Australian National University

RESEARCH SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS

First Semester Final Examination (2014)

Survival Models / Biostatistics (STAT3032/7042/8003)

Writing period:	*3 hours duration*
Study period:	*15 minutes duration*
Permitted materials:	*Non-programmable calculator, dictionary, one A4 sized sheet of paper with notes on both sides*
Total marks:	*70 (undergraduates) / 78 (postgraduates) marks*

INSTRUCTIONS TO CANDIDATES:

- *Postgraduates should attempt all questions. Undergraduates should only attempt questions 1 to 7.*
- *To ensure full marks show all the steps in working out your solutions. Marks may be deducted for failure to show appropriate calculations or formulae.*
- *All questions are to be completed in the script book provided.*
- *All answers should be rounded to 4 decimal places.*

Question 1 [7 marks]

For a particular population it is shown that $l_x = 100 - x$, $0 \leq x \leq 100$. Using this information about the number of lives aged x exact, calculate the following:

(a) **[2 marks]** the force of mortality at age 50.

Solution:

$$\mu_x = -\frac{1}{l_x}\frac{dl_x}{dx} = \frac{1}{100-x}$$
$$\mu_{50} = \frac{1}{100-50} = 0.02$$

(b) **[2 marks]** the complete expectation of life at age 50.

Solution:

$$e^0_{50} = \int_0^{50} {}_tp_{50}dt = \int_0^{50} (1 - 0.02t)dt = 25$$

(c) **[3 marks]** the average age of individuals who die between ages 40 and 45.

Solution:

$$40 + \int_0^5 \frac{t l_{40+t}\mu_{40+t}}{l_{40} - l_{45}}dt = 40 + \int_0^5 \frac{t(100-40-t)(1/(100-40-t))}{5} = 42.5$$

Question 2 [8 marks]

(For each part, you will gain 2 marks for a correct answer, be penalized 1 mark for an incorrect answer, and score 0 if no answer is given.) Answer each question "TRUE" or "FALSE". In each case, write the whole word. It is **not** acceptable to write only "T" or "F" and answers presented in this form **will be graded incorrect**.

(a) **[2 marks]** The force of mortality function μ_x can be greater than 1.

(b) **[2 marks]** If the force of mortality function μ_x is assumed to follow the Makehams's law, this means that for each one year increment in age, μ_x increases by a constant additive amount.

(c) **[2 marks]** Increasing the bandwidth of a Kernel function will lead to a rougher smoothed curve.

(d) **[2 marks]** Simultaneously conducting n ($n > 1$) multiple independent hypothesis tests (all at α % significance level) will lead to a new significance level which is theoretically equal to α^n %.

Solution:

TRUE FALSE FALSE FALSE

Question 3 [9 marks]

In this question you will be looking at the breastfeeding data that was discussed in class. As a reminder, this dataset contains information on the duration of breastfeeding for over 900 mothers as well as information on a number of covariates. To investigate the effects of these covariates a Cox regression model was fitted. The following is output from a Cox regression analysis conducted in R:

```
                        coef exp(coef)  se(coef)      z Pr(>|z|)
as.factor(race)2     0.187823  1.206620  0.105469  1.781 0.074940 .
as.factor(race)3     0.296366  1.344962  0.097186  3.049 0.002293 **
as.factor(poverty)1 -0.226792  0.797087  0.094469 -2.401 0.016363 *
as.factor(smoke)1    0.246100  1.279027  0.079599  3.092 0.001990 **
as.factor(alcohol)1  0.167771  1.182666  0.123253  1.361 0.173454
agemth              -0.173193  0.840975  0.186168 -0.930 0.352212
I(agemth^2)          0.003615  1.003621  0.004251  0.850 0.395118
ybirth               0.079672  1.082932  0.020505  3.886 0.000102 ***
yschool             -0.056739  0.944841  0.023167 -2.449 0.014320 *
```

The variables *race*, *poverty*, *smoke*, and *alcohol* are categorical variables representing the race of the mother, whether the mother is living in poverty, whether the mother smokes, and whether the mother consumes alcohol. The variables *agemth*, *ybirth*, and *yschool* are continuous variables representing the age of the mother, the year of birth of the child and the numbers of years that the mother attended school. Based on the above output answer the following questions:

(a) **[2 marks]** Use critical value 2 to provide a 95% confidence interval for the multiplicative change in the hazard for a 1 year increase in the variable *yschool*, everything else held constant.

