Survival Data Analysis

Exercises

Dr Nick Fieller

Department of Probability & Statistics

University of Sheffield

visiting

2012

N.B. *a ✭ indicates that the question is beyond the standard scope of the course.*

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) 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

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).

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

(The data are given in file leukrem.Rdata)

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.

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 1and 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.

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

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

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

ii) 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.

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

iv) 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.

v) Comment on the effect of the drug.

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?

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)?

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)
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.

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.

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

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?

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).

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.

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.

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.

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

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

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

vii) ★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).

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

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

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.