Queen Mary
University of London

B. Sc. Examination by course unit 2013

MTH6134 Statistical Modelling II

Duration: 2 hours

Date and time: xx yyyy 2013, zz:wwh–aa:bbh

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

You should attempt all questions. Marks awarded are shown next to the questions.

Calculators ARE permitted in this examination. The unauthorized use of material stored in pre-programmable memory constitutes an examination offence. Please state on your answer book the name and type of machine used.

Statistical functions provided by the calculator may be used provided that you state clearly where you have used them.

The New Cambridge Statistical Tables are provided.

Complete all rough workings in the answer book and cross through any work which is not to be assessed.

Important note: the Academic Regulations state that possession of unauthorized material at any time by a student who is under examination conditions is an assessment offence and can lead to expulsion from QMUL.

Please check now to ensure you do not have any notes, mobile phones or unauthorised electronic devices on your person. If you have any, then please raise your hand and give them to an invigilator immediately. Please be aware that if you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. Disruption caused by mobile phones is also an examination offence.

Exam papers must not be removed from the examination room.

Examiner(s): H. Maruri-Aguilar and D. S. Coad

 TURN OVER

Question 1 (26 marks)
The cuckoo has the distinct habit of laying its eggs in nests of birds of other species. Measurements of cuckoo eggs found in the nests of three different species are given below (lengths in millimetres).

Hedge sparrow	22.0	23.9	20.9	22.8	23.1	23.1
Robin	21.8	23.0	23.3	23.9	22.3	22.0
Wren	19.8	22.1	21.5	20.3	20.9	22.0

A baseline regression analysis was performed, for which the data for Wren was used as the baseline. Below is the Genstat output.

```
Regression analysis
Response variate:  length
Fitted terms:  Constant, beta1, beta2

Summary of analysis
Source       d.f.    s.s.    m.s.    v.r.
Regression      2    9.94  4.9717    5.67
Residual       15   13.16  0.8774
Total          17   23.11  1.3591

Estimates of parameters
Parameter   estimate  s.e.    t(15)  t pr.
Constant      21.100  0.382   55.18  <.001
beta1          1.533  0.541    2.84  0.013
beta2          1.617  0.541    2.99  0.009
```

(a) Write and describe the baseline model used and give an explicit version of the matrix X used in the baseline regression model $Y = X\beta + \epsilon$. As part of your response, explain the meaning of the baseline model parameters in relation to the standard model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, where parameters $\alpha_1, \alpha_2, \alpha_3$ correspond to Hedge sparrow, Robin and Wren respectively. [8]

(b) Describe the contents of the Genstat output given and interpret the results. As part of your response test the hypothesis of equality of mean egg sizes between the three bird species. [7]

(c) Perform the hypothesis test for the difference between mean sizes of eggs laid in Robin's nests and in Hedge sparrow's nests. [5]

(d) A scientist is interested in comparing size of eggs laid in Wrens nests against the average of Hedge sparrow and Robin. To this end, he proposed the contrast $L_1 = \alpha_1 + \alpha_2 - 2\alpha_3$. He wants to test another contrast $L_2 = \lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_3\alpha_3$. Using the condition of orthogonality of contrasts, propose values $\lambda_1, \lambda_2, \lambda_3$ such that contrast L_2 is orthogonal to L_1. [6]

Question 2 (25 marks)

There is interest in studying the effect of five types of barley. For that purpose, three farms were selected at random from a large candidate set in Minnesota. In each farm, five fields were sown, each with a different type of barley, and the yield was recorded after harvesting. The data is given below.

	Yield (bushels per acre)				
Farm	Manchuria	Svansota	Velvet	Trebi	Peatland
1	27	31	33	33	30
2	41	43	44	57	42
3	31	30	32	45	37

(a) Describe an appropriate model for the data and assumptions required. [4]

(b) Compute the analysis of variance table and test the hypothesis concerning the effect of barley. [9]

(c) Compute the standard error of a difference between treatment means. [3]

(d) Perform the comparisons between means using the least significant difference method. [6]

(e) Based on the results of your analysis, suggest and argue for a contrast of interest. [3]

Question 3 (19 marks)

The effect of three types of rubber and two types of additive for the manufacture of tyres was investigated in an experiment. The response was percentage of wear. A table with the data is given below.

