Imperial College London

M2S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2016

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

Statistical Modelling I

Date: **Thursday 12th May 2016**

Time: **14.00 – 16.00**

Time Allowed: **2 Hours**

This paper has Four 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.

Statistical tables will not 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. (a) Suppose we observe some data $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, where each y_i is an observed realisation from the random variable Y_i. Define the following terms, introducing notation where appropriate:

   (i) statistical model,

   (ii) random sample,

   (iii) estimator,

   (iv) covariates.

   (b) Consider a sequence of estimators $(T_n)_{n\in\mathbb{N}}$. Write down the definition for each of the following concepts.

   (i) $(T_n)_{n\in\mathbb{N}}$ is consistent,

   (ii) $(T_n)_{n\in\mathbb{N}}$ is asymptotically unbiased.

   (c) Prove that if the sequence of estimators $(T_n)_{n\in\mathbb{N}}$ is asymptotically unbiased for $g(\theta)$, and $\mathrm{Var}_\theta(T_n) \to 0$ $(n \to \infty)$, $\forall\theta \in \Theta$, then $(T_n)_{n\in\mathbb{N}}$ is consistent for $g(\theta)$.

   Recall: Markov's inequality, $p(X \geq a) \leq \frac{E(X)}{a}$, *for a non-negative random variable* X *and* $a > 0$, *which you can use without proof.*

   (d) Consider a random sample $Y_1, \ldots, Y_n \sim N(\theta, 1)$, $\theta \in \mathbb{R}$. Propose a consistent and normally distributed estimator for θ and prove that your estimator does indeed have these two properties.

2. (a) State the Cramer-Rao inequality and explain its importance.

   (b) Suppose a random sample $(x_1, ..., x_n)$ is observed. Show that the Fisher Information is proportional to the sample size, n.

   (c) Give two different expressions for the expected Fisher Information and sketch a proof that they are equivalent.

   (d) Suppose a random sample $(x_1, \ldots, x_n)$ is observed, where x_i is a realisation of $X_i \sim N(\mu, \theta)$, and where the mean μ is given and the variance θ is unknown. Calculate a lower bound on the variance of an estimator using the Cramer-Rao inequality.

3. (a) Consider a statistical model parameterised by a parameter θ and with data $\mathbf{Y}$. Define the following terms:

   (i) pivotal quantity,

   (ii) confidence interval.

   (b) Describe how a pivotal quantity may be used to obtain a confidence interval.

   (c) Consider a random sample $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. We wish to estimate the mean, μ. Suggest an appropriate pivotal quantity and derive an expression for a $(1-\alpha)$ confidence interval. State the distribution of the pivotal quantity without proof.

   (d) Let $\theta \in \Theta$ be an unknown parameter and let $\mathbf{y}$ denote the observed data. Consider the null hypothesis $H_0 : \theta \in \Theta_0$ and the alternative hypothesis $H_1 : \theta \in \Theta_1 := \Theta/\Theta_0$. Define the likelihood ratio test statistic and describe how it may be used to accept or reject the null hypothesis, H_0.

4. (a) A full-rank linear model $\mathrm{E}\boldsymbol{Y} = X\boldsymbol{\beta}$ (with standard second-order assumptions) is fitted and $\widehat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$. The same model, but with an extra covariate added, is then fitted to the data, i.e. the new model is $\mathrm{E}\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{x}\gamma$, where γ is a new unknown parameter and $\boldsymbol{x}$ is a known $n \times 1$ vector which is linearly independent of the columns of X. The least squares estimate of $\boldsymbol{\beta}$ in this new model is $\widehat{\boldsymbol{\beta}}_N$.

   (i) Show that
$$\widehat{\boldsymbol{\beta}}_N = \widehat{\boldsymbol{\beta}} - \widehat{\gamma}(X^TX)^{-1}X^T\boldsymbol{x},$$
where $\widehat{\gamma} = (\boldsymbol{x}^TA\boldsymbol{Y})/(\boldsymbol{x}^TA\boldsymbol{x})$, $A = I_n - X(X^TX)^{-1}X^T$ and I_n is the $n \times n$ identity matrix.

   (ii) Show that $A = I_n - X(X^TX)^{-1}X^T$ is a projection matrix.

   (b) Prove that if $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, I_n)$ and A is an $n \times n$ projection matrix of rank r, then
$$\boldsymbol{Z}^TA\boldsymbol{Z} \sim \chi_r^2(\delta) \quad \text{with } \delta^2 = \boldsymbol{\mu}^TA\boldsymbol{\mu}$$
You may assume that for a positive semi-definite matrix A with rank r, there exists a matrix L such that $A = LL^T$ and $L^TL = I_r$.