Solution:

$$exp(-0.056739 \pm 2 \times 0.023167) = (0.9021, 0.9896)$$

(b) **[3 marks]** Provide an estimate of the quantity $\frac{1}{\beta_9}$, where β_9 is the parameter corresponding to *yschool*. You must provide a standard error for your estimate.

Solution:

Our estimate is 1/(-0.056739)=-17.6246 and the standard error is $1/(0.056739^2) \times 0.023167 = 7.1963$

(c) **[2 marks]** After the above analysis had been conducted you are told that the values of duration were mistakenly reported using units of months rather than days. For example, a duration of 4 days was accidentally recorded as a duration of 4 months etc. How would this information change the parameter estimates reported above? You must provide a reason for your answer.

Solution:
Estimates would not change. All that matters is the order of the deaths.

(d) **[2 marks]** What would happen to the value of the partial likelihood if the coefficient estimate for the variable *yschool* was replaced by the value 0.05 (note all other estimates remain unchanged). You must provide a reason for your answer.

Solution:
The partial likelihood will decrease. The reason is that -0.056739 is the maximum likelihood estimate of β_9.

Question 4 [16 marks]

A set of crude mortality rates were graduated using a particular smoothing technique. **The smoothing technique that was used required 5 parameters to be estimated**. The table below provides details of the graduation:

Age	Crude Rate	Graduated Rate	Exposed to Risk	Deaths	Expected Deaths	Standardised Deviation	Squared Std. Dev.
20	0.028884	0.02287	1004	29	22.96	1.2748	1.6251
21	0.016260	0.02525	984	16	24.85	-1.7975	3.2310
22	0.040123	0.02778	972	39	27.00	2.3416	5.4831
23	0.018386	0.03049	979	18	29.85	-2.2027	4.8519
24	0.043750	0.03339	960	42	32.05	1.7867	3.1923
25	0.022293	0.03648	942	21	34.36	-2.3225	5.3940
26	0.052295	0.03978	937	49	37.27	1.9601	3.8420
27	0.031083	0.04332	933	29	40.42	-1.8361	3.3713
28	0.061159	0.04712	932	57	43.92	2.0226	4.0909
29	0.039474	0.05122	912	36	46.71	-1.6092	2.5895
30	0.060241	0.05566	913	55	50.82	0.6037	0.3645
31	0.048206	0.06047	892	43	53.94	-1.5367	2.3614
32	0.078409	0.06570	880	69	57.82	1.5217	2.3156
33	0.067045	0.07139	880	59	62.82	-0.5006	0.2506
34	0.087558	0.07759	868	76	67.35	1.0977	1.2049
35	0.077816	0.08434	861	67	72.62	-0.6888	0.4744
36	0.104094	0.09167	855	89	78.38	1.2589	1.5848
37	0.092216	0.09963	835	77	83.19	-0.7153	0.5117
38	0.112961	0.10824	841	95	91.03	0.4406	0.1941
39	0.102871	0.11752	836	86	98.25	-1.3153	1.7300
40	0.142327	0.12747	808	115	103.00	1.2663	1.6035
Total	**1.327451**	**1.31738**	**19024**	**1167**	**1158.61**	**1.0500**	**50.2670**

(a) **[3 marks]** Perform the chi-square test to check whether the graduated rates are appropriate. State the required steps discussed in the tutorial (the test statistic,

critical value and conclusion of the statistical test). Use a significance level of 5%.

Solution:
Test Statistic = 50.2672 which is chi-squared with 21-5=16 degrees of freedom. Critical Value at the 5 % significance level is 26.29. Reject H_0 and the graduates rates are inappropriate.

(b) **[3 marks]** Perform the standardized deviation test with four regions $(-\infty$,-1.5], (-1.5,0], (0,1.5] and (1.5, $\infty)$. State the required steps as in part (a) with a significance level of 5%.

