Imperial College
London

M2S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2017

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling I

Date: Thursday 18 May 2017

Time: 14:00 - 16:00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should start their solutions to each question in a new main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers to each question as per the table below.

Raw Mark	Up to 12	13	14	15	16	17	18	19	20
Extra Credit	0	½	1	1 ½	2	2 ½	3	3 ½	4

- Each question carries equal weight.
- Calculators may not be used.

1. Let us consider a random sample of size n, such that the random variables $Y_1, \ldots, Y_n$ follow an Exponential(λ) distribution, with some unknown parameter $\lambda > 0$.
Recall: The probability density function (pdf) of a random variable $Z \sim Exp(\lambda)$ is $f(z) = \lambda exp(-\lambda z)$ for $z > 0$, $\lambda > 0$. The cumulative distribution function of Z is $F(z) = 1 - exp(-\lambda z)$. We note that Z has mean λ^{-1} and variance λ^{-2}.

   (a) Two possible estimators for the mean $1/\lambda$ of such an Exponential(λ) distribution are $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $T = n\bar{Y}/(n+1)$. For both of these estimators find the
      (i) biases,
      (ii) variances,
      (iii) MSEs.

      Discuss why you might choose each of these estimators for estimating the mean.

   (b) Now let us consider i.i.d. samples $Y_1, \ldots, Y_n$ from a distribution with pdf $e^{-(y-\theta)}$, for $y \geq \theta$. Show that
      (i) $X = n(Y_{(1)} - \theta)$ has pdf e^{-x}, for $x \geq 0$, where $Y_{(1)} = \min(Y_1, \ldots, Y_n)$,
      (ii) $Y_{(1)}$ is the maximum likelihood estimator (MLE) of θ and is consistent for θ.
      Hint: Show consistency via the definition of consistency.

2. Suppose we observe a realisation $y_1, \ldots, y_n$ of the iid random variables $Y_1, \ldots, Y_n$ which follow a uniform distribution on $[\theta, \theta + 1]$. The unknown parameter is $\theta \in \mathbb{R}$.

   (a) Carefully write down the pdf f_θ of Y_1.

   (b) Write down the likelihood function L of the observation, including its domain. Simplify the likelihood function as far as possible.

   (c) Find the maximum likelihood estimator(s).

   (d) Suppose $n = 4$ and suppose we observe the data $0.5, 0.3, 1.1, 1.2$. Give all maximum likelihood estimates of θ.

   (e) Suppose we observed the data $0.5, 0.1, 1.1, 1.2$. What can be said about the assumed model?

   (f) Consider $T = \max(Y_1, \ldots, Y_n) - \theta$. Is T a pivotal quantity for θ?

   (g) Construct a finite 95% confidence interval for θ.

3. (a) Consider a linear model parameterised by parameters $\boldsymbol{\beta}$, and with full rank design matrix X (with rank r), data $\mathbf{Y}$ and error $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. Define the following terms:

(i) vector of fitted values,

(ii) vector of residuals.

Show that the vector of fitted values is orthogonal to the vector of residuals.

(b) Write down three expressions for the residual sum of squares; one involving Q (the projection matrix onto the space orthogonal to the space spanned by the linear model), one involving $\mathbf{e}$ (the vector of residuals) and another involving $\widehat{Y}^T\widehat{Y}$.
Show that these three expressions are indeed equivalent.

(c) Prove that $\frac{\text{RSS}}{n-r}$ is an unbiased estimator of σ^2.

(d) By considering a particular linear model, show that the sample variance for an i.i.d. sample can be written as $\frac{\text{RSS}}{n-r}$.

4. (a) Describe in detail how you would draw a sample on a computer from an $F_{4,4}(2)$ distribution. Clearly define any intermediate results and definitions you use. You may only assume that you have access to a function that allows you to draw samples from a univariate Normal distribution with a mean, μ, and variance, σ^2, of your choice.

(b) A surveyor measures once each of the angles α, β, γ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that $\text{Var}(Y_i) = \sigma^2$, $i = 1, 2, 3$. The surveyor constructs the following model,

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon},$$

and derives the least squares estimators of α, β and γ. However, he wonders if it might be possible to obtain a better estimator for the angles?

Construct an alternative linear model for the surveyor, derive an estimator for the model parameters and for σ^2, then prove that the resulting estimator for the angles has a lower variance than the naive model given above.

