Dept of Mathematics and Statistics
King Fahd University of Petroleum & Minerals
AS475: Survival Models for Actuaries
Dr. Mohammad H. Omar
Final Exam Term 182 FORM A
Monday April 22 2019
7.00pm-9.30pm

Name ____ID#:___________________ Serial #:_______

Instructions.

1. Please turn off your cell phones and place them under your chair. Any student **caught with mobile phones** on during the exam will be considered under the cheating rules of the University.
2. If you need to leave the room, please do so quietly so not to disturb others taking the test. No two person can leave the room at the same time. NO EXTRA exam time will be provided for the time spent outside the room.
3. Only materials provided by the instructor can be present on the table during the exam.
4. DO NOT spend TOO MUCH TIME on any one question. If a question seems too difficult, leave it and go on.
5. Use the blank portions of each page for your work. Extra blank pages can be provided if necessary. If you use an extra page, indicate clearly what problem you are working on.
6. Only answers supported by work will be considered. Unsupported guesses will not be graded.
7. While every attempt is made to avoid defective questions, sometimes they do occur. In the rare event that you believe a question is defective, the instructor cannot give you any guidance beyond these instructions.
8. Mobile calculators, I-pad, or communicable devices are disallowed. Use regular scientific calculators, financial calculators, or SOA approved calculators only. Write important steps to arrive at the solution of the exam problems.

The test is 2 parts each is 75 minutes, GOOD LUCK, and you may begin now!

Part	Question	Total Marks	Marks Obtained	Comments
I	1	7		
I	2	6+5 = 11		
I	3	8		
I	4	3+5+3=11		
I	5	5+5=10		
II	6	6+5+4=15		
II	7	3+3+2+4+2=14		
II	8	3+3+3=9		
II	9	4+5+2=11		
II	Output	4		
	Total	100		

Extra blank page

1) (7 marks) You are given the following ages at time of death for 10 individuals:

25	30	35	35	37	39	45	47	49	60

Using a uniform kernel with a bandwidth of $b = 10$, determine the **kernel density estimate** of the probability of **survival to age 40**.

2) (6+5=11marks) Five observations are made from a random variable. They are 1, 2, 3, 5, and 13.

a) Determine the value of the **Kolmogorov-Smirnov** test statistics for the null hypothesis that $f(x) = 2x^{-2}e^{-2/x}, x > 0$.

b) **Test** that the data came from the distribution in (a) above.

3) (8 marks) One hundred and fifty policyholders were observed from the time they arranged a **viatical settlement** until their death. A **viatical settlement** is an arrangement in which someone with a terminal disease sells his or her life insurance policy at a discount from its face to get immediate insurance cash benefit. No observations were censored.

year	1st	2nd	3rd	4th
deaths	21	27	39	63

The survival model

$$S(t) = 1 - \frac{t(t-1)}{20}, \qquad 0 \le t \le 4,$$

is being considered. At a 5% significance level, conduct the chi-square goodness of fit test.

4) (3+5+3=11 marks) The following is an edited printout of the results obtained by fitting an extended Cox model containing two heaviside functions of the time-dependent sex variable on the "Anderson.dat" leukemia data:

Time-Dependent Cox Regression Analysis

Analysis time_t: surv_t	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% interval	Confidence]
log WBC	1.567	0.333	0.000	4.794	2.498	9.202
Rx	1.341	0.466	0.004			
0–15 wks	0.358	0.483	0.459	1.430	0.555	3.682
15+ wks	0.182	0.992	0.855	0.834	0.119	5.831

No. of subjects = 42 Log likelihood = -71.980

Using the above computer results, adjusted for log WBC and the two time-dependent Sex variables,

a) estimate the hazard ratio for the **treatment effect** (Rx), and
b) obtain **its** 95% confidence interval
c) test the hypothesis of **no treatment effect**.

5) (5+5=10 marks) Suppose that Al (A), Sam (S), and Charlie (C) are the only three subjects in the dataset shown below. All three subjects have two recurrent events that occur at different times.

ID	Status	Stratum	Start	Stop	tx
A	1	1	0	65	1
A	1	2	65	90	1
S	1	1	0	20	0
S	1	2	20	30	0
C	1	1	0	10	1
C	1	2	10	40	1

a) Fill in the following data layout describing survival (in weeks) to the **first event** (stratum 1). Recall that m_f and q_f denote the number of failures and censored observations at time $t_{(f)}$. The survival probabilities in the last column use the KM product limit formula.

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_1(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

b) Fill in the following data layout describing survival (in weeks) from the **first to second event** using the **Gap Time** approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_2(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

Bonus question

B1) (5 marks) The Akaike Information Criterion (AIC) is a method designed to compare the fit of different models. The three survival models are below compared using the same 5 predictors:

1. A Weibull model without frailty
2. A Weibull model containing a frailty component and
3. The log-logistic model.

Below is a table containing the log-likelihood statistic for each model.

Model	Frailty	Number of parameters	Log likelihood
1. Weibull	No	7	-206.204
2. Weibull	Yes	8	-200.193
3. Log-logistic	No	7	-200.196

Which of the three models should be selected based on the AIC?
(Hint: AIC= -2log likelihood +2p where p=# of parameters)

Name ____ID#:__________________ Serial #:_______

6) (6+5+4=15 marks) A generalized gamma model using the Veterans' (vets) data in R with five predictor variables were conducted. The generalized gamma distribution contains two shape parameters (kappa and sigma) that allow great flexibility in the shape of the hazard. The output is shown below.

