B.3 Graduation, Markov models, basic survival analysis

1. (a) Crude estimates from the data are subject to stochastic fluctuation. Smoothing (graduating) the estimates may make more reliable predictions.

   (b) $\mu_x = a + be^{\alpha x}$ for Gompertz–Makeham. This is generally considered a reasonable model for the hazard rate (force of mortality) from middle age onward. Note, though, that the mortality rate doubling times (which would be approximately constant under Gompertz–Makeham) lengthen progressively. The parameters a, b, α will have to be fitted from the data.

   We apply the chi-squared test. To begin with, we combine the last two rows to have ≥ 5 expected deaths in each row. The last row becomes

   $$99 \qquad 17.5 \qquad 5 \qquad 0.2857 \qquad 0.3027 \qquad -0.1293$$

   (We interpolate by weighting the two rows by their central exposed to risk.) The χ^2 statistic is then 4.96 on 8 observations. Since we have estimated 3 parameters, we compare this to the table with 5 degrees of freedom, obtaining p-value 0.42.

   To test for bias we use the cumulative deviations test, obtaining $Z = 0.96$, and a p-value of 0.3375. Thus, the model seems to fit. Notice that graduated hazard is generally lower — it is strongly affected by the mortality plateau a very late ages — which would lead to an overestimate of benefits paid. This is a relatively good error to make, though it would be reversed if the company were selling life insurance!

2. (a) We are not given the size n of the group. However, this is not essential since the factorised form of the likelihood does not depend on n. We obtain

   $$\prod_{i\in\mathbb{S}}\prod_{j\neq i} q_{ij}^{N_{ij}} \exp\{-q_{ij}E_i\} = \sigma^{N_{HS}}e^{-\sigma E_H}\mu^{N_{H\Delta}}e^{-\mu E_H}\rho^{N_{SH}}e^{-\rho E_S}\nu^{N_{S\Delta}}e^{-\nu E_S}. \tag{B.11}$$

   This can be maximised parameter by parameter. To maximise in σ, we maximise $\sigma^{N_{HS}}e^{-\sigma E_H}$ or, passing to logs,

   $$\ell(\sigma) = N_{HS}\log(\sigma) - \sigma E_H \Rightarrow \ell'(\sigma) = \frac{N_{HS}}{\sigma} - E_H \Rightarrow \ell''(\sigma) = -\frac{N_{HS}}{\sigma^2} < 0 \tag{B.12}$$

   and ℓ is maximized for $\hat{\sigma} = N_{HS}/E_H$. For $N_{HS} = 15$ and $E_H = 625$ this is $\hat{\sigma} = 15/625 = 0.024$.

   (b) For the asymptotic distribution, we require the Fisher Information. The likelihood factorises, so the log likelihood is the sum of functions of single parameters, so the Fisher Information matrix is diagonal, and we calculate, approximating the Fisher Information by the observed information and its estimate

   $$I_{\sigma\sigma} = -\mathbb{E}(\ell''(\sigma)) \approx \frac{N_{HS}}{\sigma^2} \approx \frac{E_H^2}{N_{HS}}. \tag{B.13}$$

   From the asymptotic theory, $\hat{\sigma} \sim \mathcal{N}(\sigma, N_{HS}/E_H^2)$.

   (c) In particular, $\sqrt{N_{HS}}/E_H = \sqrt{15}/625 = 0.0062$ is an estimate of the standard deviation of $\hat{\sigma}$.

   (d) Then $[\hat{\sigma} - 1.96\sqrt{N_{HS}}/E_i, \hat{\sigma} + 1.96\sqrt{N_{HS}}/E_i] = [0.012, 0.036]$ is an approximate 95% confidence interval.

   (e) For solvency the company requires aggregate contributions to be greater than or equal to aggregate benefits, i.e. here $CE_H \geq BE_S$, so $C/B \geq E_S/E_H = 35/625$. Second order considerations (variances) can be used to give confidence intervals around such a value. We neglect here administrative expenses and further risk considerations.