		Rubber								
		1			2			3		
Additive	1	16	15	17	20	20	19	18	20	19
	2	20	18	18	21	19	22	21	22	22

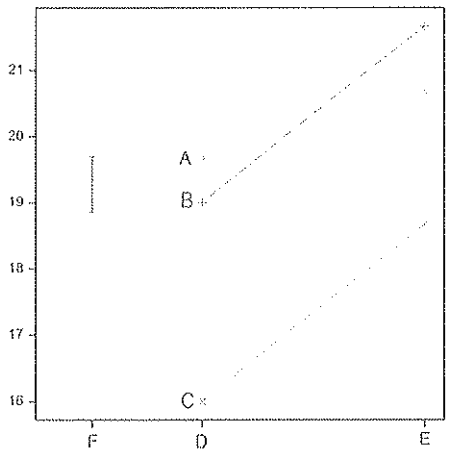
`Genstat` output was also given.

```
Source of variation d.f.  s.s.    m.s.    v.r.
additive               1  20.056  20.056  19.00
rubber                 2  34.111  17.056  16.16
additive.rubber        2   2.778   1.389   1.32
Residual              12  12.667   1.056
Total                 17  69.611
```

(a) Describe an appropriate model for the data, including assumptions. [4]

(b) Perform the hypothesis tests of the analysis of variance table. [4]

(c) The following interaction graph was produced as part of the analysis.



(i) Some labels are missing and you are required to complete the plot by identifying factor and labels for lines (A,B,C); identify factor and labels for horizontal axis (D,E); compute length of the bar shown (F), which is the standard error for the difference of means depicted in the plot. [8]

(ii) Interpret the graph. Is your interpretation compatible with your analysis? [3]

Question 4 (19 marks)

A study was carried out to measure composition of soil. To that effect, soil samples were taken at two sites, each site selected at random from a plot. Two plots were studied this way, each plot in turn being selected randomly from a large set of candidates. The data is given below, where the response of interest was the percentage of dry matter after digging the weeds.

Plot	1						2					
Site	1			2			1			2		
Dry matter	57	59	61	69	69	68	70	70	76	83	80	84

Partial `Genstat` output for the analysis is also available.

```
Variate: drymatter
Source of variation  d.f.   s.s.     m.s.
plot                    1   533.333  533.333
plot.site               2   300.333  150.167
Residual                8    41.333    5.167
Total                  11   875.000
```

(a) Describe an appropriate model for the data. [6]

(b) Perform the hypothesis tests of factors plot and site. [6]

(c) Estimate the variance components, recall that expected mean squares satisfy the following relations: $E(M_A) = \sigma^2 + n_B r\sigma_A^2 + r\sigma_B^2$, $E(M_{B(A)}) = \sigma^2 + r\sigma_B^2$ and $E(M_E) = \sigma^2$. [4]

(d) How would you analyze this data in `Genstat`? [3]

Question 5 (11 marks)
Consider a completely randomized design with two treatments and replications per treatment $r_1 = 2, r_2 = 3$. Let the vector of response values for this design be $y^T = (y_{11}, y_{12}, y_{21}, y_{22}, y_{23})$ and for $i = 1, 2$ and $j = 1, \ldots, r_i$ consider the standard model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$.

(a) Define the treatment subspace V_T and the null subspace V_0. [2]

(b) List the basis vectors $\{u_1, u_2\}$ for the treatment subspace V_T and compute the projections $P_{V_T} y$, $P_{V_T^\perp} y$ and $P_{V_0} y$. [6]

(c) Show that $P_{V_0} y$ is orthogonal to $P_{V_T^\perp} y$. [3]

End of Paper