Solution:
Since $P(Z < -1.5) = P(Z > 1.5) = 0.0668$, the expected numbers are 1.4028, 9.0972, 9.0972 and 1.4028. The observed numbers are 6,4,6,5. The test statistic is

$$\chi^2 = \frac{(6-1.4028)^2}{1.4028} + \frac{(4-9.0972)^2}{9.0972} + \frac{(6-9.0972)^2}{9.0972} + \frac{(5-1.4028)^2}{1.4028} = 28.2005$$

Degree of freedom is 3 and critical value is 7.815. Since 28.2005 > 7.815, we reject H_0 at 5% and the graduates rates are inappropriate.

(c) **[3 marks]** Consider the ages 20-25 only, calculate the kernel smoothed rate for age 22 from the crude rate provided in the table, using the triangle kernel with bandwidth 5.

Solution:

Age	Crude Rate	t	$K(t)$	w_j	$w_j y_j$
20	0.028884	-0.400000	0.4	0.076923	0.002222
21	0.016260	-0.200000	1.2	0.230769	0.003752
22	0.040123	0.000000	2	0.384615	0.015432
23	0.018386	0.200000	1.2	0.230769	0.004243
24	0.043750	0.400000	0.4	0.076923	0.003365
25	0.022293	0.600000	0	0	0
		Total	5.2		**0.0290**

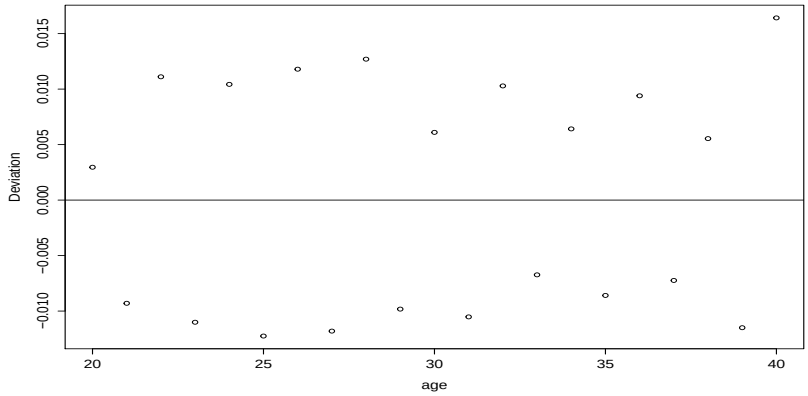
(d) **[3 marks]** The figure below shows the deviations resulted from the triangle kernel based on the entire sample. Using this figure to perform the sign test. State the required steps as in part (a) with a significance level of 5%.

Solution:

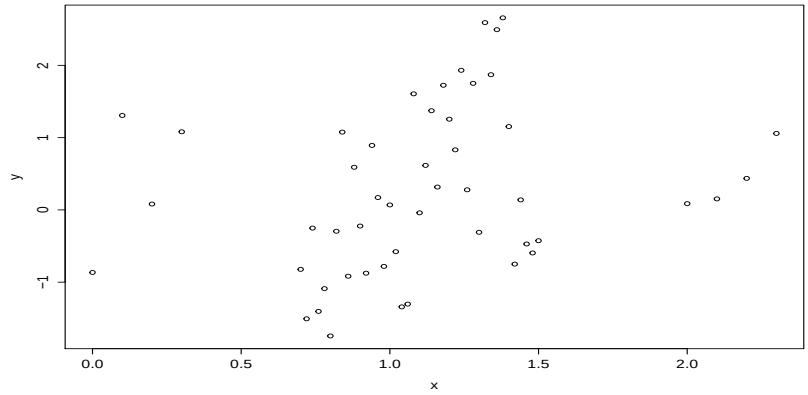
There are 11 positive deviations. Test statistic is

$$z = \frac{11 - 21 \times 0.5 - 0.5}{\sqrt{21 \times 0.5 \times 0.5}} = 0$$

Since $0 < 1.96$, we fail to reject H_0 at 5% and the graduates rates are appropriate.

