Imperial College London

Course: M2S2
Setter:
Checker:
Editor:
External:
Date: April 23, 2014

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2014

M2S2

Statistical Modelling I

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

Imperial College London

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2014

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

M2S2

Statistical Modelling I

Date: Monday, 12 May 2014 Time: 14:00-16:00

Credit will be given for all questions attempted but extra credit will be given for complete or nearly complete answers.

Calculators may not be used.

1. (a) Suppose $\theta \in \mathbb{R}$ is the unknown parameter in a statistical model and suppose that T is an estimator for θ.

   (i) Define the bias and the mean squared error of T.

   (ii) How are bias, variance and mean squared error related? Prove this relationship.

   (b) In a Bayesian model, suppose that $\theta \sim N(0,1)$, and that $X_1, \ldots, X_n \sim N(\theta, 1)$ are independent given θ.

   Compute the posterior distribution of θ given $X_1, \ldots, X_n$.

   (c) Suppose $X_1, \ldots, X_n$ are i.i.d. and follow a uniform distribution on the interval $[a,b]$ where $a, b \in \mathbb{R}, a < b$ are unknown.
   Derive the maximum likelihood estimator for the parameter vector (a,b).

2. You are standing at a bus stop close to the Royal Albert Hall. You notice that there are several bus lines passing by and you begin to wonder if two of those, the number 9 and the number 10, have the same frequency. To investigate this you record the time between the arrival of busses for both lines.

   Suppose you observe n_1 inter-arrival times for line number 9 and n_2 inter-arrival times for line number 10. You decide to let $y_{11}, \ldots, y_{1n_1}$ denote the inter-arrival times for line number 9 and to let $y_{21}, \ldots, y_{2n_2}$ denote the inter-arrival times for line number 10.

   You decide to work with the model

$$Y_{11}, \ldots, Y_{1n_1} \sim \text{Exp}(\lambda_1), \quad Y_{21}, \ldots, Y_{2n_2} \sim \text{Exp}(\lambda_2),$$

   where all random variables are independent and the unknown parameters are $\lambda_1 > 0, \lambda_2 > 0$.

   Consider the hypotheses

$$H_0 : \lambda_1 = \lambda_2 \quad \text{against} \quad H_1 : \lambda_1 \neq \lambda_2.$$

   Recall: The probability density function pdf of a random variable $Z \sim Exp(\lambda)$ is $f(z) = \lambda \exp(-\lambda z)$ for $z > 0$.

   (a) Write down and simplify the likelihood.

   (b) Work out the maximum likelihood estimator under the assumption that H_0 is true.

   (c) Work out the maximum likelihood estimator under the full model.

   (d) Describe a likelihood ratio test for the above hypotheses. Clearly describe (and simplify if possible) the test statistic, and state the decision rule.

3. (a) In a linear model with second order assumptions as stated in the lectures and a full rank design matrix X, derive the covariance matrix $\text{cov}(\widehat{\beta})$ of the least squares estimator $\widehat{\beta}$.

(b) A scientist has 3 objects whose weights w_1, w_2, w_3 are to be determined. She decides to do this in one of two way, namely by *either*

(i) weighing each object individually, *or*,

(ii) selecting all possible 3 subsets of 2 objects and measuring the combined weight of the objects in each subset.

Assuming that all observations are uncorrelated, are unbiased and have equal variance σ^2, write down suitable linear models for (i) and (ii), with w_1, w_2, w_3 as unknown parameters. In each case find an expression for $\text{Var}(\widehat{w}_i)$, where $\widehat{w}_i$ is obtained from your linear model by least squares estimation. Which method would you recommend and why?

[Hint: You may need to make use of the fact that if a and b are constant and I is the $n \times n$ identity matrix then $(aI + bJ)^{-1} = cI + dJ$ for some constants c and d, where J is the $n \times n$ matrix of all ones.]

4. (a) Define the non-central χ^2-distribution and the non-central F-distribution.

(b) Define the term "projection matrix".