   (f) Particularly, if the age range is wide, suggestion i. can improve the predictive power. ii. is likely to be of minor importance since the effects will average out over the year anyway. iii. must be expected to have impact on the variances since long-term illnesses cause a lot of benefit payments from only few individuals.

Markov models with age-dependent transition rates such as i. were done in the lectures. For ii., the techniques can be easily adapted if appropriate data are recorded (birthdays or numbers at risk every season, say, and transitions and waiting times every season). For iii., a bit more work is required to adapt the methods, since the Markovian character is lost.

We probably have about 650 subjects in the group. With a total of 27 transitions, this is not anywhere near enough to fit a refined model, neither to distinguish ages nor, say, the four seasons to discretize the time of the year or rates changing with duration of illness in months.

3. (a) i. Since only transitions from i to $i+1$ and $i-1$ are possible, the likelihood can be written as

$$\begin{aligned}&\lambda^{N_{01}}\exp\{-\lambda E_0\}\prod_{i=1}^{\infty}\lambda^{N_{i,i+1}}\mu^{N_{i,i-1}}\exp\{-(\lambda+\mu)E_i\}\\&=\lambda^{N_+}\exp\{-\lambda E_+\}\mu^{N_-}\exp\{-\mu E_-\},\end{aligned}$$

where

$$N_+=\sum_{i=0}^{\infty}N_{i,i+1},\quad N_-=\sum_{i=1}^{\infty}N_{i,i-1},\quad E_+=\sum_{i=0}^{\infty}E_i,\quad E_-=\sum_{i=1}^{\infty}E_i.$$

From the form of the likelihood we easily compute $\hat\lambda=N_+/E_+$ and $\hat\mu=N_-/E_-$.

ii. We use the observed information matrix $\tilde I$ to estimate the Fisher information matrix I. It is diagonal, since the likelihood fully factorises. We obtain

$$\tilde I_{\lambda\lambda}=\frac{N_+}{\lambda^2},\qquad \tilde I_{\mu\mu}=\frac{N_-}{\mu^2}. \tag{B.14}$$

By the asymptotic theory of maximum likelihood estimators, $(\hat\lambda,\hat\mu)\sim\mathcal{N}((\lambda,\mu),\tilde I^{-1})$, approximately, for large n. Since Normally distributed random variables are independent if and only if they are uncorrelated, we deduce that $\hat\lambda$ and $\hat\mu$ are asymptotically independent.

iii. Asymptotically, $\lambda-\hat\lambda$ is normal with mean 0 and variance $\sigma_1^2:=\hat\lambda^2/N_+$ and $\mu-\hat\mu$ is normal with mean 0 and variance $\sigma_2^2:=\hat\mu^2/N_-$. Thus, a (1-$\alpha$)-CI for λ is $\hat\lambda\pm u_{\alpha/2}\sigma_1$, and a 95% CI for μ is $\hat\mu\pm u_{\alpha/2}\sigma_2$.

We can represent these errors as approximately $\sigma_1Z_1+\sigma_2Z_2$, where Z_1 and Z_2 are independent standard normal random variables. The curves of constant probability are thus ellipsoids, and the approximate (1-α)-confidence region of minimal area will be

$$\left\{(\lambda,\mu):\frac{(\lambda-\hat\lambda)^2}{\sigma_1^2}+\frac{(\mu-\hat\mu)^2}{\sigma_2^2}\le c_\alpha\right\},$$

where c_α is the $1-\alpha$ quantile of a χ^2 distribution with 2 degrees of freedom.

(b) i. The state space is $\mathbb{S}=\{0,\dots,m\}$. We allow tridiagonal Q-matrices with parameters $q_{i,i+1}=\lambda_i$ and $q_{i+1,i}=\mu$, $i=0,\dots,m-1$.

ii. The likelihood function is now

$$\begin{aligned}&\lambda_0^{N_{01}}\exp\{-\lambda E_0\}\mu^{N_{m,m-1}}\exp\{-\mu E_m\}\prod_{i=1}^{m-1}\lambda_i^{N_{i,i+1}}\mu^{N_{i,i-1}}\exp\{-(\lambda_i+\mu)E_i\}\\&=\left(\prod_{i=0}^{m-1}\lambda_i^{N_{i,i+1}}\exp\{-\lambda_iE_i\}\right)\mu^{N_-}\exp\{-\mu E_-\}.\end{aligned}$$

