A.2 Life expectancy, graduation, and survival analysis

Questions 1–3 are to be done for discussion in class. Questions 4–7 are to be handed in for marking.

1. (a) Explain what is meant by right censoring, left censoring, right truncation, left truncation.

   (b) In a study of the elderly, individuals were enrolled in the study, at varying times, if they had already had one episode of depression. The event of interest was the onset of a second episode. An individual could be enrolled if at some previous time an episode of depression had been diagnosed. Which of the above mechanisms are relevant if it is also known that the study finished after four years?

   (c) In 1988 a study was published of the incubation time (waiting time from infection until symptoms develop) of AIDS. The sample was of 258 adults who were known to have contracted AIDS from blood transfusion. The data reported were the date of the transfusion, and the time from infection until the disease was diagnosed. Which of the above mechanisms are relevant for analysing these data?

2. (a) Suppose you are given estimates for a population of remaining life expectancy e_x and e_{x+t}, corresponding to ages x and $x+t$ (years). You wish to compute the mortality probability ${}_tq_x$. Under the assumption that mortality rates are constant over this interval, show that

   $$ {}_tq_x \approx \frac{t + e_{x+t} - e_x}{t/2 + e_{x+t}}. \qquad (*) $$

   Explain why this equation is only approximate, and what assumption would make it a good approximation.

   (b) The following is an estimated table of e_x (in years) in ancient Rome, as computed by Tim Parkin *Demography and Roman Society*, available at `http://www.utexas.edu/depts/classics/documents/Life.html`.

x	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70
e_x	25	33	43	41	37	34	32	29	26	23	20	17	14	10	8	6

   Assuming these estimates to be correct, and assuming the mortality rates to be constant over the age intervals, use equation (*) to estimate the annual mortality rates ${}_1q_x$ over the age intervals 0–1, 1–5, 6–10.

3. The data set `ovarian`, included in the `survival` package, presents data for 26 ovarian cancer patients, receiving one of two treatments, which we will refer to as the *single* and *double* treatments. (They appear in the data set as the `rx` variable, taking on values 1 and 2 respectively.)

   (a) Create a survival object for the times in this database.

   (b) Compute and plot the Kaplan–Meier estimator for the survival curves. (For a small extra challenge, plot the single-treatment survival curve black, and the double-treatment curve red.) You may use the `survfit` function.

   (c) Compute the Nelson–Aalen survival curve estimate. Make a table of the relevant data (time of events, number of events, number at risk).

(d) Compute the standard error for the probability of survival past 400 days in each group, as estimated by the Nelson–Aalen and Kaplan–Meier estimators.

4. The following is an investigation carried out by a (medium-sized) UK pension scheme into the mortality of its pensioners between 2000-2002.

   (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	$\mu_{x+0.5}$	$\overset{\circ}{\mu}_{x+0.5}$	z_x
60–64	1388.9	10	0.0072	0.0061	0.5249
65–69	1188.8	17	0.0143	0.0131	0.3615
70–74	880.5	28	0.0318	0.0262	1.0266
75–79	841.6	34	0.0404	0.0487	-1.0912
80–84	402.8	41	0.1018	0.0839	1.2394
85–89	123.9	19	0.1533	0.1338	0.5949
90–94	27.9	7	0.2509	0.1975	0.6346
95–99	10.0	3	0.3000	0.2706	0.1787
100+	7.5	2	0.2666	0.3455	-0.3673

   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.

5. Attached is an excerpt from a cohort life table for men in England and Wales born in 1894, including curtate life expectancies. (Data from the Human Mortality Database.) Using the given data:

   (a) Estimate the change to e_0, the curtate life expectancy at birth, if the mortality rate in the first two years of life were reduced to modern-day levels (say $q_0 = 0.005$, $q_1 = 0.0004$).

   (b) Make a rough estimate of the change to e_0 if the increases in mortality due to the 1914-18 war and the 1918-19 influenza pandemic had not occurred.

Age x	l_x	q_x	e_x
0	100000	0.16134	44.82
1	83866	0.05398	52.39
⋮	⋮	⋮	⋮
14	74067	0.00220	45.99
15	73904	0.00237	45.09
16	73729	0.00260	44.20
17	73538	0.00301	43.31
18	73316	0.00313	42.44
19	73087	0.00787	41.57
20	72512	0.01836	40.90
21	71181	0.03218	40.65
22	68890	0.04424	40.98
23	65842	0.06194	41.86
24	61764	0.02088	43.59
25	60474	0.00551	43.51
26	60141	0.00385	42.75
27	59910	0.00384	41.91
28	59680	0.00391	41.07
29	59446	0.00377	40.23
30	59222	0.00386	39.38
31	58994	0.00367	38.53
32	58777	0.00380	37.67
33	58554	0.00399	36.81
34	58320	0.00445	35.96
35	58061	0.00460	35.11

6. If x is the observed value of a random variable $X \sim \text{Binom}(n, p)$, with known n, find the maximum-likelihood estimator $\hat{p}$, and deduce that

$$\text{Var}(\hat{p}) \approx \frac{x(n-x)}{n^3}.$$

If $\hat{S}(t)$ is the Kaplan-Meier estimator, an alternative estimator for the variance is

$$\text{Var}\big(\hat{S}(t)\big) = \frac{\hat{S}(t)^2(1-\hat{S}(t))}{n(t)}$$

where $n(t)$ is the number at risk at time $t+$. If $d(t)$ is the number of failures up to and including time t, justify the estimation

$$\hat{S}(t) \approx \frac{n(t)}{n(t)+d(t)} = \frac{n(t)}{n(0)},$$

making the conservative assumption that all the censoring in the interval $[0, t)$ takes place at $t = 0$. What is the distribution of $d(t)$ given this assumption? Explain how this can be used to justify the expression for $\text{Var}\,\hat{S}(t)$ in terms of a binomial proportion estimator (as $\hat{p}$ above). In the special case of no censoring, what is the connection between this estimator and Greenwood's estimator for the variance?

7. We are carrying out a hypothetical study of the survival of Alzheimer patients. We enrol 30 subjects in a clinic, and follow them over five years. We record their age at being enrolled in the study and the age at which they left, and the cause of exit, whether death (1) or something else (0).

Entry Age	Exit Age	Death Indicator	Entry Age	Exit Age	Death Indicator
67	72	0	69	74	1
70	71	0	69	71	0
70	73	1	66	68	0
65	70	0	73	76	1
65	68	1	67	68	0
73	78	1	66	70	1
69	74	1	69	73	1
76	78	1	66	70	1
66	67	0	78	81	1
72	76	1	66	70	1
65	70	1	68	73	1
71	75	1	70	74	1
69	71	0	66	68	0
71	74	1	89	92	1
68	73	0	68	72	1

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

(b) Make a table indicating the number of subjects at risk at ages from 65 to 75.

(c) Estimate the survival curve over this age range.

(d) Compute a 95% confidence interval for the survival probability from age 70 to 75.

(e) Enter the data into `R` and use the `survival` package to estimate and plot the survival curve.