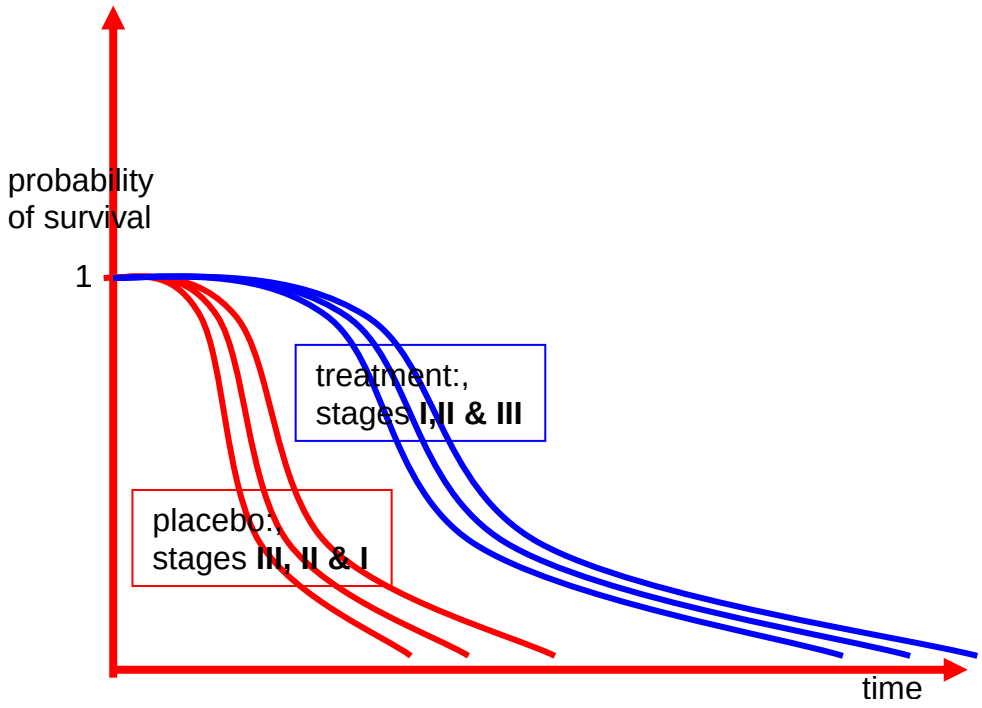
Thus, on untreated patients the hazards increase with stage of disease and so their survival prospects decrease with stage of disease. For patients on the treatment the effect of stage of disease is negated and indeed perhaps slightly reversed, though looking att he CIs the evidence for an actual improvement 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:



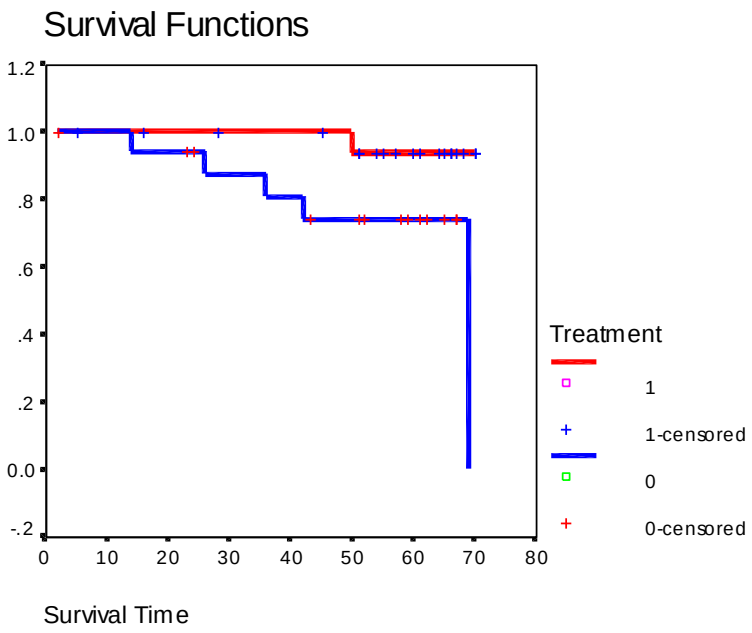
This allows a sketch of the relative positions of the survival times



13) 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.