Therefore, the maximum likelihood estimators are $\hat\lambda_i=N_{i,i+1}/E_i$, $i=0,\dots,m-1$, and $\hat\mu=N_-/E_-$.

iii. As in (a), the asymptotic distribution of the maximum likelihood estimators is multivariate Normal with diagonal variance-covariance matrix. In particular, for $m = 2$,

$$\frac{\hat\lambda_0 - \hat\lambda_1}{\sqrt{\mathrm{Var}(\hat\lambda_0) + \mathrm{Var}(\hat\lambda_1)}} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad \frac{(\hat\lambda_0 - \hat\lambda_1)^2}{\mathrm{Var}(\hat\lambda_0) + \mathrm{Var}(\hat\lambda_1)} \sim \chi^2_1.$$

iv. The problem is to compute the variance of N_i/E_i. It is reasonable to treat E_i as fixed, and simply compute the conditional variance. There are several ways to approach this. We might first say that the conditional distribution of N_i, conditioned on a particular value of E_i, is Poisson with parameter λE_i, so the conditional variance of N_i is also $\lambda_i E_i$, and the conditional variance of $\hat\lambda$ is λ_i/E_i. This is not exactly true, since we have also assumed that there are exactly n transitions, so $N_- + N_0 + N_1 = n$. This means that the transitions during a period of time E_i do not really have a Poisson distribution. How much does this matter?
The answer is, not much, asymptotically. We would expect that, since the dependence between two transitions will be very small when the total number is very large. The correct calculation is as follows: Consider just $\hat\lambda_0$, and assume n is even. Conditioned on N_0, there is a gamma distribution for E_i, with parameters N_0 and λ_1. We have then

$$\begin{aligned}
\mathbb{E}\big[(E_i)^{-1}\big] &= \frac{\lambda_0}{N_0 - 1} \\
\mathbb{E}\big[(E_i)^{-2}\big] &= \frac{(\lambda_0)^2}{(N_0-1)(N_0-2)} \\
Var(E_i^{-1}) &= \frac{(\lambda_0)^2}{(N_0-1)^2(N_0-2)} \approx (N_0)^{-3}(\lambda_0)^2.
\end{aligned}$$

Thus, conditioned on N_0, we have $Var(\hat\lambda_0) \approx \lambda_0^2/N_0$. Since $N_0/E_0 \to \lambda_0$, this is very close to λ_0/E_0 for large n, which is the result we had before.

v. What is the generalised question? One reasonable alternative is that we are testing the null hypothesis that all the λ_i are equal to the same (unknown) value λ. Asymptotically, under the null hypothesis, $\hat\lambda_i$ will be approximately independent normally distributed with mean λ and variance N_i/E_i. Our joint estimate for λ is then

$$\begin{aligned}
\hat\lambda_* :&= \Big(\sum_{i=0}^{m-1} \frac{1}{\mathrm{Var}(\hat\lambda_i)}\Big)^{-1} \sum_{i=0}^{m-1} \frac{\hat\lambda_i}{\mathrm{Var}(\hat\lambda_i)} \\
&= \Big(\sum_{i=0}^{m-1} \frac{E_i^2}{N_i}\Big)^{-1} \sum_{i=0}^{m-1} E_i.
\end{aligned}$$

Under the null hypothesis, then,

$$\tilde\lambda_j := \left(\frac{N_j}{E_j^2} + \left(\sum \frac{E_i^2}{N_i}\right)^{-1}\right)^{-1/2} \Big(\hat\lambda_j - \hat\lambda_*\Big)$$

are approximately independent standard normal, except that they are subject to a single linear condition. Thus $X = \sum_{j=0}^{m-1} \tilde\lambda_j^2$ should be approximately chi-squared distributed with $m-1$ degrees of freedom.

4. (a) See lecture notes.