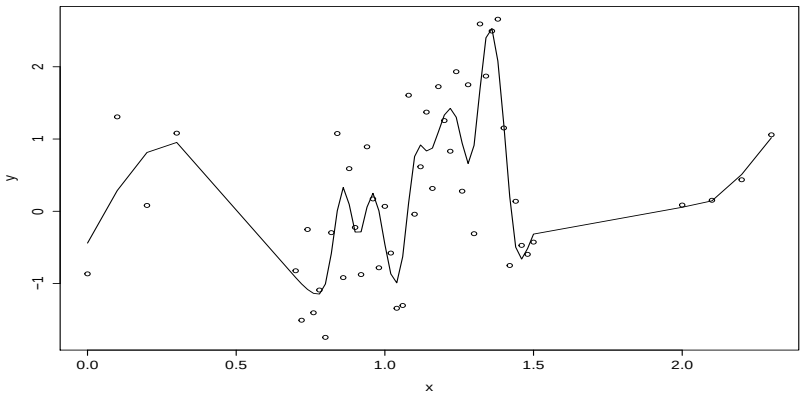


(e) **[2 marks]** The figure below shows a plot of a particular set of data that is going to be smoothed using kernel smoothing (in conjunction with the Normal kernel). Would you have any concerns smoothing the data depicted in this figure using kernel smoothing?



Solution:
Low density of points in some areas.

(f) **[2 marks]** We use the natural cubic spline to generate the smoothed curve as shown below. Do you think this is an optimal choice of knots? If not, how would you like to adjust the number of knots?



Solution:
No. Decrease the number of knots.

Question 5 [10 marks]

The table gives data on 18 lives from a portfolio of policies on impaired lives. The lives are classified in two categories, Smoking and Non-smoking. The table gives the time in months until a claim is made; the * indicates that the observation was censored.

Smoking	2 7* 9 10* 10* 10* 11 14 17
Non-smoking	4* 5 10* 10* 10* 12* 13 15* 18*

(a) **[4 marks]** Find the Nelson-Aalen estimate of the survivor function for all lives combined. *Note: You don't need to calculate the standard errors.*

Solution:

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	18	1	0.946	0.0541	0.8457	1
5	16	1	0.889	0.0766	0.7505	1
9	14	1	0.827	0.0941	0.6621	1
11	7	1	0.717	0.1375	0.4926	1
13	5	1	0.587	0.1729	0.3297	1
14	4	1	0.457	0.1886	0.2038	1
17	2	1	0.277	0.2271	0.0558	1

(b) **[2 marks]** In a Cox regression $\mu_x(t) = \mu_0(t)e^{\beta x}$, where $x = 0$ for smoking and $x = 1$ for non-smoking, show the **partial log likelihood function** $l(\beta)$. Simplify your results as much as possible.

Solution:

The partial likelihood is

$$L(\beta) = \frac{1}{9+9e^{\beta}} \cdot \frac{e^{\beta}}{8+8e^{\beta}} \cdot \frac{1}{7+7e^{\beta}} \cdot \frac{1}{3+4e^{\beta}} \cdot \frac{e^{\beta}}{2+3e^{\beta}} \cdot \frac{1}{2+2e^{\beta}} \cdot \frac{1}{1+e^{\beta}}$$

with Partial log likelihood

$$l(\beta) = 2\beta - 5log(1+e^{\beta}) - log(2+3e^{\beta}) - log(3+4e^{\beta})$$

(c) **[4 marks]** Apply the Maximum Likelihood method to estimate β. In the calculation, use the approximation $a+(a+1)e^{\beta} \approx (a+1)(1+e^{\beta})$. Provide the corresponding standard error. Use Z-test to test whether $\beta = 0$, you only need to **report the p-value** and write a conclusion based on 5 % significance level.

Solution:

$$\begin{aligned}
\frac{dl(\beta)}{d\beta} &= 2 - e^{\beta}\left(\frac{5}{1+e^{\beta}} + \frac{3}{2+3e^{\beta}} + \frac{4}{3+4e^{\beta}}\right) \\
&\approx 2 - e^{\beta}\left(\frac{5}{1+e^{\beta}} + \frac{1}{1+e^{\beta}} + \frac{1}{1+e^{\beta}}\right) \\
&= \frac{2-5e^{\beta}}{1+e^{\beta}}
\end{aligned}$$

