## MA2261 Examination — Draft Questions and Solutions

This copy generated 10th May 2022.

| | |
|---|---|
| Title of paper | MA2261 — Linear Statistical Models |
| Version | 1 |
| Candidates | All candidates |
| Department | School of Computing and Mathematical Science |
| Examination Session | **June Examinations 2022** |
| Time allowed | 2 hours |
| Instructions | Answer all questions. |
| Calculators | Approved calculators may be used. |
| Books/statutes | Statistical tables are provided |
| Own Books/statutes/notes | No |
| Additional Stationery | No |
| Number of questions | 4 |

**Question 1.**

A continuous random variable $X$ has a probability density function given by

$$f(x,\lambda) = \begin{cases} \dfrac{xe^{-\left(\frac{x}{\lambda}\right)}}{\lambda^2}, & x > 0 \\ 0, & x \leq 0\,, \end{cases}$$

where $\lambda$ is a parameter and $\lambda > 0$.

In a factory of electronic components, the life time $X$ is assumed to follow the above distribution. Twenty components were randomly selected, tested and their following life time observations $x_1, \dots x_{20}$ of $X$ (in thousand hours) were recorded.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 6 | 1.2 | 4.2 | 3.2 | 7.5 | 11 | 16 | 7.5 | 21 |
| 6.5 | 2 | 18 | 14 | 9.5 | 13.2 | 13.9 | 6.2 | 12.2 | 31 |

In your answers you can use, without proof, the following calculus facts:

$$\frac{d}{dz}\left(-e^{-z}(z+1)\right) = ze^{-z},$$
$$\sum_{i=1}^{20} x_i = 220.1$$

a) **[5 marks]**

Find the cumulative distribution function of $X$.

b) **[5 marks]**

Show that the log-likelihood function $\ell(\lambda)$ for the given dataset is

$$\ell(\lambda) = \ln \mathcal{L}(\lambda) = -2n\ln\lambda - \sum_{i=1}^{n} \frac{x_i}{\lambda} + \sum_{i=1}^{n} \ln x_i$$

c) **[5 marks]**

Apply the method of maximum likelihood to find the maximum likelihood estimate of the parameter $\lambda$ for the given dataset. You may assume any critical point of the log-likelihood function is a maximum.

**Total 15 marks**

**Answer — total marks for this question: 15**

Parts a), b) and c) are standard questions, but the function considered is different from coursework and workbook questions.

a) **[5 marks]**

To calculate the CDF we need to integrate the $f(x,\lambda)$:

$$F(x,\lambda) = \int_{-\infty}^{x} \frac{te^{-(\frac{t}{\lambda})}}{\lambda^2} dt = \int_{0}^{x} \frac{te^{-(\frac{t}{\lambda})}}{\lambda^2} dt$$

Let $z = \dfrac{t}{\lambda}$. Then $dt = \lambda dz$ and

$$\int_0^x \frac{te^{-\frac{t}{\lambda}}}{\lambda^2} dt = \int_0^{\frac{x}{\lambda}} \frac{1}{\lambda} ze^{-z} \lambda dz = \int_0^{\frac{x}{\lambda}} ze^{-z} dz =$$
$$= -e^{-z}(z+1)\Big|_0^{\frac{x}{\lambda}} = -e^{-(\frac{x}{\lambda})}(\frac{x}{\lambda}+1)+1 \ .$$

b) **[5 marks]**

The likelihood function is

$$\mathcal{L} = \mathcal{L}(\lambda, x_1, \ldots, x_n) = \begin{cases} \dfrac{1}{\lambda^{2n}} \displaystyle\prod_{i=1}^{n} (x_i e^{-\frac{x_i}{\lambda}}), & x_i \geq 0, \quad 1 \leq i \leq n \\ 0, & \text{otherwise} \end{cases}$$

Hence

$$\ell(\lambda) = \ln \mathcal{L}(\lambda) = -2n \ln \lambda - \sum_{i=1}^{n} \frac{x_i}{\lambda} + \sum_{i=1}^{n} \ln x_i$$

c) **[5 marks]**

After deriving with respect to $\lambda$ and equating to zero we get:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -\frac{2n}{\lambda} + \sum_{i=1}^{n} \frac{x_i}{\lambda^2} = 0 \iff -2n\lambda + \sum_{i=1}^{n} x_i = 0 \ .$$

Therefore, we can calculate the estimate $\hat{\lambda}$:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{2n} \text{ with } n = 20.$$

We use $\displaystyle\sum_{i=1}^{20} x_i = 220.1$, hence

$$\hat{\lambda} = \frac{220.1}{40} = 5.5025 \ .$$

**Question 2.**

A team of scientists from the University of Oxford is on a mission in Antarctica to study the dust concentration $(Y)$ in the ice of different geological eras. To this end, they drill the ice and obtain samples at different depths (between 300 and 1600 m) corresponding to periods $(x)$ in the past between 3000 and 280,000 years ago. The collected data and statistical analysis, using simple linear regression, were sent back to Oxford. kYr means 1000 years.

