Imperial College
London

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May – June 2015

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

Statistical Modelling I

Date: Wednesday, 20 May 2015. Time: 2.00pm – 4.00pm. 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

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	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	$2\frac{1}{2}$	3	$3\frac{1}{2}$	4

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

1. Consider the model $X \sim \text{Binomial}(n, p)$, where n is a known positive integer and $p \in [0, 1]$ is an unknown parameter.
Recall: The probability mass function (pmf) of a random variable $Z \sim Binomial(n,p)$ is $f(z) = \binom{n}{z} p^z (1-p)^{n-z}$, where $z \in \{0, 1, ..., n\}$. $E(Z) = np$ and $Var(Z) = np(1-p)$.

   (a) Consider $S = \frac{X}{n}$ as an estimator for the unknown parameter p.
      (i) Define the bias of an estimator.
      (ii) Show that S is an unbiased estimator for p.
      (iii) Calculate the variance of the estimator S, and hence write down the mean squared error of S.

   (b) Now consider an alternative estimator, $T = \frac{X+1}{n+2}$.
      (i) Calculate the bias of T.
      (ii) Calculate the variance of T.

   (c) Compare the mean squared error of these two estimators, S and T. Comment on their relative performance for different values of p.

2. Let $Y_1, ..., Y_n \sim \text{Exp}(\lambda)$ independently for 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 $z > 0$, $\lambda > 0$.

   (a) Derive the maximum likelihood estimator for λ.

   (b) Calculate the large sample properties of this maximum likelihood estimator.

   (c) Write down an asymptotic 95% confidence interval for λ.

   (d) We could also employ a Bayesian approach for estimating λ, in which case we must define a prior distribution for the unknown parameter. It is often convenient to choose one that is conjugate to the likelihood.
      (i) Define the term *conjugate prior*.
      (ii) Show that the gamma distribution is a conjugate prior for an exponential likelihood.
      Recall: The probability density function (pdf) of a random variable $Z \sim Gamma(\alpha, \beta)$ is $f(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp(-\beta z)$ for $z > 0$, $\alpha > 0$, $\beta > 0$.

3. (a) Write down the general form of a linear model (using matrix notation) and fully describe each term in the equation.

(b) The least squares estimator for a linear model is given by $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(i) Write down an expression for the vector of fitted values, $\widehat{\mathbf{Y}}$.

(ii) Define the term *projection matrix*.

(iii) Show how the vector of fitted values may be written as a projection of the original data vector, $\mathbf{Y}$, and prove that this matrix does indeed satisfy the required properties of a projection matrix.

(c) (i) State the Gauss Markov Theorem.

(ii) Give a specific and concrete example of a modelling problem that could be tackled with a linear model. State clearly what you are trying to estimate and describe whether or not the Gauss Markov Theorem would influence your choice of estimator.

4. (a) Define the non-central t-distribution.

(b) In this part we consider the linear model you defined in question 3 part (a) and assume the Normal Theory Assumptions.

(i) Calculate the expected value and variance of the maximum likelihood estimator of some linear combination of the parameters, i.e. $\mathrm{E}(\mathbf{c}^T\widehat{\beta})$ and $\mathrm{Var}(\mathbf{c}^T\widehat{\beta})$, for some deterministic column vector $\mathbf{c}$.

(ii) Show that $\mathrm{RSS} = \mathbf{Y}^T\mathbf{Q}\mathbf{Y}$.
Recall: RSS is the residual sum of squares, $\mathbf{Y}$ is the measurement vector and $\mathbf{Q}$ is the projection onto the complement of the space spanned by the columns of the design matrix.

(iii) State what distribution the following test statistic has, and sketch out how you would prove this.

$$\frac{\mathbf{c}^T\widehat{\beta} - \mathbf{c}^T\beta}{\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\frac{\mathrm{RSS}}{n-p}}}$$

Recall: $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$, where p is the rank of the design matrix.

(c) Define the non-central F-distribution.

(d) Explain how the Fisher-Cochran Theorem can be used to prove that a test statistic has a non-central F distribution.