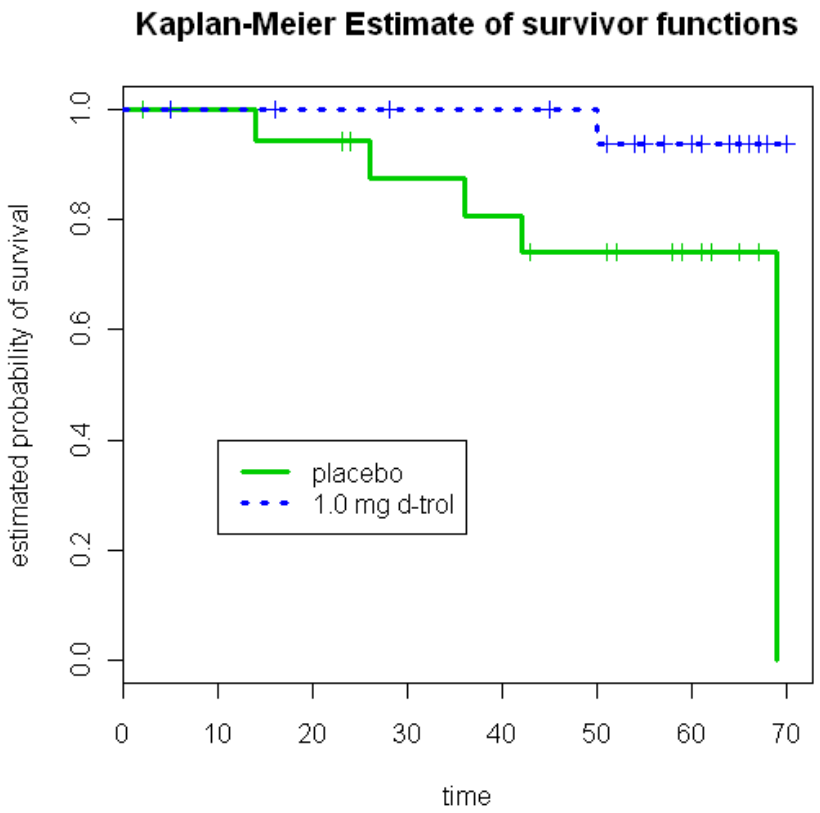
Plot produced by SPSS, Analyze>Survival>Kaplan-Meier, (and then edited to make the lines thicker and more distinct colours — double click on picture in SPSS to call up chart editor to do this).



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.

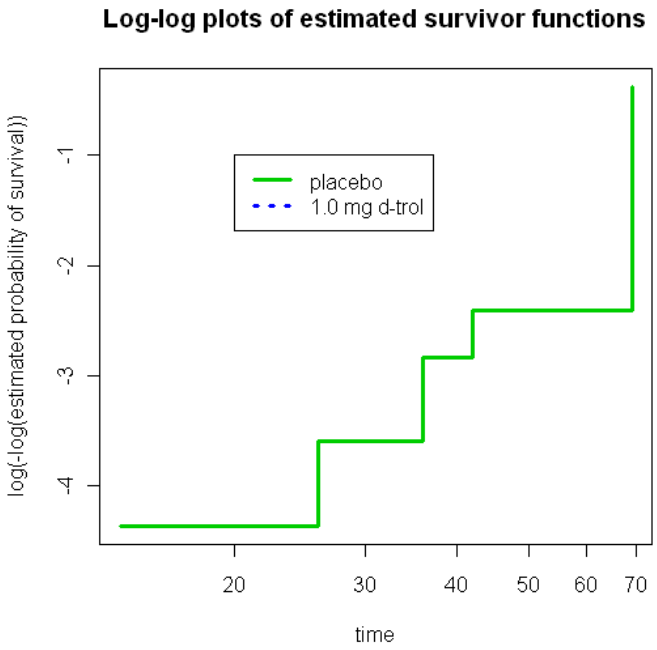
In **R** we have

```
> 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 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
>
```


14) *Returning to the data on ovarian cancer given in Q4 (data ovarian.Rdata), assess the evidence for a difference between the two treatments after making an allowance for the various covariates using*

i) a Cox proportional hazards regression model

ii) an exponential regression model

iii) ⋆*a Weibull regression model*

Solution to follow later

15) ⋆⋆⋆*The data in file methtrex.Rdata arise from a study of treatment for primary biliary cirrhosis. Sixty subjects were randomized into two treatment groups, one receiving Methotrexate and the other receiving a placebo.*

The data consist of 10 variables measured on 60 subjects. The variables are:

AGE:– Age (Years)
ALBUMIN:– Serum albumin (g/L)
AMA:– AMA Antimitochondrial antibody, (0=negative, 1=positive)
BILIRUBIN:– Serum bilirubin (μmol/L)
FOLLOWUP:– months of survival of liver to either transplant or death of subject or end of study
LUDWIG:– Ludwig stage on 4 point scale.
MAYO:– Mayo Clinic score
PROTHROM:– Prothrombin time (seconds)
STATUS:– censoring status (1=event occurred, 0=event not occurred before end of study)
TREATMNT:– Treatment (1= Methotrexate, 0= Placebo)

The prime question of interest is whether there is evidence of a difference in survival patterns of those receiving the two treatments, after making due allowance for any relevant covariates.

Notes

i) It may be noted that an attempt to fit a Cox proportional hazards model which includes the raw Ludwig categories usually results in a message indicating that convergence has not been achieved, although parameter estimates are given — this lack of convergence suggests that the estimates are not to be trusted.

ii) Discussion with the clinician involved in the study has suggested that the Ludwig stage might usefully be regarded as either a distinction between 1&2 vs 3&4 or as a distinction between 1&2&3 vs 4 (or even 1 vs 2&3&4), i.e. that the 4 categories on this variable may need to be condensed onto a two-point scale.

iii) Further discussions suggest that there is some doubt about whether to include subjects who have tested negative to antimitochondrial antibodies.

iv) Additionally, he has pointed out that the Mayo Clinic Score is based at least in part on the values of several of the other recorded variables.

*Please note that these data are confidential. They have been provided for educational purposes **ONLY** by a clinician at the Gastroenterology & Liver Unit, Royal Hallamshire Hospital, Sheffield. The data remain the property of Royal Hallamshire Hospital and may not be used for any purpose whatsoever beyond work for this course. They should not be copied to any third person for any reason whatsoever. ALL electronic copies (including those on back-up files and including any derived data files in any format) should be deleted when you have finished with your studies.*

Solution to follow later