A.2 Life expectancy, multiple decrements model, and introduction to survival analysis

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

$$ {}_tq_x \approx \frac{t + e_{x+t} - e_x}{(t-1)/2 + e_{x+t}}. \qquad (*) $$

   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.

   (c) Suppose we know that 20% of Roman infants died of dysentery in their first year. Under the competing risks assumption, estimate the change in life expectancy at birth that would have resulted from a cure for this illness.

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. 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?

5. (From [29, Chapter 8]) Suppose that in some population ${}_5q_{90}^{Cancer*} = 0.097$ and ${}_5q_{90}^{Other*} = 0.132$. What is the overall ${}_5q_{90}$, and what is the fraction of deaths in this age range due to cancer? What assumptions are required?

6. 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 the age interval 70 to 75. Think of this as a survival-analysis problem and as a life-table problem. How does this change your approach? How do the two methods for estimating survival curves relate to the two methods for estimating life-table parameters?

(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.