Gamma regression — accelerated failure-time form

LR chi2(5) = 52.86

Log likelihood = -200.626 Prob > chi2 = 0.0000

_t	Coef.	Std. Err.	z	p>|z|
tx	[ ]	.1908	[ ]	0.491
perf	0.039	.0051	7.77	0.000
dd	0.0004	.0097	0.04	0.965
age	0.008	.0095	0.89	0.376
priortx	0.004	.0229	0.17	0.864
_cons	1.665	.7725	2.16	0.031
/ln_sig	.0859	.0654	1.31	0.189
/kappa	[ ]	.2193	[ ]	0.279
sigma	1.0898		.0714	

a) Estimate the acceleration factor γ with a 95% confidence interval comparing the **test and standard treatment (TX =2 vs. TX = 1)**.

b) Use the output to test the null hypothesis that a lognormal distribution is appropriate for this model. (Hint: If kappa = 0 the model reduces to a lognormal distribution)

c) A lognormal model and log-logistic models were run with the same five predictors above and yielded the results below

Model	Number of parameters	Log likelihood
1. log normal	7	[ ]
2. Log-logistic	7	-200.196

Comparing the AIC of the generalized gamma model, the lognormal model, and the log-logistic model, select the **best model**.

7) (3+3+2+4+2=14 marks) The R data "bladder2" contains 85 subjects with nonzero follow-up who were assigned to either thiotepa or placebo, and only the first four recurrences for any patient. The status variable is 1 for recurrence and 0 for everything else (including death for any reason). The data is laid out in the (start, stop] format.

a) The following Recurrent data Cox model was conducted. Complete the blanks in the output below.

```
Call:
coxph(formula = Y ~ rx + number + cluster(id), data = bladder2)

            coef exp(coef) se(coef) robust se      z       p
rx      [        ]   0.62845  0.20015  [       ] -1.730 [       ]
number   0.18161   1.19915  0.04579   0.06088  2.983 0.00285

Likelihood ratio test=17.1  on 2 df, p=0.0001931
n= 178, number of events= 112
```

b) A similar model as (a) with extra predictor variable "size" has likelihood ratio test value of 17.52 with 3 df. Compare this model against the model in (a) above to see if "size" is a significant predictor.

c) The actual number of patients is 85 but n=178 in the printout above. Explain why this discrepancy.

d) The following Recurrent data Cox model was conducted. Complete the blanks in the output below.

```
Call:
coxph(formula = Y ~ rx + number + strata(enum) + cluster(id),
    data = bladder2)

            coef exp(coef) se(coef) robust se      z      p
rx      [        ]   0.71789  0.21559  [       ] -1.620 [      ]
number   0.12135   1.12902  0.05124   0.04843  2.505 0.0122

Likelihood ratio test=[   ]  on 2 df, p=0.03885
n= 178, number of events= 112
```

e) Based on the model in (d), does the treatment, rx, appear to be effective.

8) (3+3+3=9 points) You are given:

(i) A sample of insurance claim payments is:

29	64	90	135	182

(ii) Claim sizes are assumed to follow an exponential distribution.

(iii) The mean of the exponential distribution is estimated using the maximum likelihood estimation to be 100. That is, the rate $\hat{\lambda} = \frac{1}{100} = 0.01$.

Use R to

a) calculate the value of the **Kolmogorov-Smirnov** test statistic and complete the following output

```
One-sample Kolmogorov-Smirnov test

data:
D = [        ], p-value = [      ]
alternative hypothesis: two-sided
```

b) calculate the value of the **Anderson-Darling** test statistic and complete the following output. (Hint: use the Anderson darling test in the "goftest" package).

```
Anderson-Darling test of goodness-of-fit
    Null hypothesis: exponential distribution
    with parameter rate = 0.01

data:  e
An = [        ], p-value = [       ]
```

c) What is your conclusion regarding whether or not the data follows the exponential distribution.

9) (4+5+2=11 marks) Ten payments were recorded as follows: 4, 4, *5+*, *5+*, *5+*, 8, *10+*, *10+*, 12, and 15, with the italicized + values representing payments at a policy limit. There were no deductibles.

a) Kaplan Meier estimates were done through R below. Complete the missing product limit estimates below and the Greenwood's approximation of variance for S(12).

```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   4     10       2     0.80   0.126       0.5868            1
   8      5       1    [    ]   0.175       0.3742            1
  12      2       1    [    ] [     ]       0.0724            1
  15      1       1     0.00     NaN           NA           NA
```

b) The exponential model was fitted to the survival data in part (a) above. Complete the missing information in R output below.

```
Call:
survreg(formula = surdays ~ 1, dist = "exponential")
            Value Std. Error    z       p
(Intercept) [    ]      [    ] 6.14 8.1e-10

Scale fixed at 1

Exponential distribution
Loglik(model)= [    ]    Loglik(intercept only)= [    ]
Number of Newton-Raphson Iterations: 4
n= 10
```

c) Write the form of cox model's baseline hazard, $h_0(t)$, as defined by the exponential hazard in (b) above.

END OF TEST