   (b) There is right censoring: The depression may not have recurred at the time that the study ended, or the patient died or dropped out. There is left truncation: The first episode of depression made the patients eligible for the study, but not immediately. Thus, the event of interest — the recurrence of depression — could already have happened before the patient was enrolled in the study.

(c) This study design involves right truncation: The entire study population has already experienced the event of interest (AIDS diagnosis). Any individual whose incubation period extended beyond the truncation time would not have appeared in the study.

5. The log likelihood is

$$\ell(p) = \log \binom{n}{x} + x \log p + (n-x)\log(1-p).$$

This has solution $0 = \ell'(\hat{p}) = x/\hat{p} - (n-x)/(1-\hat{p})$, implying $\hat{p} = x/n$. We know that the variance of a binomial random variable is $np(1-p)$. Substituting $\hat{p}$ for p yields the estimate

$$\operatorname{Var}(\hat{p}) = \operatorname{Var}(x/n) = n^{-2}\operatorname{Var}(x) = n^{-1}p(1-p) = n^{-1}\frac{x}{n}\frac{n-x}{n} = \frac{x(n-x)}{n^3}.$$

If all the censoring occurs at $t = 0$ then the number of individuals at risk of dying in $(0, t)$ is actually $n(t) + d(t)$. Thus alive at time t is binomial with parameters $n = n(0) = n(t) + d(t)$ and $p = S(t)$. The MLE for p is thus

$$\hat{S}(t) = \hat{p} = \frac{n(t)}{n(t) + d(t)} = \frac{n(t)}{n(0)}.$$

(If the censoring all happens at time 0, then the number at risk at time 0+ will be the same as the sum of the number who die up to time t, and the number still at risk at time t.) The variance estimate is

$$\frac{d(t)n(t)}{n(0)^3} = n(t)^{-1}\frac{d(t)}{n(0)}\frac{n(t)}{n(0)}\frac{n(t)}{n(0)} = n(t)^{-1}(1-\hat{S}(t))\hat{S}(t)^2.$$

Greenwood's estimate in the case of no censoring is

$$\begin{aligned}
\operatorname{Var}\hat{S}(t) &\approx \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)} \\
&= \hat{S}(t)^2 \sum_{t_j \le t} \frac{n_{j+1} - n_j}{n_j(n_{j+1})} \\
&= \hat{S}(t)^2 \sum_{t_j \le t} \left(\frac{1}{n_{j+1}} - \frac{1}{n_j}\right) \\
&= \hat{S}(t)^2 \left(\frac{1}{n(t)} - \frac{1}{n(0)}\right) \\
&= \hat{S}(t)^2 \frac{n(t) - n(0)}{n(t)n(0)} \\
&= n(t)^{-1}\hat{S}(t)^2(1 - \hat{S}(t))
\end{aligned}$$

as before.

6. (a) Right censoring and left truncation.

(b) If individuals who enter at age x are considered immediately available to count at risk at age x, and those who die at age x are also at risk.

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	3	9	11	13	14	17	14	12	12	8	4

We are planning to use the actuarial estimator — so we count those who are censored or died as having had half a year at risk, and count those who entered at a given age as having half a year at risk in that year, we get the following counts:

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	1.5	6.0	9.5	9.5	11.5	13.0	11.5	10.5	9.0	6.0	3.5

(c) Again, counting whole years at risk for those who enter, die, or are right-censored, we have

Age	n_i	d_i	h_i	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
70	17	4	0.235	0.765	0.790
72	12	1	0.083	0.701	0.727
73	12	3	0.250	0.526	0.566
74	8	4	0.500	0.263	0.343
75	4	1	0.250	0.197	0.268

The actuarial estimate gives us

Age	n_i	d_i	h_i	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
70	13.0	4	0.308	0.692	0.735
72	10.5	1	0.095	0.626	0.668
73	9.0	3	0.333	0.418	0.479
74	6.0	4	0.667	0.139	0.246
75	3.5	1	0.286	0.099	0.185

