A.3 Survival regression models and testing for survival distributions

Questions 1–3 are to be done for discussion in class. Questions 4–6 are to be handed in for marking.

1. (a) Suppose we have a random sample that includes right-censored data (censoring assumed non-informative). We wish to decide whether or not a Weibull distribution is appropriate. Using an estimator of the survival function how might we graphically inves- tigate the appropriateness of the model? Given that the model appears to be appropriate how would you test whether or not the special case of an exponential model is valid? Suppose that the Weibull model does not appear to be appropriate what graph would you use to consider a log-logistic model?

   (b) Now suppose that there are two groups to be considered (eg smokers v. non-smokers). What graphs would be appropriate for consideration of a proportional hazards model, accelerated life model respectively?

   (c) Gehan (1965) studied 42 leukaemia patients. Some were treated with the drug *6-mercaptopurine* and the rest are controls. The trial was designed as matched pairs, but both members of a pair observed until both came out of remission or the study ended. (The data are included under the name `gehan` in the `R` package `MASS`. The description attached to these data there says that in each pair both were withdrawn from the trial when either came out of remission. If you have a look at the data, you can see that this is clearly not true.) The observed times to recurrence (in months) were:

$$\begin{aligned}\text{Controls: } & 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23\\ \text{Treatment: } & 6+,6,6,6,7,9+,10+,10,11+,13,16,17+,19+,20+,22,23,25+,\\ & 32+,32+,34+,35+\end{aligned}$$

   Here + indicates censored times. Investigate these data in respect of both a) and b).

2. The object `tongue` in the package `KMsurv` lists survival or right-censoring times in weeks after diagnosis for 80 patients with tongue tumours. The `type` random variable is 1 or 2, depending as the tumour was aneuploid or diploid respectively.

   (a) Use the log-rank test to test whether the difference in survival distributions is significant at the 0.05 level.

   (b) Repeat the above with a test that emphasises differences shortly after diagnosis.

3. The following is an investigation carried out by a (medium-sized) UK pension scheme into the mortality of its pensioners over the course of three years

   (a) Explain why the crude rates are usually graduated.

   (b) The data used to produce the crude rates and the proposed graduated rates are as follows.

Age	Central ExpRisk	Deaths	crude hazard	graduated hazard
x	E_x^c	d_x	μ_x	$\overset{\circ}{\mu}_x$
60–64	1388.9	10	0.0072	0.0061
65–69	1188.8	17	0.0143	0.0131
70–74	880.5	28	0.0318	0.0262
75–79	841.6	34	0.0404	0.0487
80–84	402.8	41	0.1018	0.0839
85–89	123.9	19	0.1533	0.1338
90–94	27.9	7	0.2509	0.1975
95–99	10.0	3	0.3000	0.2706
100+	7.5	2	0.2666	0.3455

Assume the Gompertz–Makeham model has been used for graduation. Is this a sensible choice? Test the proposed graduation for

i. Overall goodness of fit; and

ii. Bias.

4. (a) Sketch the shape of the hazard function in the following cases, paying attention to any changes of shape due to changes in value of κ where appropriate.

   i. Weibull: $S(t) = e^{-(\rho t)^\kappa}$.

   ii. Log-logistic: $S(t) = \frac{1}{1+(\rho t)^\kappa}$.

   (b) Suppose that it is thought that an accelerated life model is valid and that the hazard function has a maximum at a non-zero time point. Which parametric models might be appropriate?

   (c) Suppose that $t_1, \ldots, t_n$ are observations from a lifetime distribution with respective vectors of covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$. It is thought that an appropriate distribution for lifetime y is Weibull with parameters ρ, κ, where the link is $\log \rho = \beta \cdot \mathbf{x}$. Explain what x_{i0} is, and the interpretation of the parameter β_0. In the case that there is no censoring write down the likelihood and, using maximum likelihood, give equations from which the vector of estimated regression coefficients β (and also the estimate for κ) could be found.

   What would be the asymptotic distribution of the vector of estimators? How would the likelihood differ if some of the observations t_i were right censored (assuming independent censoring)?

5. (a) Describe the proportional hazards model, explaining what is meant by the partial likelihood and how this can be used to estimate resgression coefficients. How might standard errors be generated?

   (b) Drug addicts are treated at two clinics (clinic 0 and clinic 1) on a drug replacement therapy. The response variables are the time to relapse (to re-taking drugs) and the status relapse =1 and censored =0. There are three explanatory variables, clinic (0 or 1), previous stay in prison (no=0, yes=1) and the prescribed amount of the replacement dose. The following results are obtained using a proportional hazards model, $h(t, x) = e^{\beta x} h_0(t)$.

Variable	Coeff	St Err	p-value
clinic	-1.009	0.215	0.000
prison	0.327	0.167	0.051
dose	-0.035	0.006	0.000

What is the estimated hazard ratio for a subject from clinic 1 who has not been in prison as compared to a subject from clinic 0 who has been in prison, given that they are each assigned the same dose?

(c) Find a 95% confidence interval for the hazard ratio comparing those who have been in prison to those who have not, given that clinic and dose are the same.

6. Coronary Heart Disease (CHD) remains the leading cause of death in many countries. The evidence is substantial that males are at higher risk than females, but the role of genetic factors versus the gender factor is still under investigation. A study was performed to assess the gender risk of death from CHD, controlling for genetic factors. A dataset consisting of non-identical twins was assembled. The age at which each person died of CHD was recorded. Individuals who either had not died or had died from other causes had censored survival times (age). A randomly selected subsample from the data is as follows. (* indicates a censored observation.)

Age male twin	Age female twin
50	63*
49*	52
56*	70*
68	75
74*	72
69*	69*
70*	70*
67	70
74*	74*
81*	81*
61	58
75*	73*

(a) Write down the times of events and list the associated risk sets.

(b) Suppose the censoring mechanism is independent of death times due to CHD, and that the mortality rates for male and female twins satisfy the PH assumption, and let β be the regression coefficient for the binary covariate that codes gender as 0 or 1 for male or female respectively. Write down the partial-likelihood function. Using a computer or programmable calculator, compute and plot the partial-likelihood for a range of values of β. What is the Cox-regression estimate for β? What does this mean?

(c) Estimate the survival function for male twins.

(d) Suppose now only that the censoring mechanism is independent of death times due to CHD, perform the log-rank test for equivalence of hazard amongst these two groups. Contrast the test statistic and associated p-value with the results from the Fleming–Harrington test using a weight $W(t_i) = \hat{S}(t_{i-1})$.

(e) Do you think the assumption of a non-informative censoring mechanism is appropriate? Give reasons.