$\bar{x} = 92.4, \quad \bar{y} = 86.805, \; S_{xx} = 85,114.4, \; S_{yy} = 19,179.1, \; S_{xy} = 33,123.77,$
$RSS = 6,288.401$

| x (kYr) | Y Dust concentration (ppm) |
|---|---|
| 3 | 39.36 |
| 6 | 24.25 |
| 15 | 36.91 |
| 25 | 97.32 |
| 46 | 60.5 |
| 66 | 114.91 |
| 102 | 81.5 |
| 161 | 121.3 |
| 220 | 164.8 |
| 280 | 127.2 |

The Oxford team has a deadline to submit their findings to a Government agency and they need to produce, without delay, a statistical analysis using the information they have.

a) **[3 marks]**

Obtain the estimated simple linear regression line.

b) **[10 marks]**

Decide by conducting a statistical test with significance level $\alpha = 5\%$ if the period variable $x$ has a statistically significant effect on the mean dust concentration.

c) **[2 marks]** Calculate a point estimate for the mean increase in dust concentration corresponding to an increase in one unit (that is 1000 years) in $x$.

**Total: 15 marks**

**Answer — total marks for this question: 15**

Part a) and b) are entirely standard, but the dataset and context are different from the ones of coursework and workbook questions. Part c) is a new type of question.

a) **[3 marks]**

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{33,123.77}{85,114.4} \approx 0.389 \,,$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x} \approx 50.846 \,.$$

b) **[10 marks]**

We now want to test the hypothesis $H_0 : b = 0$. If $H_0$ is true, then we need to check if the $t$ value:

$$t = \frac{\hat{b}}{\hat{\sigma}/\sqrt{S_{xx}}}$$

is less or equal to the critical value $t_{0.0975,n-2} = t_{0.0975,8} = 2.306$ (assuming a confidence level of 95%).

To calculate $\hat{\sigma}$ we use

$$\hat{\sigma} = \sqrt{\frac{RSS}{(n-2)}} = \sqrt{\frac{6288.401}{(10-2)}} \approx 28.037$$

Therefore

$$t = \frac{0.389}{28.037/\sqrt{85114.4}} \approx 4.048$$

Since $t = 4.048$ is greater than the critical value $2.306$, $H_0$ is rejected and $x$ is statistically significant.

c) **[2 marks]**

$$\hat{y}(x+1) = \hat{a} + \hat{b}(x+1), \quad \hat{y}(x) = \hat{a} + \hat{b}(x) \,.$$

Hence

$$\hat{y}(x+1) - \hat{y}(x) = \hat{b} \approx 0.389.$$

**Question 3.**

Hooke's law states that the elongation L of a spring subjected to a weight force W is given by $W = kL$, where $W$ is measured in $kg$, $L$ is measured in $mm$, $k$ is called the elastic constant of the spring and is expressed in $kg/mm$. A group of Physics students wants to measure the elastic constant of a spring. Not having very precise instruments they perform an experiment by applying to the spring weights of increasing intensity, from $10$ to $50\,kg$, for five times. The measurements are influenced by the approximation of the reading of the lengths and by the fact that the spring does not behave like a perfect spring and the same weight applied several times does not give the same elongation. The measurements of elongation, in millimeters, for each test done are shown below.

| Weight (kg) | Measured elongation L (mm) | | | | |
|---|---|---|---|---|---|
| W | L1 | L2 | L3 | L4 | L5 |
| 10 | 48.6 | 47.6 | 48.8 | 51.5 | 49.8 |
| 15 | 78.4 | 77.5 | 71.6 | 77.5 | 73.6 |
| 20 | 95.7 | 98.6 | 100.4 | 102.4 | 97.3 |
| 25 | 123.5 | 131.1 | 118.9 | 130.6 | 128.3 |
| 30 | 150.6 | 154.5 | 148.3 | 146.0 | 153.3 |
| 35 | 175.5 | 176.3 | 173.2 | 181.8 | 181.8 |
| 40 | 209.4 | 199.8 | 197.8 | 195.9 | 203.5 |
| 45 | 230.9 | 233.2 | 230.5 | 218.7 | 222.6 |
| 50 | 245.2 | 249.2 | 257.0 | 256.7 | 244.3 |

a) **[4 marks]**

Write the equation of a simple linear regression model with repeated observations for this dataset.

b) **[8 marks]**

Perform a statistical test with significance level $\alpha = 5\%$ to establish if a linear regression model with repeated observations is a good fit for these data. You can use the following values:

$SST = 192093$, $SSM = 191345$ and $SSB = 191387$.

c) **[3 marks]**

Write the general formula of the test statistic and its distribution you can use to decide if Hooke's law can be assumed to be valid for this dataset.

