Main Examination period 2019

MTH6134 / MTH6134P: Statistical Modelling II

Duration: 2 hours

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 available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

The New Cambridge Statistical Tables are provided.

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

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. 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. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: D. S. Coad, H. Maruri-Aguilar

Question 1. [29 marks] The root yield of sugar beet, in tons per acre, was recorded for plots fertilised with six levels of nitrogen. Below are the data.

Nitrogen	Yield					Total
1	31.3	33.4	29.2	32.2	33.9	160.0
2	38.8	37.5	37.4	35.8	38.4	187.9
3	40.9	39.2	39.5	38.6	39.8	198.0
4	40.9	41.7	39.4	40.1	40.0	202.1
5	39.7	40.6	39.2	38.7	41.9	200.1
6	40.6	41.0	41.5	41.1	39.8	204.0

The sum of squares of the observations is $\sum_{i=1}^{6}\sum_{j=1}^{5} y_{ij}^2 = 44,555.61$.

(a) Write down a suitable model for these data and any necessary assumptions, explaining your notation. **[5]**

(b) Compute the analysis of variance table and test factor nitrogen. **[10]**

(c) Use the least significant difference method to compare pairs of nitrogen means. **[10]**

(d) Describe how the model and analysis would have changed if the levels of nitrogen had been selected at random from a large number of possible levels available. **[4]**

Question 2. [23 marks] An experiment was designed to study the performance of four different detergents in cleaning clothes. The cleanness readings were obtained with specially designed equipment for three different types of common stains. The data are given below, where a higher reading indicates that the clothes are cleaner.

	Detergent			
Stain	1	2	3	4
1	45	47	48	42
2	43	46	50	37
3	51	52	55	49

The model to be used for these data is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

for $i = 1,2,3,4$ and $j = 1,2,3$, where α_i is the effect of the ith detergent and β_j is the effect of the jth stain. The constraints on the parameters are $\sum_{i=1}^{4}\alpha_i = 0$ and $\sum_{j=1}^{3}\beta_j = 0$, and it is assumed that $\varepsilon_{ij} \sim N(0,\sigma^2)$, all independent.

(a) Derive expressions for the least squares estimates of μ, α_i and β_j. **[8]**

(b) Derive expressions for the fitted values and residuals. **[3]**

(c) What is the advantage of using a randomised block design here rather than a completely randomised design? **[4]**

(d) Using the above constraints, write down in matrix form the multiple regression model that is equivalent to the analysis of variance model for the data. **[8]**

Question 3. [15 marks] A bacteriologist is interested in the effect of two different culture mediums at two different times on the growth of a particular virus. She performs six replicates of a factorial design, making the runs in random order. Below are the results, in plaque-forming units.

	Culture Medium			
Time	1		2	
12	21	22	25	26
	23	28	24	25
	20	26	29	27
18	37	39	31	34
	38	38	29	33
	35	36	30	35

An analysis of variance was performed using GenStat and the output is as follows:

```
Analysis of variance

Variate: growth

Source of variation d.f.     s.s.      m.s.    v.r.  F pr.
medium                1     9.375     9.375    1.84  0.191
time                  1   590.042   590.042  115.51  <.001
medium.time           1    92.042    92.042   18.02  <.001
Residual             20   102.167     5.108
Total                23   793.625
```

(a) Briefly explain how you would enter these data into GenStat. What expression should be entered in Treatment Structure in the analysis of variance Dialogue Box? [**4**]

(b) Draw conclusions from the above output, illustrating your answer by calculating and commenting on the treatment means. [**5**]

(c) What residual plots would you look at in GenStat, and why? [**4**]

(d) Explain why the runs are made in random order. [**2**]

Question 4. [22 marks] Three regional health authorities were chosen at random to participate in a health awareness programme. Within each authority, three cities were randomly selected for participation. To evaluate the effectiveness of the programme, five households within each city were randomly selected. All members of the selected households were interviewed before and after participation, and a composite index was formed for each household in order to measure the impact of the programme. The data are given below, where the larger the index, the greater the awareness.

Authority	1			2			3		
City	1	2	3	1	2	3	1	2	3
	42	26	34	47	56	68	19	18	16
	56	38	51	58	43	51	36	40	28
	35	42	60	39	65	49	24	27	45
	40	35	29	62	70	71	12	31	30
	28	53	44	65	59	57	33	23	21

The sums of squares of the observations and the treatment totals are $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{5} y_{ijk}^2 = 89,246$ and $\sum_{i=1}^{3}\sum_{j=1}^{3} T_{ij}^2 = 426,764$.

(a) Write down a suitable model for these data and any necessary assumptions, explaining your notation. **[5]**

(b) Compute the analysis of variance table, and test factors authority and city. **[13]**

(c) Estimate the variance components. **[4]**

Question 5. [11 marks] Consider a completely randomised design with two treatments. Let the data vector be $\mathbf{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23})^\top$.

(a) Define the treatment subspace V_T and the null subspace V_0. **[3]**

(b) Compute the projections $P_{V_T}\mathbf{y}$ and $P_{V_0}\mathbf{y}$. **[4]**

(c) Define the sum of squares and the degrees of freedom for V_T. **[2]**

(d) In the usual analysis of variance notation, what are the formulae for the quantities that you defined in part (c)? **[2]**

End of Paper.