Imperial College London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2015

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) $\text{bias}_\theta(S) = \mathrm{E}_\theta(S) - \theta$. $\boxed{2}$

(ii) $\forall p, \text{bias}_p(S) = \mathsf{E}_p(S - p) = \frac{1}{n}\mathsf{E}(X) - p = 0$. Thus, S is unbiased for p. $\boxed{2}$

(iii) $\mathsf{Var}_p(S) = \frac{1}{n^2}\mathsf{Var}_p(X) = \frac{p(1-p)}{n}$

$\mathsf{MSE}_p(S) = \mathsf{Var}(S) + \text{bias}(S)^2 = \frac{p(1-p)}{n}$ $\boxed{4}$

(b) *This is an example in the lecture notes.*

(i) $\text{bias}_p(T) = \mathsf{E}_p(T - p) = \frac{\mathsf{E}_p(X)+1}{n+2} - p = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2} \neq 0$ $\boxed{3}$

(ii) $\mathsf{Var}_p(T) = \frac{1}{(n+2)^2}\mathsf{Var}_p(X) = \frac{np(1-p)}{(n+2)^2}$ $\boxed{3}$

(c) *This is an example in the lecture notes.*

$\mathsf{MSE}_p(T) = \mathsf{Var}_p(T) + \text{bias}_p(T)^2 = \frac{np(1-p)}{(n+2)^2} + \frac{(1-2p)^2}{(n+2)^2}$

For $p = 0$ and $p = 1$, $\mathsf{MSE}_p(T) = \frac{1}{(n+2)^2} > 0 = \mathsf{MSE}_p(S)$.

But for $p = 0.5$, $\mathsf{MSE}_p(T) = \frac{n}{4(n+2)^2} < \frac{n}{4n^2} = \frac{1}{4n} = \mathsf{MSE}_p(S)$.

The performance of the estimator therefore depends on the true value of p.

Sometimes a biased estimator gives a better estimate than an unbiased estimator, according to the MSE criterion. $\boxed{6}$

2. (a) *This part appeared in the lecture notes.* sim. seen ⇓

The likelihood is $p(\mathbf{Y}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda \mathbf{y}_i)$.
The log-likelihood is $\log p(\mathbf{Y}|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} \mathbf{y}_i$.
The derivative of the log-likelihood follows as $\frac{d}{d\lambda} \log p(\mathbf{Y}|\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} \mathbf{y}_i$.
Setting this expression for the derivative equal to zero implies that $\lambda_{MLE} = \frac{n}{\sum_{i=1}^{n} \mathbf{y}_i}$.
We then confirm it is a maximum by showing that the Hessian is always negative, $\frac{d^2}{d\lambda^2} \log p(\mathbf{Y}|\lambda) = -\frac{n}{\lambda^2} \leq 0$. 4

(b) *This part is similar to an example in the lecture notes.*
The Fisher Information is $-\mathsf{E}\left(\frac{d^2}{d\lambda^2} \log p(\mathbf{Y}|\lambda)\right) = -\mathsf{E}(-\frac{n}{\lambda^2}) = \frac{n}{\lambda^2}$.
If λ_0 is the "true" parameter, then $\sqrt{n}(\lambda_{MLE} - \lambda_0) \to^d N(0, \lambda_0^2)$. 4

(c) *This part is similar to an example in the lecture notes.*
From the previous part we can conclude that, $Pr\left(c_1 < \frac{\sqrt{n}(\lambda_{MLE} - \lambda_0)}{\lambda_0} < c_2\right) = 1 - \alpha$.
Choosing $\alpha = 0.05$, c_1 such that $\Phi(c_1) = \frac{\alpha}{2}$, and c_2 such that $\Phi(c_2) = 1 - \frac{\alpha}{2}$, then rearranging the inequality results in $\frac{c_1/\sqrt{n}+1}{\lambda_{MLE}} < \frac{1}{\lambda_0} < \frac{c_2/\sqrt{n}+1}{\lambda_{MLE}}$, and so the random interval is given by $\left(\frac{\lambda_{MLE}}{c_2/\sqrt{n}+1}, \frac{\lambda_{MLE}}{c_1/\sqrt{n}+1}\right)$ 6

(d) (i) A family of prior probability distributions P is said to be conjugate to a family of observational distributions L, if for every prior $p \in P$ and every observational distribution $l \in L$, the resulting posterior distribution also belongs to P. 2

(ii) $p(y_1|\lambda)p(\lambda) = \lambda \exp(-\lambda y_1) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \propto \lambda^{(\alpha+1)-1} \exp(-\lambda(\beta + y_1))$ which is also gamma distributed with parameters $\alpha_{new} = \alpha + 1$ and $\beta_{new} = \beta + y_1$. 4

3. (a) $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ seen ⇓

$\mathbf{Y}$ is an $n \times 1$ vector of observations.

$\mathbf{X}$ is an $n \times p$ design matrix.

β is an $p \times 1$ vector of parameters.