(c) State the Fisher-Cochran Theorem.

(d) Give a concrete example of a linear model and hypotheses that can be tested with an F-test. Describe how you would conduct the test in your example and what assumptions you need to make. Clearly state the test statistic and the rejection rule.

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2014

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) $\text{bias}_\theta(T) = \mathrm{E}_\theta(T) - \theta$. $\boxed{\text{seen} \Downarrow}$
$\text{MSE}_\theta(T) = \mathrm{E}_\theta((T-\theta)^2)$. $\boxed{2}$

(ii) $\text{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + (\text{bias}_\theta(T))^2$ $\boxed{2}$

Let $X = T - \theta$. Then $\mathrm{Var}_\theta(X) = \mathrm{E}_\theta(X^2) - (\mathrm{E}_\theta\, X)^2 = \text{MSE}_\theta(T) - (\text{bias}_\theta(T))^2$. Rearranging shows the relationship. $\boxed{3}$

(b) *(this is an example in the lecture notes)*
The a-priori density of θ is $\pi(\theta) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\theta^2) \propto \exp(-\frac{1}{2}\theta^2)$ and the conditional density of the observations is

$$p(x|\theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right) \propto \exp\left(-\frac{1}{2}(n\theta^2 - 2\theta\sum_{i=1}^{n} x_i)\right).$$

Hence,

$$\begin{aligned} p(\theta|x) \propto & p(x|\theta)\pi(\theta) \propto \exp\left(-\frac{1}{2}\left((n+1)\theta^2 - 2\sum_{i=1}^{n} x_i\theta\right)\right) \\ \propto & \exp\left(-\frac{1}{2}\frac{(\theta - \sum_{i=1}^{n} x_i/(n+1))^2}{1/(n+1)}\right) \end{aligned}$$

and thus $\theta|x_1, \dots, x_n \sim N(\frac{1}{n+1}\sum_{i=1}^{n} x_i, \frac{1}{n+1})$. $\boxed{6}$

(c) *(students have seen an example in which the MLE for the uniform distribution on $[0, \theta]$ was derived)* $\boxed{\text{sim. seen} \Downarrow}$
The likelihood function is

$$\begin{aligned} L(a,b) = & \prod_{i=1}^{n} \frac{1}{b-a} \mathrm{I}\,(a \le X_i \le b) \\ = & \frac{1}{(b-a)^n} \mathrm{I}\,(a \le X_i \le b, i = 1, \dots, n) \\ = & \frac{1}{(b-a)^n} \mathrm{I}\,(a \le \min(X_1, \dots, X_n) \text{ and } \max(X_1, \dots, X_n) \le b) \end{aligned}$$

This is increasing in a for $a \le \min(X_1, \dots, X_n)$ and equal to 0 otherwise. Similarly, L is decrasing in b for $b \ge \max(X_1, \dots, X_n)$. Thus the MLE is $(\widehat{a}, \widehat{b}) = (\min(X_1, \dots, X_n), \max(X_1, \dots, X_n))$. $\boxed{7}$

2. *(in an example of the lecture notes, students have seen the derivation of the likelihood ratio test statistic for m groups of exponentially distributed random variables)* sim. seen ⇓

(a) The likelihood is

$$L(\lambda_1,\lambda_2) = \prod_{j=1}^{n_1} \lambda_1 \exp(-\lambda_1 Y_{1j}) \prod_{j=1}^{n_2} \lambda_2 \exp(-\lambda_1 Y_{2j})$$
$$= \lambda_1^{n_1} \exp\left(-\lambda_1 \sum_{j=1}^{n_1} Y_{ij}\right) \lambda_2^{n_2} \exp\left(-\lambda_2 \sum_{j=1}^{n_2} Y_{ij}\right).$$

3

(b) To find the MLE under the null hypothesis: Under H_0, we have $\lambda_1 = \lambda_2$. Denoting this joint parameter by λ the likelihood simplifies to

$$L(\lambda) = \lambda^{n_1+n_2} \exp\left(-\lambda \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}\right).$$