**Total: 15 marks**

**Answer — total marks for this question: 15**

Similar to previous years' exam paper.

a) **[4 marks]** Familiar definition.

The model has equation $L_{ij} = a + bw_i + \varepsilon_{ij}$ for $i = 1, \cdots, 9,\ j = 1, \cdots, 5$, where the $\varepsilon_{ij}$ are independent random variables with normal distribution $N(0, \sigma^2)$, and $a$, $b$, and $\sigma^2$ are unknown parameters, $\sigma^2$ does not depend on x.

b) **[8 marks]** Unseen example, familiar techniques.

$$SSE = SST - SSM = 192093 - 191345 = 748$$
$$SSW = SST - SSB = 192093 - 191387 = 706$$
$$SSL = SSE - SSW = 748 - 706 = 42$$

Since $k = 9$, $N = 45$ we obtain

$$\frac{SSL/7}{SSW/36} = 0.306 < 2.28.$$

Because $2.28$ is the critical value for $\alpha = 5\%$ of a $F_{7,36}$-distribution, the $p$-value > $5\%$.Therefore the hypothesis that the linear regression model is a good fit is accepted.

c) **[3 marks]** Newer kind of question.

We test the null hypothesis $a = 0$ with the following test statistic $t$:

$$t = \frac{\hat{a}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2} = t_{43} \ \text{(t-distribution with 43 degrees of freedom)}$$

**Question 4.**

In this question, $X$ and $Y$ are jointly distributed discrete random variables. The probability $P(X = i, Y = j)$ is given by the entry in column $i$, row $j$ of the following table with $p, q \in [0, 1]$:

| | | $X$ | | |
|---|---|---|---|---|
| | | 0 | 1 | 2 |
| $Y$ | 0 | 0 | $pq$ | 0 |
| | 1 | $\frac{p(1-q)}{2}$ | $1-p$ | $\frac{p(1-q)}{2}$ |

a) **[5 marks]**

Calculate $P(Y = 0)$ and $P(Y = 1)$.

b) **[8 marks]**

For which values of $p$ and $q$ are the **events** $\{X = 0\}$ and $\{Y = 1\}$ independent?

c) **[2 marks]**

Give a formal definition of what it means for two random variables to be independent.

**Total: 15 marks**

**Answer — total marks for this question: 15**

a) **[5 marks]** Unseen example, standard techniques.

The events $\{X=0\}$, $\{X=1\}$ and $\{X=2\}$ partition the sample space, so

$$\begin{aligned}
P(Y=0) &= P(X=0,Y=0)+P(X=1,Y=0)+P(X=2,Y=0) = 0+pq+0 = pq\ ,\\
P(Y=1) &= P(X=0,Y=1)+P(X=1,Y=1)+P(X=2,Y=1)\\
&= \frac{p(1-q)}{2}+1-p+\frac{p(1-q)}{2} = 1-pq
\end{aligned}$$

,

b) **[8 marks]** Unseen example, standard techniques.

$$\begin{aligned}
P(X=0) &= \frac{p(1-q)}{2},\\
P(Y=1) &= 1-pq,\\
P(X=0,Y=1) &= \frac{p(1-q)}{2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
& P(X=0,Y=1) = P(X=0)P(Y=1)\\
\iff\ & \frac{p(1-q)}{2} = \frac{p(1-q)}{2}(1-pq)\\
\iff\ & p(1-q) = p(1-q)(1-pq)\\
\iff\ & p=0 \lor q=1 \lor (p\neq 0 \land q\neq 1 \land 1 = 1-pq)\\
\iff\ & p=0 \lor q=1 \lor (p\neq 0 \land q\neq 1 \land pq = 0)\\
\iff\ & p=0 \lor q=1 \lor (p\neq 0 \land q\neq 1 \land (p=0 \lor q=0))\\
\iff\ & p=0 \lor (p\neq 0 \land q=0) \lor q=1\\
\iff\ & p=0 \lor q=0 \lor q=1.
\end{aligned}$$

$\{X=0\}$ and $\{Y=1\}$ are independent iff $p=0 \lor q=0 \lor q=1$.

c) **[2 marks]** Familiar definition.

Random variables $X$ and $Y$ are independent if for all $x,y$ we have:

$$P(X\le x, Y\le y) = P(X\le x)P(Y\le y).$$

It's ok to give a definition here that only works in the discrete case, i.e. for all $x,y$:

$$P(X=x,Y=y) = P(X=x)P(Y=y).$$