ϵ is an $n \times 1$ vector of random variables describing the error. $\boxed{4}$

(b) *(Seen before in class)*

(i) $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ $\boxed{2}$

(ii) Let L be a linear subspace of $\mathbb{R}^n$, $\dim L = r \leq n$. $P \in \mathbb{R}^{n \times n}$ is a projection matrix onto L, if

1. $P\boldsymbol{x} = \boldsymbol{x} \quad \forall \boldsymbol{x} \in L$
2. $P\boldsymbol{x} = \mathbf{0} \quad \forall \boldsymbol{x} \in L^{\perp} = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^T\boldsymbol{y} = 0 \, \forall \boldsymbol{y} \in L\}$

Alternatively, candidates can define a projection matrix as follows.

Let $A \in \mathbb{R}^{n \times n}$. A is called a projection matrix if it is symmetric ($A^T = A$) and idempotent ($AA = A$). $\boxed{3}$

(iii) $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = P\mathbf{Y}$, where $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a projection matrix, since $P^T = P$ and $P^2 = P$, both of which should be shown algebraically. $\boxed{3}$

(c) (i) Let $c \in \mathbb{R}^p$ and let $\widehat{\beta}$ be a least squares estimator of β in a linear model, where we assume full rank and second order assumptions. Then the estimator $c^T\widehat{\beta}$ has the smallest variance among all linear unbiased estimators for $c^T\beta$. $\boxed{3}$

(ii) This is an open question and any reasonable description of a linear model, where we are interested in estimating some linear combination of parameters, i.e. $c^T\beta$, is acceptable. unseen ⇓

If we want an unbiased estimator for $c^T\beta$, then the Gauss Markov theorem says that we should use $c^T\widehat{\beta}$, as per part (i).

However, we may be able to find a biased estimator with lower variance, and hence lower MSE, in which case we might choose to ignore the Gauss Markov theorem. $\boxed{5}$

4. (a) If $X \sim N(\delta, 1)$, and $U \sim \chi^2_n$ independently then $\boxed{\text{seen} \Downarrow}$

$$Y = \frac{X}{\sqrt{U/n}}$$

is said to have a non-central t-distribution with n d.f. and n.c.p.$=\delta$.

$\boxed{3}$

(b) (i) Since $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$, we have that

$$\mathsf{E}(c^T\boldsymbol{\beta}) = \mathsf{E}(c^T(X^TX)^{-1}X^TY) = c^T(X^TX)^{-1}X^TX\boldsymbol{\beta} = c^T\boldsymbol{\beta}$$

$$\begin{aligned}\mathsf{Var}(c^T\boldsymbol{\beta}) &= \mathsf{Var}(c^T(X^TX)^{-1}X^TY) \\ &= c^T(X^TX)^{-1}X^T\mathsf{Cov}(Y)X(X^TX)^{-1}c \\ &= c^T(X^TX)^{-1}c\sigma^2\end{aligned}$$

$\boxed{3}$

(ii)

$$\begin{aligned}\mathsf{RSS} &= e^Te \\ &= ((I-P)Y)^T((I-P)Y) \\ &= Y^TQ^TQY \\ &= Y^TQY\end{aligned}$$

$\boxed{2}$

(iii) This statistic is t_{n-p}-distributed. From part (i) we know $c^T\widehat{\beta} \sim N(c^T\beta, c^T(X^TX)^{-1}c\sigma^2)$, and so $\boxed{\text{sim. seen} \Downarrow}$

$$A = \frac{c^T\widehat{\beta} - c^T\beta}{\sqrt{c^T(X^TX)^{-1}c\sigma^2}} \sim N(0,1)$$

Let $B = \frac{\mathsf{RSS}}{\sigma^2} \sim \chi^2_{n-p}$. We can first prove that A and B are independent, then use the fact from part (a) that $\frac{A}{\sqrt{B/n}} \sim t_n$. $\boxed{4}$

(c) If $W_1 \sim \chi^2_{n_1}(\delta)$, $W_2 \sim \chi^2_{n_2}$ independently then

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

is said to have a non-central F distribution with (n_1, n_2) d.f. and n.c.p.$=\delta$. $\boxed{3}$

(d) The Fisher-Cochran theorem states that if $A_1, \ldots, A_k$ are $n \times n$ projection matrices such that $\sum_{i=1}^{n} A_i = I_n$, and if $Z \sim N(\mu, I_n)$, then $Z^T A_1 Z, \ldots, Z^T A_k Z$ are independent and $\boxed{\text{seen} \Downarrow}$

$$Z^T A_i Z \sim \chi^2_{r_i}(\delta_i), \quad \text{where } r_i = \operatorname{rank} A_i \text{ and } \delta_i^2 = \mu^T A_i \mu.$$

If a test statistic can be written in the form defined in part (c), then once we have shown independence of the two chi squared distributions using the Fisher-Cochran theorem, we can conclude that the test statistic is F distributed. $\boxed{5}$