This is the likelihood for iid Exp(λ) observations. To derive the MLE,

$$\log L(\lambda) = (n_1+n_2)\log(\lambda) - \lambda \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}$$

and

$$\frac{\partial}{\partial\lambda} \log L(\lambda) = \frac{n_1+n_2}{\lambda} - \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}.$$

Equating this to zero and solving for λ gives the following candiate for the MLE:

$$\widehat{\lambda} = \frac{n_1+n_2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}}.$$

5

This is indeed the MLE as

$$\left(\frac{\partial}{\partial\lambda}\right)^2 \log L(\lambda) = -\frac{n_1+n_2}{\lambda^2} < 0.$$

2

(c) To find the MLE under the full model: As the factors in the likelihood are nonnegative, and only contain different components of the parameter, we can maximise them separately. Each element is the likelihood of iid Exp(λ_i) observations, so the same argument as above holds and we get the MLE

$$\widehat{\lambda}_i = \frac{n_i}{\sum_j Y_{ij}}, \quad i = 1,2.$$

4

(d) The test statistic of the likelihood ratio test is

$$t = \frac{\sup_{\lambda_1,\lambda_2} L(\lambda_1,\lambda_2)}{\sup_\lambda L(\lambda)}.$$

Using the above, the denominator simplifies to

$$L(\widehat{\lambda}) = \left(\frac{n_1+n_2}{\sum_{i=1}^2 \sum_{j=1}^{n_i} Y_{ij}}\right)^{n_1+n_2} \exp(-n_1-n_2).$$

The numerator simplifies to

$$\sup_{\lambda_1,\lambda_2} L(\lambda_1,\lambda_2) = L(\widehat{\lambda}_1,\widehat{\lambda}_2) = \left(\frac{n_1}{\sum_{i=1}^{n_1} Y_{1i}}\right)^{n_1}\left(\frac{n_2}{\sum_{i=1}^{n_1} Y_{2i}}\right)^{n_2} \exp(-n_1-n_2).$$

Hence,

$$t = \frac{\left(\frac{n_1}{\sum_{i=1}^{n_1} Y_{1i}}\right)^{n_1}\left(\frac{n_2}{\sum_{i=1}^{n_1} Y_{2i}}\right)^{n_2}}{\left(\frac{n_1+n_2}{\sum_{i=1}^2 \sum_{j=1}^{n_i} Y_{ij}}\right)^{n_1+n_2}}.$$

(This could be written a bit more elegantly as $t = \frac{\bar{Y}^{n_1+n_2}}{\bar{Y}_1^{n_1}\bar{Y}_2^{n_2}}$ where $\bar{Y} = \frac{1}{n_1+n_2}\sum_{i=1}^2 \sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, but students are not required to do this) $\boxed{3}$

From the lectures we know

$$2\log t \xrightarrow{d} \chi_1^2,$$

as n_1 and n_2 tend to ∞.

So, to get an asymptotic test to the level α one would reject H_0 if $2\log t > c$ where c is such that $\mathrm{P}(Z > c) = \alpha$ with $Z \sim \chi_1^2$. $\boxed{3}$

3. (a) *(Students have seen this in the lecture notes)* $\boxed{\text{seen} \Downarrow}$

Let $\mathrm{E}(\boldsymbol{Y}) = X\boldsymbol{\beta}$ be the linear model under consideration. As we are dealing with a full rank linear model, the least squares estimator is

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y}.$$

Thus,

$$\begin{aligned}\operatorname{cov}(\widehat{\boldsymbol{\beta}}) =& (X^TX)^{-1}X^T\operatorname{cov}(\boldsymbol{Y})((X^TX)^{-1}X^T)^T\\ =& (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1}\\ =& \sigma^2(X^TX)^{-1}.\end{aligned}$$

$\boxed{5}$

(b) *(a similar question, but with 5 objects, was on one of the non-assessed problem sheets)* $\boxed{\text{sim. seen} \Downarrow}$