   (c) Consider a linear model with Normal theory assumptions. Making use of the result in part (b), prove that
$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi_{n-r}^2$$
where $r = \operatorname{rank} X$.

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2016

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) $\boxed{\text{seen} \Downarrow}$

(i) Given observed data $\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}$, we define the random variable version, $\mathbf{Y} = (Y_1, ..., Y_n)$ a random vector, where y_i is a realisation of Y_i. A statistical model is a specification of the joint distribution of $\mathbf{Y}$ up to an unknown parameter θ. $\boxed{3}$

(ii) If $Y_1, ..., Y_n$ are independent and identically distributed (i.i.d.) then $Y_1, ..., Y_n$ is called a random sample. $\boxed{1}$

(iii) Let $t(y_1, ..., y_n)$ be a function of the observed data, then its random variable version $T \equiv t(Y_1, ..., Y_n)$ is called an estimator of θ. $\boxed{2}$

(iv) Covariates are non-random/deterministic quantities that $Y_1, ..., Y_n$ may depend on. $\boxed{1}$

(b) (i) A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called consistent if for all $\theta \in \Theta$, $T_n \to^{P_\theta} g(\theta)$ $(n \to \infty)$. $\boxed{2}$

(ii) A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called asymptotically unbiased if $E_\theta(T_n) \to \theta$ $(n \to \infty)$, $\forall \theta \in \Theta$. $\boxed{1}$

(c) *This is bookwork from the lecture notes.*

$$\begin{aligned}
\forall \epsilon > 0 : p_\theta(|T_n - g(\theta)| \geq \epsilon) &= p_\theta\left((T_n - g(\theta))^2 \geq \epsilon^2\right) \\
&\leq \frac{E_\theta(T_n - g(\theta))^2}{\epsilon^2} \\
&= \frac{\mathsf{MSE}_\theta(T_n)}{\epsilon^2} \\
&= \frac{1}{\epsilon^2}\left[\mathrm{Var}_\theta(T_n) + [E_\theta(T_n - g(\theta))]^2\right] \\
&\to 0\ ,\ (n \to \infty)
\end{aligned}$$

$\boxed{5}$

(d) Consider the estimator $T_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then $\boxed{\text{sim. seen} \Downarrow}$

$$E(T_n) = E\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \theta$$

$$\mathrm{Var}(T_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Y_i) = \frac{1}{n}$$

As the sum of independent normally distributed random variables is also normally distributed, this implies $T_n \sim N(\theta, \frac{1}{n})$. Hence, since the proposed estimator is asymptotically unbiased and $\mathrm{Var}_\theta(T_n) \to 0$, $(n \to \infty)$, then (T_n) is a consistent estimator, by part (c). $\boxed{5}$

2. (a) *This part appeared in the lecture notes.* $\boxed{\text{seen} \Downarrow}$

Let X be the observed data, and suppose $T = T(X)$ is an unbiased estimator for θ. The Cramer-Rao lower bound gives the following lower bound on the variance of any such estimator T,

$$\mathrm{Var}_\theta(T) \geq \frac{1}{I_f(\theta)}$$

where $I_f(\theta) = E_\theta[(\frac{d}{d\theta}\log f_\theta(X))^2]$ is the expected Fisher Information and f_θ is the joint density of X. $\boxed{4}$

(b) *This part appeared in the lecture notes.*

For a random sample $\mathbf{x} = (x_1, \ldots, x_n)$, the joint pdf is $f_\theta(x) = \prod_{i=1}^n f_\theta^{(1)}(x_i)$, where $f_\theta^{(1)}$ is the pdf/pmf of a single observation. This implies

$$I_f(\theta) = -E_\theta\left[\left(\frac{\partial}{\partial\theta}\right)^2 \log f_\theta(X)\right] = \sum_{i=1}^n -E_\theta\left[\left(\frac{\partial}{\partial\theta}\right)^2 \log f_\theta^{(1)}(X_i)\right] = nI_{f^1}(\theta)$$

