A.3 Graduation, Markov models, basic survival analysis

1. 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.

2. (a) A life office uses the three-state healthy-sick-dead model in the pricing of its long term sickness policies. The transition rates are assumed to be constant. Denote the state space by $\mathbb{S} = \{H, S, \Delta\}$ and transition rates by $\sigma = q_{HS}$, $\rho = q_{SH}$, $\mu = q_{H\Delta}$, $\nu = q_{S\Delta}$.

   For a group of policy holders, over a one-year period the following data were recorded:

transition from	number
H to S	15
H to Δ	6
S to H	5
S to Δ	1

   The total times spent in states H and S were 625 years and 35 years, respectively.

   i. Write down the likelihood function for this model and show that this is maximized when $\sigma = 0.024$.

   ii. Determine the asymptotic distribution of $\hat{\sigma}$, the MLE of σ.

   iii. Calculate an estimate of the standard deviation of $\hat{\sigma}$.

   iv. Construct an approximate 95% confidence interval for σ.

   v. The policyholder pays contributions at rate C when in state H and receives benefits at rate B when in state S. No death benefit is payable. The life office uses the model to set the ratio of contributions to benefits. Briefly explain how this can be done.

vi. A trainee believes that the model is too simplistic. For each of the trainee's suggestions below, comment on whether following the suggestion would be likely to improve the model's predictive power.

A. The transition rates should depend on the age of the policyholder.

B. The transition rates should vary according to the time of year.

C. ρ and ν should also depend on the duration of the sickness to date.

Discuss briefly which suggestions you could carry out, in principle, with techniques you have learned in the lectures. What additional data do you need?
Outline the principal difficulty in fitting a model with parameters dependent on all these factors.

3. Consider a single-server queueing system.

(a) Denote the arrival rate by λ, the service rate by μ. Starting from an empty system and given observations up to the nth transition,

i. write down the likelihood function and determine the maximum likelihood estimator $(\hat{\lambda}, \hat{\mu})$;

ii. calculate the bivariate asymptotic distribution of $(\hat{\lambda}, \hat{\mu})$. Deduce that $\hat{\lambda}$ and $\hat{\mu}$ are asymptotically independent.

iii. Derive separate approximate $(1-\alpha)$-CIs for λ and μ, and joint approximate $(1-\alpha)$-confidence regions, either rectangular or with minimal area.**[Optional]**

(b) Suppose that the queue length cannot increase beyond m and the length of the queue has an impact on the arrival rates (but not on the service times).

i. How do you model this situation?

ii. Derive maximum likelihood estimators.

iii. Suppose $m = 2$. If $\lambda_0 = \lambda_1$, what is the large-sample distribution of

$$\frac{\hat{\lambda}_0 - \hat{\lambda}_1}{\sqrt{\mathrm{Var}(\hat{\lambda}_0) + \mathrm{Var}(\hat{\lambda}_1)}} \quad \text{or} \quad \frac{(\hat{\lambda}_0 - \hat{\lambda}_1)^2}{\mathrm{Var}(\hat{\lambda}_0) + \mathrm{Var}(\hat{\lambda}_1)}? \tag{A.5}$$

iv. By estimating the variance, deduce a test for $H_0 : \lambda_0 = \lambda_1$ vs $H_1 : \lambda_0 \neq \lambda_1$. **[Optional]**

v. Generalise iii. and iv. to $m \geq 3$. **[Optional]**

4. (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?

5. 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?

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 this age range.

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

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