Appropriate linear models:

(i)

$$\mathrm{E}\boldsymbol{Y} = \underbrace{\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}}_{=:X}\underbrace{\begin{pmatrix}w_1\\ w_2\\ w_3\end{pmatrix}}_{=:\boldsymbol{\beta}}$$

(ii)

$$\mathrm{E}\boldsymbol{Y} = \underbrace{\begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 1\\ 1 & 0 & 1\end{pmatrix}}_{=:X}\underbrace{\begin{pmatrix}w_1\\ w_2\\ w_3\end{pmatrix}}_{=:\boldsymbol{\beta}}$$

(The rows of X in both models can of course be presented in a different order.)

The covariance matrices of the least squares estimators in these models are as follows: $\boxed{5}$

(i) $\operatorname{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^TX)^{-1} = \sigma^2 I$

Thus, for $i = 1, 2, 3$, $\operatorname{Var}(\widehat{w}_i) = \sigma^2$. $\boxed{3}$

(ii) $\operatorname{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^TX)^{-1} = \sigma^2\begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{pmatrix}^{-1} = \sigma^2(I+J)^{-1}$

Using the hint from the question, we know that the inverse of $I + J$ is of the form $aI + bJ$ for some $a, b \in \mathbb{R}$. To determine the constants,

$$I = (I+J)(aI+bJ) = aI + (a+b)J + bJ^2 = aI + (a+4b)J.$$

Thus, $a = 1$ and $b = -1/4$. Hence,

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2\begin{pmatrix}3/4 & -1/4 & -1/4\\ -1/4 & 3/4 & -1/4\\ -1/4 & -1/4 & -3/4\end{pmatrix}.$$

Thus, for $i = 1, 2, 3$, $\operatorname{Var}(\widehat{w}_i) = 0.75\sigma^2$. $\boxed{4}$

The effort for both methods (number of weighings) is the same. Both result in unbiased estimators. Thus, they can be compared by comparing the variance of the resulting estimators. The variance of method (ii) is smaller and thus method (ii) is to be preferred. $\boxed{3}$

4. (a) Let $Z \sim N(\mu, I_n)$, where $\mu \in \mathbb{R}^n$. $\boxed{\text{seen} \Downarrow}$

$U = Z^T Z = \sum_{i=1}^n Z_i^2$ is said to have a non-central χ^2-distribution with n degrees of freedom (d.f.) and non-centrality parameter

$$\delta = \sqrt{\mu^T \mu}.$$

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.$=\delta$. $\boxed{4}$

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

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

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 ($AA = A$). $\boxed{3}$

(c) If $A_1, \dots, A_k$ 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 $Z^T A_1 Z, \dots, Z^T A_k Z$ are independent and

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

$\boxed{3}$

(d) This is an open question, so any reasonable example will be acceptable. Points will be allocated as follows: unseen ⇓

* Statement of an example and the corresponding hypotheses. 4
* The normal theory assumption, i.e. the assumption that the error satisfies $\epsilon \sim N(0, \sigma^2 I)$. 1
* Test statistic and the rejection rule, including appropriate degrees of freedom. 5

The marking of this question will put particular emphasis on clarity of expression.
One example solution follows. This is a deliberately simple example:
Suppose we observe the body and the brain weight of $n = 60$ mammals. *(a similar data set has been used as an example in the lecture notes, but not for F-tests)*
Let $Y_i = \log(\text{brain weight of the } i\text{th mammal})$ and let $o_i = \log(\text{body weight of the } i\text{th mammal})$. Consider the linear model

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

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$.
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 also be written as

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

The test statistic of the F-test is

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

where $n = 60$, $r = \operatorname{rank} X = 2$, $s = \operatorname{rank} X_0 = 1$. 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.
Under H_0, the test statistics satisfies $F \sim F_{r-s,n-r} = F_{1,59}$ 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 $Z \sim F_{1,59}$.