Set it to 0, and $\hat{\beta} = log(2/5) = -0.9163$

Fisher Information is approximately

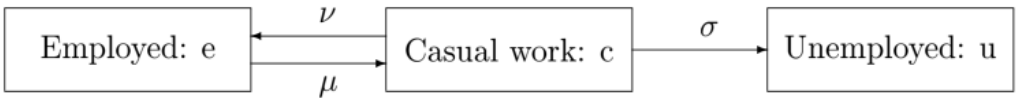
$$\begin{aligned}
-\frac{d^2l(\beta)}{d\beta^2} &= -7e^{\beta}\left(\frac{1}{(1+e^{\beta})^2} - \frac{1}{1+e^{\beta}}\right) \\
&= \frac{7e^{\beta}}{(1+e^{\beta})^2}
\end{aligned}$$

Since $e^{\hat{\beta}} = 2/5$, the variance is $(1+0.4)^2/7 \times 0.4 = 0.7$. Then standard deviation is 0.8367.

Test statistic is 0.9163/0.8367=-1.10. P-value is $P(|Z| > 1.10) = 0.2713$. Fail to reject H_0 and smoking is not a significant factor to affect mortality.

Question 6 [12 marks]

The following three state Markov model with constant transition intensities μ, ν and σ is used to represent the transition from employment to long-term unemployment.



(a) **[6 marks]** Show that ${}_tp_x^{eu}$ satisfies the differential equation

$$\frac{\partial^2}{\partial t^2}\,{}_tp_x^{eu} + (\mu+\sigma+\nu)\frac{\partial}{\partial t}\,{}_tp_x^{eu} + \sigma\mu\,{}_tp_x^{eu} = \sigma\mu$$

[Hint: show the first order of derivatives of ${}_tp_x^{eu}$ and ${}_tp_x^{ec}$ and then consider ${}_tp_x^{eu} + {}_tp_x^{ee} + {}_tp_x^{ec}$.]

Solution:

Apply the Kolmogorov forward equation and we have

$$\frac{\partial}{\partial t}{}_tp_x^{eu} = \sigma\,{}_tp_x^{ec}$$

and

$$\frac{\partial}{\partial t}{}_tp_x^{ec} = -(\sigma+\nu){}_tp_x^{ec} + \mu\,{}_tp_x^{ee}$$

Also, we know that ${}_tp_x^{ee} = 1 - {}_tp_x^{ec} - {}_tp_x^{eu}$. Substitute it into the second equation, we have

$$\frac{\partial}{\partial t}{}_tp_x^{ec} = -(\sigma+\nu){}_tp_x^{ec} + \mu(1 - {}_tp_x^{ec} - {}_tp_x^{eu}) = \mu - (\sigma+\nu+\mu){}_tp_x^{ec} - \mu\,{}_tp_x^{eu}$$

From the first equation, $\frac{\partial^2}{\partial t^2}{}_tp_x^{eu} = \sigma\frac{\partial}{\partial t}{}_tp_x^{ec}$, substitute it to the LHS of above equation. Also, substitute $\frac{\partial}{\partial t}{}_tp_x^{eu} = \sigma\,{}_tp_x^{ec}$ to the RHS will lead to the required result.

Now only consider the two-state case (the Casual work and Unemployment). Suppose we have the below observations ($\delta = 1$ for death and $\delta = 0$ for censoring) for a group of people aged x and the force of mortality is a constant (σ) in $[x, x+1]$:

x	a_i	b_i	δ_i	T_i
1	0	1	0	1
2	0.4	0.8	0	0.8
3	0.5	0.7	1	0.6
4	0	1	1	0.8

(b) **[2 marks]** Using the binomial model of mortality obtain the exact likelihood arising from the above data. Simplify your results as much as possible.

Solution:
The mortality rate $q = e^{-\mu}$. The waiting times are 1, 0.4, 0.2 and 1. The reason is that binomial model does not use exact time of death. Then, the likelihood is

$$\begin{aligned} L(\sigma) &= (1 - e^{-\sigma}) \cdot (1 - e^{-0.4\sigma}) \cdot e^{-0.2\sigma} \cdot e^{-\sigma} \\ &= (1 - e^{-\sigma})(1 - e^{-0.4\sigma})(e^{-1.2\sigma}) \end{aligned}$$

(c) **[4 marks]** Using the two-state model of mortality to calculate the maximum likelihood estimate of σ, and hence the maximum likelihood estimate of q_x. Provide a standard error of q_x.