$\boxed{4}$

(c) *This part appeared in the lecture notes.*
We wish to show that $E_\theta[(\frac{\partial}{\partial\theta}\log f_\theta(X))^2] = -E_\theta[(\frac{\partial}{\partial\theta})^2 \log f_\theta(X)]$.
Let us denote $f'_\theta = \frac{\partial}{\partial\theta} f_\theta$ and $f''_\theta = \frac{\partial}{\partial\theta} f'_\theta$. Then

$$\begin{aligned} E_\theta\left[(\frac{\partial}{\partial\theta})^2 \log f_\theta(X)\right] &= E_\theta\left[\frac{\partial}{\partial\theta}\frac{f'_\theta(X)}{f_\theta(X)}\right] \\ &= E_\theta\left[-\frac{f'_\theta(X)}{f_\theta^2(X)} f'_\theta(X) + \frac{f''_\theta(X)}{f_\theta(X)}\right] \\ &= E_\theta\left[-\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] + E_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)}\right]. \end{aligned}$$

We complete the proof by observing that

$$E_\theta\left[\frac{f''_\theta(X)}{f_\theta(X)}\right] = \int \frac{f''_\theta(x)}{f_\theta(x)} f_\theta(x)dx = \int f''_\theta(x)dx = \left(\frac{\partial}{\partial\theta}\right)^2 \int f_\theta(x)dx = 0.$$

$\boxed{6}$

(d)

unseen $\Downarrow$

$$\begin{aligned}
f(x|\theta) &= \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x-\mu)^2}{2\theta}\right\} \\
\therefore l(x|\theta) \equiv \log f(x|\theta) &= -\frac{(x-\mu)^2}{2\theta} - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\theta \\
\therefore l'(x|\theta) &= \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta} \\
\therefore l''(x|\theta) &= -\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2} \\
\therefore I(\theta) &= -E\left(l''(x|\theta)\right) = -E\left(-\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}\right) = \frac{1}{2\theta^2} \\
\therefore I_n(\theta) &= nI(\theta) = \frac{n}{2\theta^2}
\end{aligned}$$

The Cramer-Rao lower bound on the variance of such unbiased estimators is therefore $\frac{2\theta^2}{n}$.

$\boxed{6}$

3. (a) (i) Consider a statistical model with a single unknown parameter θ and data $\mathbf{Y}$. A pivotal quantity for θ is a function $t(\mathbf{Y}, \theta)$ of the data and θ, such that the distribution of $t(\mathbf{Y}, \theta)$ is known and does not depend on any further unknown parameters. seen $\Downarrow$ 2

(ii) A $(1-\alpha)$ confidence interval for θ is a random interval I that contains the "true" parameter with probability $\geq 1-\alpha$, i.e. $P_\theta(\theta \in I) \geq 1-\alpha$, $\forall \theta \in \Theta$. 2

(b) Suppose $t(\mathbf{Y}, \theta)$ is a pivotal quantity for θ. Then we can find constants a_1 and a_2 such that,

$$p(a_1 \leq t(\mathbf{Y}, \theta) \leq a_2) \geq 1-\alpha$$

since we know the distribution of $t(\mathbf{Y}, \theta)$. In many case we can rearrange the terms to obtain the form,

$$p(h_1(\mathbf{Y}) \leq \theta \leq h_2(\mathbf{Y})) \geq 1-\alpha$$

Then $[h_1(\mathbf{Y}), h_2(\mathbf{Y})]$ is a random interval and the observed interval $[h_1(\mathbf{y}), h_2(\mathbf{y})]$ is a $(1-\alpha)$ confidence interval for θ. 6

(c) We can't use $\frac{\bar{\mathbf{Y}}-\mu}{\sigma/\sqrt{n}}$ as a pivotal quantity, where $\bar{\mathbf{Y}}$ denotes the sample mean, since σ is also unknown, however we can replace σ^2 by the sample variance S^2, where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2$$

to give the statistic $T = \frac{\sqrt{n}}{S}(\bar{\mathbf{Y}} - \mu) = \frac{(\bar{\mathbf{Y}}-\mu)}{\sqrt{S^2/n}}$, which follows a Student-t distribution.

