MTH 6157 Solutions

Q1 Kaplan Meier

(a)

at time t_j minutes

d_j = nº of customers paying for shopping

c_j = nº customers leaving queue without paying

n_j = risk set, those still in queue

λ_j = hazard = d_j / n_j

K-M estimate of survival function is $\hat{S}(t)$ where

$$\hat{S}(t) = \prod_{t_j \le t} (1 - \lambda_j)$$

for checkout counter

t	n	d	c	λ	$1-\lambda$	$\hat{S}(t)$
0	16					
1	16	3		0.1875	0.8125	0.813
2	13	3	1	0.2308	0.7692	0.625
3	9	2	1	0.2222	0.7778	0.486
4	6	1		0.1667	0.8333	0.405
5	5	1		0.2	0.8	0.324
6	4	1	1	0.25	0.75	0.243
7	2	0	1	0	1	0.243
8	1	1		1	0	0
9	0					

for self-service machine

t	n	d	c	λ	$1-\lambda$	$\hat{S}(t)$
0	16					
1	16	0		0	1	1
2	16	2		0.125	0.875	0.875
3	14	3		0.2143	0.7857	0.688
4	11	3		0.2727	0.7273	0.500
5	8	2		0.25	0.75	0.375
6	6	2		0.3333	0.6667	0.250
7	4	2	1	0.5	0.5	0.125
8	1	1		1	0	0
9	0					

(b) The survival function is probability of still being in the queue at t minutes, therefore the student will choose the queue with smaller S(t) values

t	S(checkout)	S(machine)
1	0.813	1
2	0.625	0.875
3	0.486	0.688
4	0.405	0.500
5	0.324	0.375
6	0.243	0.250
7	0.243	0.125
8	0	0

$\hat{S}(t)$ is smaller at all t except $t=7$ for the checkout $\therefore$ the student will select the checkout over the machine based on these K-M estimates.

(c) There is much more censoring in the checkout data (4 of 16 observations) than the machine data (1 of 16). There is the possibility that censored people in one queue leave to join the other queue. This would be "informative" right censoring and would indicate dependence between the checkout and machine times.

Cox Proportional Hazard

Q2

(a) The baseline hazard applies to lives with all $z_i = 0$
This is a patient in Los Angeles aged under 65 given antibiotics immediately on admission to hospital

(b) (i) $z_1 = 0$
$z_2 = 0.5$
$z_3 = 1$

$\therefore$ hazard function is

$$h(t) = h_0(t)\ \exp(0 \times 0.1 + 0.5 \times 0.02 - 1 \times 0.3)$$
$$= h_0(t)\ \exp(0.29)$$

(ii) $$S(t) = \exp - \int_0^t h_0(t)\, e^{0.29}\, dt$$

(c) for life 44NY

$$S(7) = 0.88$$

$$= \exp - \int_0^7 h_0(t) \exp(1 \times 0.1 + 2 \times 0.02 - 0 \times 0.3)$$

$$= \exp - \int_0^7 h_0(t)\, e^{0.14}\, dt \qquad = 0.88$$

$$\therefore - \int_0^7 h_0(t)\, dt = \frac{\log 0.88}{e^{0.14}}$$

We are looking for $S_{72LA}(7)$

from (b) (ii)

$$S_{72LA}(7) = \exp - \int_0^7 h_0(t) \, e^{0.29} \, dt$$

$$= \exp\left(\frac{\log 0.88 \;\; e^{0.29}}{e^{0.14}} \right)$$

$$= \underline{\underline{0.862}}$$

Graduation Tests

Q3

(a) Standard table at least 10 years old so mortality improvements at oldest ages would be form of time selection

(b) The null hypothesis is the standard table remains a good representation of mortality experience. We calculate the chi-squared statistic $\sum_{ages} Z_x^2$ and compare to

$\chi^2_{0.95,10}$ with 10 degrees of freedom because we have 10 ages of data.

x	Z_x	Z_x^2
90	0.8	0.64
91	-0.4	0.16
92	-1.1	1.21
93	0.6	0.36
94	2	4
95	-0.3	0.09
96	-1	1
97	-0.6	0.36
98	-0.6	0.36
99	-0.8	0.64

$$8.82 < 18.3 = \chi^2_{0.95;10}$$

∴ we do not reject the null hypothesis at 95% level

This test suggests overall the standard table is a good representation of experienced mortality

(c) If the table overstated mortality more as age increases we would expect increasing number of negative Z_x values at higher ages. A "grouping of signs" or 'Stevens's' test could check for this at older ages.