Solution:
$\hat{\sigma} = D/V = 2/(1 + 0.4 + 0.1 + 0.8) = 0.8696$ and so $\hat{q}_x = 1 - e^{-\hat{\sigma}} = 0.5809$. The standard error of σ is $\sqrt{D/V^2} = \sqrt{2/2.3^2} = 0.6149$. Using the delta method, the standard error of q_x is $e^{-\hat{\sigma}} \cdot SE(\hat{\sigma}) = e^{-0.8696} \times 0.6149 = 0.2577$.

Question 7 [8 marks]

A life office has carried out an investigation of the mortality experience of certain of their policyholders. Part of the data is given below.

	Census Data					
Age	31/12/95	30/6/96	31/12/96	30/6/97	31/12/97	Total Deaths
50	5451	4420	4515	4773	5934	70
51	6002	5722	5534	4631	4428	76
52	5789	6121	6087	5322	5172	83

(a) **[3 marks]** Suppose first that the definitions of age for the census data and death are the same. Estimate the central exposed to risk E^c_{51}. Hence estimate μ_{51} and

q_{51}. When calculate q_{51}, assuming that μ_{51} is a constant over the age 51 and avoid using the initial exposure to risk.

Solution:
The central exposure to risk is $0.5\times(6002+5722)/2+0.5\times(5722+5534)/2+0.5\times(5534+4631)/2+0.5\times(4631+4428)/2 = 10551$. Then, $\hat{\mu}_{51} = 76/10551 = 0.0072$ and $\hat{q}_{51} = 1 - e^{-\hat{\mu}_{51}} = 1 - e^{-0.0072} = 0.0072$.

(b) **[3 marks]** Suppose now that the definition of age for the census data is: age x nearest birthday; the definition of age for those who died is: age x next birthday. Use this additional information to re-estimate E^c_{51}, μ_{51} and q_{51}. When calculate q_{51}, assuming that μ_{51} is a constant over the age 51 and avoid using the initial exposure to risk.

Solution:
Define the numbers alive aged x for death and census are $P^D_{x,t}$ and $P^C_{x,t}$, respectively. For the age definition, we have $P^D_{x,t} = 0.5(P^C_{x-1,t} + P^C_{x,t})$.
Then, the central exposure to risk is

$$0.25(P^D_{x,0} + 2P^D_{x,0.5} + 2P^D_{x,1} + 2P^D_{x,1.5} + P^D_{x,2})$$

Substitute $P^D_{x,t}$ by $P^C_{x,t}$, we have

$$\begin{aligned} E^c_{51} =& 0.125[P^C_{50,0} + P^C_{51,0} + 2(P^C_{50,0.5} + P^C_{51,0.5}) + 2(P^C_{50,1} + P^C_{51,1}) \\ &+ 2(P^C_{50,1.5} + P^C_{51,1.5}) + P^C_{50,2} + P^C_{51,2}] \\ &= 10125.625 \end{aligned}$$

Then $\hat{\mu}_{51} = 76/10125.625 = 0.0075$ and $\hat{q}_{51} = 0.0075$

(c) **[2 marks]** Using the poisson model to calculate an approximate standard error for $\hat{\mu}_{51}$ derived in part (b).

Solution:
The standard error of $\hat{\mu}_{51}$ is $\sqrt{d_x}/E^c_x = \sqrt{76}/10125.625 = 0.0009$.

Question 8 [8 marks] (For students enrolled in STAT7042/8003 ONLY)

(For each part, you will gain 2 marks for a correct answer, be penalized 1 mark for an incorrect answer, and score 0 if no answer is given.) Answer each

question "TRUE" or "FALSE". In each case, write the whole word. It is **not** acceptable to write only "T" or "F" and answers presented in this form **will be graded incorrect**.

(a) **[2 marks]** Lasso regression can lead to exact 0 for the estimated parameters.

(b) **[2 marks]** Estimates of Ridge regression are unbiased.

(c) **[2 marks]** We cannot use Bootstrap to calculate confidence interval for the median.

(d) **[2 marks]** The test statistic r_j for the jth Serial Correlation asymptotically follows the standard normal distribution.

Solution:
TRUE FALSE FALSE FALSE

END OF EXAMINATION