We may then consider,

$$\begin{aligned} 1-\alpha &= p(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) \\ &= p(\bar{\mathbf{Y}} - \frac{S}{\sqrt{n}}t_{\alpha/2} \leq \mu \leq \bar{\mathbf{Y}} + \frac{S}{\sqrt{n}}t_{\alpha/2}) \end{aligned}$$

to obtain the $(1-\alpha)$ confidence interval $(\bar{\mathbf{y}} - \frac{S}{\sqrt{n}}t_{\alpha/2}, \bar{\mathbf{y}} + \frac{S}{\sqrt{n}}t_{\alpha/2})$. 6

(d) Suppose we observe data $\mathbf{y}$, then the likelihood ratio test statistic is defined by comparing the maximised likelihood under H_0 to the unrestricted maximum likelihood, as follows

$$t(\mathbf{y}) = \frac{\sup_{\theta \in \Theta} L(\theta; \mathbf{y})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{y})}$$

We note that $t(\mathbf{y}) \geq 1$ and if $t(\mathbf{y})$ is large, this will indicate support for H_1. Therefore we reject H_0 when $t(\mathbf{y}) \geq k$, where k is chosen such that

$$\sup_{\theta \in \Theta_0} p_\theta(t(\mathbf{Y}) \geq k) \leq \alpha$$

which gives us an α level test. 4

4. (a) unseen $\Downarrow$

(i) The new model is $\mathrm{E}\boldsymbol{Y} = (X \quad \boldsymbol{x})\begin{pmatrix}\boldsymbol{\beta}\\ \gamma\end{pmatrix}$. This is also of full rank because $\boldsymbol{x}$ is linearly independent of the columns of X. The new least squares equations are

$$\begin{pmatrix} X^TX & X^T\boldsymbol{x} \\ \boldsymbol{x}^TX & \boldsymbol{x}^T\boldsymbol{x}\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_N \\ \widehat{\gamma}\end{pmatrix} = \begin{pmatrix}X^T \\ \boldsymbol{x}^T\end{pmatrix}\boldsymbol{Y}.$$

Now write this out as 2 (vector) equations. Rearranging the first of these gives

$$\widehat{\boldsymbol{\beta}}_N = \widehat{\boldsymbol{\beta}} - \widehat{\gamma}(X^TX)^{-1}X^T\boldsymbol{x}$$

and substituting into the second equation gives

$$\widehat{\gamma} = \frac{\boldsymbol{x}^TA\boldsymbol{Y}}{\boldsymbol{x}^TA\boldsymbol{x}}.$$

(ii) We show that $A = I_n - X^T(X^TX)^{-1}X$:
is symmetric, i.e. $A = A^T$, (1 mark) and
is idempotent, i.e. $A^2 = A$ (2 marks).

6

seen $\Downarrow$

3

(b) All nonzero eigenvalues of A are equal to 1.
$\exists\ L \in \mathbb{R}^{n\times r}$ such that $A = LL^T$ and $L^TL = I_r$. Let $\boldsymbol{V} = L^T\boldsymbol{Z}$. Then $\boldsymbol{V} \sim N(L^T\boldsymbol{\mu}, \underbrace{I}_{=L^TL})$ and

$$\boldsymbol{Z}^TA\boldsymbol{Z} = \boldsymbol{Z}^TLL^T\boldsymbol{Z} = \boldsymbol{V}^T\boldsymbol{V} \sim \chi^2_r(\delta),$$

where $\delta^2 = \left(L^T\boldsymbol{\mu}\right)^T L^T\boldsymbol{\mu} = \boldsymbol{\mu}^T\underbrace{LL^T}_{=A}\boldsymbol{\mu} = \boldsymbol{\mu}^TA\boldsymbol{\mu}.$

5

(c) Let P denote the projection matrix onto $\operatorname{span}(X)$. Then

$$\mathrm{RSS} = \boldsymbol{e}^T\boldsymbol{e} = (\underbrace{(I-P)}_{=:Q}\boldsymbol{Y})^T(I-P)\boldsymbol{Y} = \boldsymbol{Y}^T\underbrace{Q^TQ}_{=QQ=Q}\boldsymbol{Y} = \boldsymbol{Y}^TQ\boldsymbol{Y}$$

and Q is the projection onto the space orthogonal to the columns of X. Hence,

$$\frac{\mathrm{RSS}}{\sigma^2} = \frac{\boldsymbol{Y}}{\sigma}^T Q\frac{\boldsymbol{Y}}{\sigma} = \boldsymbol{Z}^TQ\boldsymbol{Z}$$

where $\boldsymbol{Z}\boldsymbol{Y}/\sigma \sim N(X\boldsymbol{\beta}/\sigma, I)$ and Q is a projection matrix.
Furthermore, $Q + P = I$ and Q and P are projection matrices. Thus $\operatorname{rank} Q + \operatorname{rank} P = n$, implying $\operatorname{rank} Q = n - r$.
Thus, by the result in part (b),

$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-r}$$

since $(X\boldsymbol{\beta}/\sigma)^T\underbrace{QX}_{=0}\boldsymbol{\beta}/\sigma = 0.$

6