(d) We hope that $Z_x \sim N(0,1)$ and the standardised deviations test checks this Normality assumption

Interval	Actual no of Z_x	Expected number	$(A-E)^2/E$
$(-3,-2)$	0	0.2	0.2
$(-2,-1)$	2	1.4	0.257
$(-1,0)$	5	3.4	0.753
$(0,1)$	2	3.4	0.576
$(1,2)$	0	1.4	1.4
$(2,3)$	1	0.2	3.2
			6.387

compare this to $\chi^2_{0.95;6} = 12.59$

$6.387 < 12.59$ $\therefore$ we do not reject the hypothesis

Again the standard table is found to be a good representation of the mortality experience

(e) Largest positive deviation (age 94) offsets larger number of negative deviations at higher ages

Exposed to Risk

Q4

(a) The principle of correspondence means a life alive at time t should be included in the Exposed to Risk calculation if and only if, were that life to die immediately, they would be counted in the deaths data at their age.

(b) Exposed to Risk = total number of person hours in the museum

for i^{th} person in the group $i = 1, 2, \ldots 20$

let

a_i = time in minutes through security
b_i = time to walk to exhibit
c_i = time waiting before exhibit
d_i = time in exhibit
e_i = time waiting after exhibit
f_i = time walking back at end

then exposed to risk $= \sum_{i=1}^{20} b_i + c_i + d_i + e_i + f_i$

$a_i = 2i$
$b_i = 4$ all i
$d_i = 25$
$e_i = 25$ for $i = 1, \ldots, 10$ $\quad e_i = 0$ for $i > 10$
$f_i = 4$ all i

10^{th} person has $c_{10} = 0$ (they go straight into exhibit)

c_i for $i = 1, \ldots 9$ depends on how long before 10^{th} person they clear security

$\therefore \quad c_9 = 2 \qquad c_8 = 4 \qquad c_7 = 6 \quad \ldots \quad c_1 = 18$

11^{th} person arrives at exhibit after $2 \times 11 + 4 = 26$ mins
first 10 leave exhibit after $2 + 4 + 18 + 25 = 49$ mins

$\therefore \quad c_{11} = 49 - 26 = 23$
$\quad c_{12} = 21 \qquad c_{13} = 19 \quad \ldots \quad c_{20} = 5$

adding $b_i \; c_i \; d_i \; e_i \; f_i$

Exposed to Rah $= 1140$ Minutes
$= 1140/60$ hours

Rate of coffee consumption per person hour

$$= \frac{6}{1140/60} = \underline{\underline{0.316}}$$

Generalized Linear Modelling

Q5 (a) Unless the age distribution is very different we would expect supermarket channel policies to exhibit higher rates of mortality than IFA policies as IFA clients will be wealthier than average supermarket customers and access to healthcare means wealth is negatively correlated with mortality.

(b) ① determine the distribution for the mortality data
② define the linear predictor
③ define a link function between the distribution and the linear predictor

(c) ① the distribution is given as Poisson

② the linear predictor has 3 covariates:-
- age
- male/female
- sales channel

age x is a variable which for life i enters the predictor in the form $V_i x$
Male/female and sales channel are factors not variables and will take values α_i and β_j respectively

As we have 2 factors we will need an interaction term γ_{ij}

$\therefore$ Linear predictor is given by $\gamma_i x + \alpha_i + \beta_j + \gamma_{ij}$

(3) we use the canonical link function for the Poisson distribution

$$g(\mu) = \log(\mu)$$

(d) We can test whether mortality is statistically different between the 2 sales channels by comparing the scaled deviance of a model with sales channel factor to one without. These are nested models. The model without will have 2 fewer parameters (β_j and γ_{ij}). Scaled deviance is approx chi-squared under the Poisson distribution $\therefore$ we compare the difference in scaled deviation with χ^2_2 (2 degrees of freedom)

A high chi-sq statistic indicates adding a sales channel factor does have significant effect on the model.