Note that we might reasonably suggest that age is not a sensible time variable here, since mortality is largely determined by time since diagnosis. We see that the estimator of survival past age 78 is 0, since the single individual who happened to be in the study at that age died. This despite the fact that there are other individuals who entered later and survived to much older ages. We might reasonably look instead at the *time-on-test* as time variable. We would then get the following calculation:

t_j	n_j	d_j	h_j	$\widehat{S}(t_j)$	$\widetilde{S}(t_j)$
2	27	1	0.04	0.96	0.96
3	22	6	0.27	0.70	0.73
4	16	8	0.50	0.35	0.44
5	8	5	0.62	0.13	0.24

(d) We use the whole-year method, rather than the actuarial estimate. Our central estimate for the probability of surviving from age 70 to age 75 is $\hat{S}(74) = 0.343$. Using Greenwood's estimate, we estimate the variance of $\log \hat{S}(74)$ to be

$$\sum_{t_i \leq 74} \frac{d_i}{n_i(n_i - d_i)} = \frac{4}{17 \cdot 13} + \frac{1}{12 \cdot 11} + \frac{3}{12 \cdot 9} + \frac{4}{8 \cdot 4}$$
$$= 0.178,$$

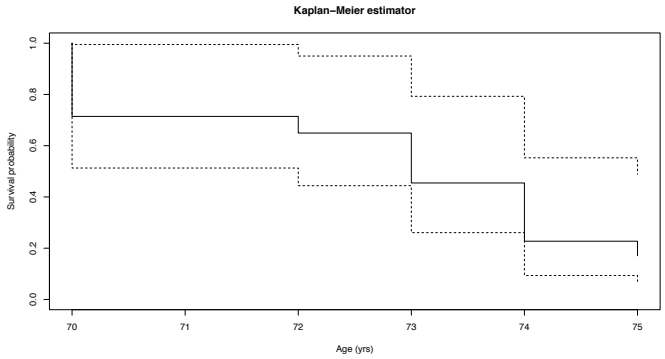
so the standard error is $\sqrt{0.178} = 0.422$. Thus an approximate 95% confidence interval for $S(74)$ is

$$\left(0.343e^{-0.422 \cdot 1.96}, 0.343e^{0.422 \cdot 1.96}\right) = (0.150, 0.784).$$

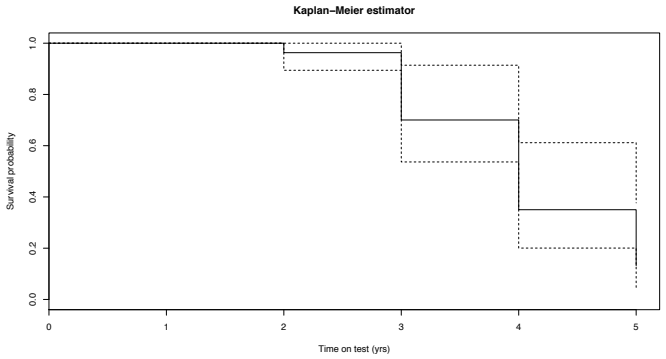
(e)

```
require('survival')
age.entry=c(67,70,70,65,65,73,69,76,66,72,65,71,69,71,68,69,69,66,
  73,67,66,69,66,78,66,68,70,66,89,68)
age.exit=c
    (72,71,73,70,68,78,74,78,67,76,70,75,71,74,73,74,71,68,76,68,70,73,
70,81,70,73,74,68,92,72)
delta=c(0,0,1,0,1,1,1,1,0,1,1,1,0,1,0,1,0,0,1,0,1,1,1,1,1,1,1,0,1,1)

clinic.surv=Surv(time=age.entry,time2=age.exit,event=delta) # left-
    truncated, right-censored is default
```

```
KM.fit=survfit(clinic.surv~1,subset=(age.exit>=70)) # Survival of those present after age 70
plot(KM.fit , firstx=70,xmax=75,ylab='Survival probability',main='Kaplan-Meier estimator',xlab='Age (yrs)')

TOT.surv=Surv(time=time.on.test,event=delta)
TOT.fit=survfit(TOT.surv~1)
plot(TOT.fit,ylab='Survival probability',main='Kaplan-Meier estimator',xlab='Time on test (yrs)')
```



(a) Survival by age



(b) Survival by time on test