THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

May 2022 Project

ST227

Survival Analysis

2020/21 and 2021/22 syllabus only

Instructions to candidates

This paper contains four questions. Answer **ALL FOUR**.
Question 1: 25 marks
Question 2: 30 marks
Question 3: 25 marks
Question 4: 20 marks

The marks in brackets reflect marks for each part of a question.

Time allowed **Reading Time:** *None*
Writing Time: *2 hours*

1. Consider the following mortality intensity function:

$$\mu(t) = \frac{\alpha\gamma t^{\gamma-1}}{1+\alpha t^{\gamma}}, \quad \alpha = 3.757 \times 10^{-3}, \gamma = 1.4243.$$

(a) Define in R the survival probability function $(t, x) \to {}_tp_x$ and calculate the probability of surviving the next 3 years for a 20-year-old individual.

[5 *marks*]

(b) Define in R the density function for T_{20}. This definition may involve a numerical integral.

[5 *marks*]

(c) Calculate the expected remaining lifetime for a 20-year-old individual.

[5 *marks*]

(d) Define in R the cumulative distribution function of T_{20}. This definition may involve a numerical integral.

[5 *marks*]

(e) Discuss how one can numerically find the median of this distribution. Outline the approach only, you are **not** required to solve for the median.

[5 *marks*]

[Total 25 *marks*]

2. This questions is divided into two parts. Both parts use the same data set of fully observed lifetimes given below:

80	75	38	45	62	65	77	92	65	60
55	67	72	46	64	35	68	52	45	94

(a) Let us suppose that this data set comes from a Log-Normal distribution, i.e:

$$f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\Big) \tag{1}$$

i. Using the results:

$$E(X) = \exp(\mu + \frac{\sigma^2}{2}), \quad \mathrm{Var}(X) = (\exp(\sigma^2)-1)\Big(\exp(\mu+\frac{1}{2}\sigma^2)\Big)^2, \tag{2}$$

derive the method of moment estimators for μ and σ.

[5 *marks*]

ii. Using your MMEs above as the initial values for `optim` or otherwise, derive the MLE for μ and σ.

[10 *marks*]

(b) We propose a lifetime model with the following mortality intensity function:

$$\mu(t) = \lambda\gamma(\lambda t)^{\gamma-1}, \quad t \geq 0.$$

i. Derive algebraically the probability density function for lifetime and write down the joint-likelihood of the given sample.

[5 *marks*]

ii. Using `optim` and the initial values $\lambda_0 = 67$ and $\gamma_0 = 0.2233$, numerically obtain the maximum likelihood estimators of the model parameters.

[8 *marks*]

iii. For the purpose of lifetime modelling, what range of values for γ would yield a sensible model?

[2 *marks*]

[Total 30 *marks*]

3. Cancer patients who are in remission are observed and the number of days until the symptoms reappear is recorded. Some records have been right-censored. The data set is provided in a spreadsheet named cancer.xlsx and the columns therein are:

- time: the time until reappearance of symptoms in number of days.
- event: an indicator variable taking value 0 if the record has been right-censored and 1 if fully observed.
- fullyObserved: logical variable indicating whether the record has been fully observed.
- sex: categorical variable with value 0 for male (the reference group) and 1 for female.

(a) Calculate the Kaplan-Meier estimate for survival probabilities.

[15 *marks*]

(b) Denote by T the time until reappearance of cancer. Using the following formula:

$$E(T^m) = \int_0^\infty m t^{m-1} P(T > t) dt,$$

propose and calculate a suitable estimation for $\mathrm{Var}(T)$. (Hint: you can use Midpoint Rule, Trapezoidal/Trapezium Rule or any geometric method of approximating the area under the curve.)

[10 *marks*]

[Total 25 *marks*]

4. In this question, we will fit a Cox Proportional Hazard model on the same data set in Question 3, with time as the response variable and sex as the categorical covariate.

(a) By using the survival package, calculate the MLE for the Cox Proportional Hazard Model.

[10 *marks*]

(b) Based on the output you have generated, perform the z-test, Score test, and Likelihood Ratio test on the following hypotheses:

$$H_0 : \beta = 0, \quad \text{vs} \quad H_1 : \beta \neq 0.$$

[7 *marks*]

(c) It is hypothesised that cases with remission time greater than 125 belong to a different class of cancer. Create a data frame containing this subset of cases.

[3 *marks*]

[Total 20 *marks*]