(c) Consider the linear model

$$\boldsymbol{Y} = \begin{pmatrix} 1 & o_1 \\ \vdots & \vdots \\ 1 & o_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \boldsymbol{\epsilon},$$

for which we wish to test the following hypotheses,

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0.$$

(i) Describe a concrete example that could be modelled using the above model and describe how the hypotheses may be interpreted.

(ii) Describe how we could test the above hypotheses. State clearly any assumptions you make, which test statistic you would use, and an appropriate rejection rule.

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2017

This paper is also taken for the relevant examination for the Associateship.

M2S2

Statistical Modelling (Solutions)

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

1. (a) (i) $\mathrm{E}(\bar{Y}) = \lambda^{-1}$, $\mathrm{bias}(\bar{Y}) = 0$,
$\mathrm{E}(T) = n\lambda^{-1}/(n+1)$, $\mathrm{bias}(T) = -\lambda^{-1}/(n+1)$

(ii) $\mathrm{Var}(\bar{Y}) = \lambda^{-2}/n$, $\mathrm{Var}(T) = n\lambda^{-2}/(n+1)^2$,

(iii) $\mathrm{MSE}(\bar{Y}) = \lambda^{-2}/n$, $\mathrm{MSE}(T) = \lambda^{-2}/(n+1)$.

We might choose $\bar{Y}$ as it is an unbiased estimator. We might choose T as it has a smaller MSE, $\mathrm{MSE}(T) < \mathrm{MSE}(\bar{Y})$.

sim. seen ⇓

6

(b) (i) Note that $Z = Y - \theta$ has an Exp(1) distribution.
For $x \geq 0$,

$$\begin{aligned} \mathrm{P}(X \geq x) = \mathrm{P}(Y_{(1)} - \theta \geq x/n) &= \mathrm{P}(Z_{(1)} \geq x/n) \\ &= \mathrm{P}(Z_i \geq x/n) \text{ (for } i = 1, \ldots, n) \\ &= (\exp(-x/n))^n \\ &= \exp(-x). \end{aligned}$$

Hence, $F_X(x) = 1 - e^{-x}$ and $f_X(x) = e^{-x}$ for $x \geq 0$.

7

(ii) The likelihood function is $L(\theta; y) = \exp(-\sum(y_i - \theta))$, for $y_{(1)} \geq \theta$ (and 0 otherwise). This is an increasing function of θ for $\theta \leq y_{(1)}$ and hence the MLE of θ is $y_{(1)}$.

Consistency can be shown as follows:

$$\begin{aligned} \forall \epsilon > 0: \ \mathrm{P}(|Y_{(1)} - \theta| \geq \epsilon) &= 1 - \mathrm{P}(-n\epsilon \leq X \leq n\epsilon) \\ &= 1 - \mathrm{P}(0 \leq X \leq n\epsilon) \\ &= e^{-n\epsilon} \to 0 \quad (n \to \infty) \end{aligned}$$

7

2. (a) $f_\theta(y) = \begin{cases} 1, & y \in [\theta, \theta+1] \\ 0, & \text{otherwise} \end{cases}$ sim. seen ⇓ 2

(b) $L : \mathbb{R} \to [0, \infty)$,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f_\theta(y_i) = \prod_{i=1}^{n} \mathrm{I}\left(\theta \le y_i \le \theta + 1\right) \\ &= \mathrm{I}\left(\theta \le \min(y_1, \ldots, y_n), \max(y_1, \ldots, y_n) \le \theta + 1\right) \\ &= \mathrm{I}\left(\max(y_1, \ldots, y_n) - 1 \le \theta \le \min(y_1, \ldots, y_n)\right). \end{aligned}$$

3

(c) Being an indicator function only, L is maximised iff $\max(y_1, \ldots, y_n) - 1 \le \theta \le \min(y_1, \ldots, y_n)$. Thus any $\widehat{\theta} \in [\max(y_1, \ldots, y_n) - 1, \min(y_1, \ldots, y_n)]$ is a maximum likelihood estimator. 2

(d) In this example, any $\widehat{\theta} \in [1.2 - 1, 0.3] = [0.2, 0.3]$ is a maximum likelihood estimate. 1

(e) For this observation, $\max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n) = 1.2 - 0.1 = 1.1$. This is impossible under the assumed model, implying that the data cannot have come from this model, i.e. the model must be wrong. 2

