Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2022	MATH2697-WE01

Title: Statistical Modelling II

Time:	2 hours	
Additional Material provided:	Statistical tables	
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Students must use the mathematics specific answer book.

Revision:	

SECTION A

Q1 Let X be a bivariate normal random vector with mean $\boldsymbol{m}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, i.e. $X \sim N_2(\boldsymbol{m}, \boldsymbol{\Sigma})$. Assume that $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are both known. For a point $\boldsymbol{x} \in \mathbb{R}^2$, the Mahalanobis distance to the mean $\boldsymbol{m}$ is defined as

$$d_M(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{m})}$$

1.1 Show that $d_M(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\Sigma}) \sim \chi^2_{(q)}$, where $\chi^2_{(q)}$ denotes a χ^2 distribution with q degrees of freedom, and give the value of q.

1.2 Explain how the Mahalanobis distance can be used to test for outliers; that is for a data set of size n we wish to test the null hypothesis H_0 : Case i is not outlying vs H_1 : Case i is an outlier, for all $i = 1, \ldots, n$. Describe the test, give the test statistic and its distribution, and state any assumptions made.

1.3 For a certain bivariate data set, the squared Mahalanobis distances d^2_M of all $n = 20$ observations are as follows:

i	1	2	3	4	5	6	7	8	9	10
d^2_M	1.453	2.756	2.430	2.593	0.545	0.816	0.812	1.250	1.052	3.729
i	11	12	13	14	15	16	17	18	19	20
d^2_M	1.058	2.039	3.421	3.408	4.900	1.812	0.245	2.193	8.470	1.335

Without using the Bonferroni correction, carry out the tests at a 10% level of significance. Explain your working, and write down the case numbers of the detected outliers.

1.4 For a given random vector $X = (X_1, X_2)^T$, where X_1 and X_2 are independent such that $X_j \sim N(m_j, \sigma^2)$, $j = 1, 2$, show that the Mahalanobis distance d_M between a realization $\boldsymbol{x} = (x_1, x_2)^T$ of X and the mean $\boldsymbol{m} = (m_1, m_2)^T$ is proportional to the corresponding Euclidean distance d_E and give the proportionality constant.

Q2 Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

2.1 Show that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric and positive semidefinite.

2.2 Assume from now on that, for a given data set, $\boldsymbol{X}^T\boldsymbol{X}$ is in fact positive definite, and let $R(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$. By minimizing $R(\boldsymbol{\beta})$, derive the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

2.3 Explain why the property positive definite is important for the existence of the least squares estimator, and provide (without proof) a necessary condition for this property.

2.4 Show that $\hat{\boldsymbol{\beta}}$ yields a minimum (rather than a maximum or saddlepoint) of $R(\boldsymbol{\beta})$.

SECTION B

Q3 An experiment was carried out to assess the effects of soy plant variety (levels: I, II and III) and planting density (levels: 5, 10, 15, and 20 thousand plants per hectare) on yield. Each of the 12 treatments was randomly applied to 3 plots. The data set is given below:

	density			
variety	5(k/ha)	10(k/ha)	15(k/ha)	20(k/ha)
I	7.8, 9.1, 10.6	11.2, 12.7, 13.3	12.1, 12.5, 14.1	9.1, 10.7, 12.6
II	8.0, 8.7, 10.0	11.3, 12.9, 13.8	13.8, 14.3, 15.4	11.3, 12.7, 14.3
III	15.3, 16.0, 17.6	16.8, 18.3, 19.2	17.9, 21.0, 20.7	17.2, 18.3, 19.1

We fit a sequence of linear models and observe the residual sum of squares, RSS:

Variables included	RSS
1	461.896
1+ `variety`	134.123
1+ `variety` + `density`	47.215
1+ `variety` + `density` + `variety:density`	39.147

Here, a 1 symbolizes the intercept term and `variety:density` symbolizes the interaction of `variety` and `density`. We denote the degrees of freedom contributed by the sources `variety`, `density`, the interaction, and the residuals of the full interaction model, by df_1, df_2, df_3, and df_{res}, respectively.

3.1 In the context of this data set, explain the terms *treatment* and *replicates*, and make clear what it means to speak of *complete* and *balanced* factorial design.

3.2 How many parameters would we have in a full unconstrained interaction model for this data set? How many parameters do we have for the constrained model? Describe an appropriate constraint which is commonly used.

3.3 Construct a sequential Analysis of Variance table. For all three sources of variation, give the F-values and test the null hypothesis: the source does not contribute to the variation in the response at the 5% level of significance. Hint: You can use $F_{0.05}(df_1, df_{res}) = 3.40$, $F_{0.05}(df_2, df_{res}) = 3.01$, and $F_{0.05}(df_3, df_{res}) = 2.51$.

3.4 Without performing calculations, state how the ANOVA table derived in the previous question changes if the order of inclusion of `variety` and `density` is interchanged. What is the RSS of a model containing only the intercept and `density`? Explain your answers.

3.5 Next, test (at the 5% level of significance) the full interaction model against a model not containing `density` at all (in other words, against a model only containing the intercept and `variety`). Give an interpretation of your result. Hint: You can use $F_{0.05}(df_1^* - df_{res}, df_{res}) = 2.30$, where df_1^* is the degrees of freedom of the model that only contains the intercept and `variety`.

Q4 In an attempt to understand what affects the rate of crime, measurements on the following variables were recorded for 30 U.S. states.

- `y` violent crime rate
- `x1` poverty rate
- `x2` percentage of those living in urban areas
- `x3` percentage of single-parent families

A linear model was fitted, the summary of which and some supplementary information are displayed **in the R output shown below.**

4.1 In the summary table, provide the missing values of A, B, and C, and a lower bound for D.

4.2 Which of the four regression parameters are significantly different from zero at the 5% level of significance?

4.3 Without carrying out calculations, specify which of the 95% confidence intervals for the four parameters contains zero and which not. Then, calculate the confidence interval for the parameter associated with the `x1` variable explicitly.

4.4 We are interested in predicting the violent crime rate for a state, with poverty rate of 12, percentage of those living in urban areas of 70% and percentage of single-parent families of 25%. Based on the full model above, find (i) the predicted value $\hat{y}$, (ii) a confidence interval for the expected violent crime rate, (iii) and a prediction interval for the individual violent crime rate.

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = crime)

Residuals:
    Min      1Q  Median      3Q     Max
-391.74  -75.44   -2.73  119.85  244.62

Coefficients:
            Estimate   Std. Error     t value      Pr(>|t|)
(Intercept)       A      177.461      -4.742     6.64e-05 ***
x1            17.899     11.598        1.543          D
x2             5.357          B        3.187     0.003725 **
x3            30.883      7.229          C        0.000229 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 148.9 on 26 degrees of freedom
Multiple R-squared:  0.7269,Adjusted R-squared:  0.6953
F-statistic: 23.06 on 3 and 26 DF,  p-value: 1.703e-07

> round(summary(crime.lm)$cov.unscaled,4)
            (Intercept)      x1      x2      x3
(Intercept)      1.4212 -0.0476 -0.0084 -0.0087
x1              -0.0476  0.0061  0.0005 -0.0025
x2              -0.0084  0.0005  0.0001 -0.0003
x3              -0.0087 -0.0025 -0.0003  0.0024
```