A.4 Model testing, proportional-hazards, accelerated life

1. (a) Adapt the delta-method procedure used in the lectures to justify Greenwood's estimate for the Kaplan-Meier curve, to give the following approximation for the Neslon-Aalen estimator of the survival distribution:

$$\operatorname{Var} \tilde{S}(t) \approx \tilde{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i(n_i - d_i)}{n_i^3}.$$

   (b) Rather than a normal approximation for the Kaplan-Meier estimator $\hat{S}(t)$ itself, or for $\log \hat{S}(t)$, some authors recommend a normal approximation for $\log\big(-\log \hat{S}(t)\big)$. One advantage is that any value in $(-\infty, \infty)$ is a possible value for this quantity (which is not the case for $\hat{S}(t)$ or $\log \hat{S}(t)$). Starting from the variance of $\log \hat{S}(t)$ as used in the justification of Greenwood's estimate, use the delta method again to obtain an approximation for the variance of $\log\big(-\log \hat{S}(t)\big)$. Find the form of the confidence interval for $\hat{S}(t)$ which results from this approach.

2. Plot (or sketch) the Kaplan-Meier curve for the following sample of size 15:

   $$1.75, 1.89, 1.92, 2.17+, 2.24, 2.27+, 2.60+, 2.88, 3.10+, 3.84+, 4.25, 4.81, 5.05+, 5.11+, 5.25$$

   where + indicates a right-censored observation. Estimate the variance of $\hat{S}(4)$ using Greenwood's formula. Find a 95% confidence interval for $S(4)$ using one (or more, if you like) of the approaches mentioned above.

3. The table below shows data collected by a group of specialist care homes on their residents. For each individual, the following data was available: (curtate) age on entering the home, and age on death or on exit from the home for another reason (or current age for those still living and resident). The following shows the number of arrivals, deaths and other exits recorded at each age:

Age	Arrivals	Deaths	Other Exits
68	1	0	0
69	5	0	2
70	10	2	1
71	24	5	3
72	24	6	3
73	11	9	9
74	11	8	3
75	4	5	4
76	4	6	2

   (a) What sorts of censoring and/or truncation do we have in this data?

   (b) Make a table showing the number of individuals at risk at each age (suitably approximated).

   (c) Find a confidence interval for the probability that an individual on his or her 71st birthday survives a further 5 years.

   (d) What assumptions about the population do we need to make in the calculation above? Are they reasonable?

4. Parametric families of survival distributions with the accelerated lifetime property include Weibull $S(t) = \exp(-(\rho t)^\alpha)$, log-logistic $S(t) = \dfrac{1}{1+(\rho t)^\alpha}$, log-normal $S(t) = 1 - \Phi(\log((\rho t)^\alpha))$.

   (a) Describe the shape of the corresponding hazard functions (increasing/decreasing with time? etc)

   (b) To assess informally whether one of these models fits data, we can inspect a plot. To make it most easy to assess visually, we would construct a plot under which the model in question would correspond to a straight-line fit.
   Show that for Weibull, it would be appropriate to plot $\log(-\log \hat{S}(t))$ against $\log t$. Find corresponding plots for the log-logistic and log-normal cases.

   (c) In an AL regression model, individual j has a rate parameter ρ_j, given by $\rho_j = \exp(\beta.x_j)$ where x_j is the vector of covariates for that individual. Let T_j be the lifetime of individual j. Show that in the Weibull case, we have the relation

$$\log T_j \stackrel{d}{=} -\log \rho_j + \frac{1}{\alpha} Y \qquad (*)$$

   where Y has the *extreme-value distribution* $S_Y(y) = \exp(-e^y)$, (independent of α or ρ_j).
   Show that (*) also holds for the log-logistic and log-normal cases, if Y is appropriately distributed.

   (d) Suppose we believe that the lifetimes of a population are well modelled by a Weibull distribution. Explain how we could then test whether the special case of an exponential distribution (i.e. the case $\alpha = 1$) is an appropriate model?

5. (a) Suppose we wish to compare the survival functions of two groups. What graphs would be appropriate for consideration of a proportional hazards model? for an accelerated life model respectively? for both together?

   (b) Show that a Weibull family with a given fixed value of α has both the PH property and the AL property. Explain this in the light of 4(b) and 5(a).

   (c) Gehan (1965) studied 42 leukaemia patients. Some were treated with the drug *6-mercaptopurine* and the rest are controls. (The data are included under the name `gehan` in the `R` package `MASS`.) The observed times to recurrence (in months) were:

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

   Here + indicates censored times. Investigate these data along the lines of 4(b), 4(d), 5(a).

6. (a) In the context of the PH model, what is meant by the *partial likelihood* and how this can be used to estimate regression coefficients. How might standard errors be generated? (The framework we considered in lectures is known as "Cox regression").

   (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.

7. Suppose that $y_1, \ldots, y_n$ are observations from a lifetime distribution with respective vectors of covariates $x_1, \ldots, x_n$. It is thought appropriate to study the data y using an AL model based on the Weibull distribution with parameters ρ, α, with the link function $\rho = \exp(\beta.x)$. 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 α could be found. What would be the asymptotic distribution of the vector of estimators? How would the likelihood differ if some of the observations y_i were right censored (assuming independent censoring)?

8. Coronary Heart Disease (CHD) is a leading cause of death in many countries. The evidence is substantial that males are at higher risk than females, but the role of gender versus other genetic factors 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	50	49*	56*	68	74*	69*	70*	67	74*	81*	61	75*
Age female twin	63*	52	70*	75	72	69*	70*	70	74*	81*	58	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, calculate 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 an independent censoring mechanism is appropriate? Give reasons.