(f) T only depends on the observed parameter and the quantity of interest, so we only need to check if the distribution of T is completely know, i.e. does not depend on any parameter. For all $0 \le t \le 1$,

$$\mathrm{P}_\theta(T \le t) = \mathrm{P}_\theta(\max(Y_1, \ldots, Y_n) - \theta \le t) = \prod_{i=1}^{n} \mathrm{P}_\theta(Y_i - \theta \le t) = t^n.$$

For $t < 0$, $\mathrm{P}_\theta(T \le t) = 0$ and for $t > 1$, $\mathrm{P}_\theta(T \le t) = 1$. Thus the cdf of T does not depend on θ. Hence, T is a pivotal quantity for θ. 5

(g) We can use the pivotal quantity T for constructing a confidence interval.
We want c such that $\mathrm{P}_\theta(T \ge c) = 0.05$. This is equivalent to $c^n = 0.95$, implying $c = 0.95^{\frac{1}{n}}$. Using this,

$$0.95 = \mathrm{P}_\theta(\max(Y_1, \ldots, Y_n) - \theta \le c) \quad \forall \theta.$$

Hence,

$$0.95 = \mathrm{P}_\theta(\max(Y_1, \ldots, Y_n) - c \le \theta) \forall \theta.$$

Hence $[\max(Y_1, \ldots, Y_n) - c, \infty]$ is a confidence interval for θ, however this is not finite. Since $Y_1 \ge \theta$ implies $\min(Y_1, \ldots, Y_n) \ge \theta$, then we also have

$$0.95 = \mathrm{P}_\theta(\max(Y_1, \ldots, Y_n) - c \le \theta \le \min(Y_1, \ldots, Y_n))$$

This then gives a finite confidence interval

$$[\max(Y_1, \ldots, Y_n) - c, \min(Y_1, \ldots, Y_n)]$$

Note that any suitably constructed confidence interval is acceptable. 5

3. (a) (i) The vector of fitted values is $\widehat{\mathbf{Y}} = X\widehat{\beta}$, where $\widehat{\beta} = (X^TX)^{-1}X^T\mathbf{Y}$ is the least squares estimator. seen ⇓

(ii) The vector of residuals is $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$. 2

1

We can see that $\mathbf{Y} - X\widehat{\beta}$ is orthogonal to $X\widehat{\beta}$ since,

$$(X\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}}^T X^T(\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}}^T \underbrace{(X^T\boldsymbol{Y} - X^TX\widehat{\boldsymbol{\beta}})}_{=\mathbf{0} \text{ by LSE}} = 0$$

3

(b) RSS $= \mathbf{e}^T\mathbf{e} = \mathbf{Y}^TQ\mathbf{Y} = \mathbf{Y}^T\mathbf{Y} - \widehat{\mathbf{Y}}^T\widehat{\mathbf{Y}}$
Since, RSS $= \mathbf{e}^T\mathbf{e} = (Q\boldsymbol{Y})^TQ\boldsymbol{Y} = \boldsymbol{Y}^TQ^TQ\boldsymbol{Y} = \boldsymbol{Y}^TQ\boldsymbol{Y}$
and $\mathrm{RSS} = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\widehat{\boldsymbol{Y}}^T\boldsymbol{Y} + \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}} = \boldsymbol{Y}^T\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}}$,
where the last equality holds because
$\widehat{\boldsymbol{Y}}^T\boldsymbol{Y} = (P\boldsymbol{Y})^T\boldsymbol{Y} = (PP\boldsymbol{Y})^T\boldsymbol{Y} = \boldsymbol{Y}^TP^TP^T\boldsymbol{Y} = (P\boldsymbol{Y})^TP\boldsymbol{Y} = \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}}$. 4

(c) From Part (b), we know that $\mathrm{RSS} = \boldsymbol{Y}^TQ\boldsymbol{Y}$, and so

$$\begin{aligned}
\mathrm{E}(\mathrm{RSS}) &= \mathrm{E}\,\mathrm{trace}\,\mathrm{RSS} = \mathrm{E}\,\mathrm{trace}(\boldsymbol{Y}^TQ\boldsymbol{Y}) = \mathrm{E}\,\mathrm{trace}(Q\boldsymbol{Y}\boldsymbol{Y}^T) = \mathrm{trace}(Q\,\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^T)) \\
&= \mathrm{trace}(Q[\mathrm{cov}\,\boldsymbol{Y} + \mathrm{E}(\boldsymbol{Y})\,\mathrm{E}(\boldsymbol{Y})^T]) \\
&= \mathrm{trace}(Q\sigma^2) + \mathrm{trace}(QX\boldsymbol{\beta}(X\boldsymbol{\beta})^T) \\
&= \sigma^2\,\mathrm{trace}(I - P) + 0 = \sigma^2(n - \mathrm{trace}(P)) \overset{*}{=} \sigma^2(n - \mathrm{rank}(P)) \\
&= \sigma^2(n - r).
\end{aligned}$$

where at $*$ we use the fact that for a projection matrix P, rank(P) = trace(P). 5

(d) We can construct the linear model $\boldsymbol{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \boldsymbol{\epsilon}$ with $\mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$ and $\mathrm{cov}\,\boldsymbol{\epsilon} = \sigma^2 I$.

Then $P = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$, thus $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - P\boldsymbol{Y} = \boldsymbol{Y} - \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$.

Hence,

$$\frac{\mathrm{RSS}}{n-r} = \underbrace{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}_{=s^2=\text{sample variance}}.$$

5

4. (a) In order to draw a sample F from an $F_{n_1,n_2}(\delta)$ distribution, we need to sample from $W_1 \sim \chi^2_{n_1}(\delta)$, $W_2 \sim \chi^2_{n_2}$ independently, then calculate

sim. seen ⇓

$$F = \frac{W_1/n_1}{W_2/n_2}$$

which is a non-central F distribution with (n_1, n_2) d.f. and n.c.p.=δ.
In order to sample W from a $\chi^2_n(\delta)$ distribution, we need to sample $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, I_n)$, such that $\delta = \sqrt{\boldsymbol{\mu}^T\boldsymbol{\mu}}$, then calculate

$$W = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{n} Z_i^2$$

which is a χ^2-distribution with n d.f. and n.c.p.=δ.
Finally we note that we can instead sample n times from the univariate Normal function we are given, by calculating a suitable value of μ as $\mu = \sqrt{\frac{\delta^2}{n}}$, i.e. the vector $\boldsymbol{\mu}$ has all elements the same.
We can therefore use this algorithm with $n_1 = n_2 = 2$ and $\delta = 4$.

5

(b) Since we can write the 3rd angle in terms of π and the other 2 angles, we can construct a better linear model as follows

unseen ⇓

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 - \pi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

The least squares estimates of the unknown angles are therefore

$$\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = (X^TX)^{-1}X^T\mathbf{Y} = \frac{1}{3}\begin{pmatrix} 2Y_1 - Y_2 - Y_3 + \pi \\ -Y_1 + 2Y_2 - Y_3 + \pi \end{pmatrix}$$

where

$$(X^TX)^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},$$

with covariance matrix

$$\text{Var}\begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \frac{\sigma^2}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

and $\widehat{\sigma}^2 = (Y_1 + Y_2 + Y_3 - \pi)^2/3$. Hence we can see that the variance of the parameters of our new model is 2/3 that of the naive model originally suggested.

6

(c) This is quite an open question, so reasonable examples will be acceptable.

3

(i) Suitable description of an example and interpretation of the corresponding hypotheses. For example, from the notes we have an example of modelling n mammals, such that Y_i = log(brain weight of i'th mammal) using the covariate o_i = log(body weight of i'th mammal). We could then test the hypothesis that there is not a linear relationship between the log brain weight and log body weight of our collection of mammals.

sim. seen ⇓

3

(ii) Firstly, we should make the normal theory assumption, i.e. the assumption that the error satisfies $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$. seen ⇓

Marks for this question should be given with a particular emphasis on the clarity of the answer. 1

The F-test can be used to test the hypotheses

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0.$$

The null hypothesis can be written as

$$\mathrm{E}\,\boldsymbol{Y} = \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{=:X_0} \beta_1$$

1

The test statistic of the F-test is

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}} \frac{n-r}{r-s}$$

where $r = \operatorname{rank} X = 2$, $s = \operatorname{rank} X_0 = 1$. RSS_0 is the residual sum of squares under H_0 (which in this case can be worked out to be $\sum_{i=1}^n (Y_i - \bar{Y})^2$). RSS is the residual sum of squares under the full model. 2

Under H_0, the test statistics satisfies $F \sim F_{r-s,n-r}$ and H_0 is rejected for large values of F.

So, for a level α test *(candidates can also pick a values for α)*, we reject H_0 if $F > c$ where c is s.t. $\mathrm{P}(Z > c) = \alpha$ for $Z \sim F_{r-s,n-r}$.

Note that the equivalent t-test could also be used as an acceptable answer. 2