Durham University

EXAMINATION PAPER

Examination Session: May/June	**Year:** 2022	**Exam Code:** MATH3051-WE01

Title: Statistical Methods III

Time:	3 hours	
Additional Material provided:	Statistical tables; Graph paper.	
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:	

ED01/2022

SECTION A

Q1 (a) Consider the bivariate random vector $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with density

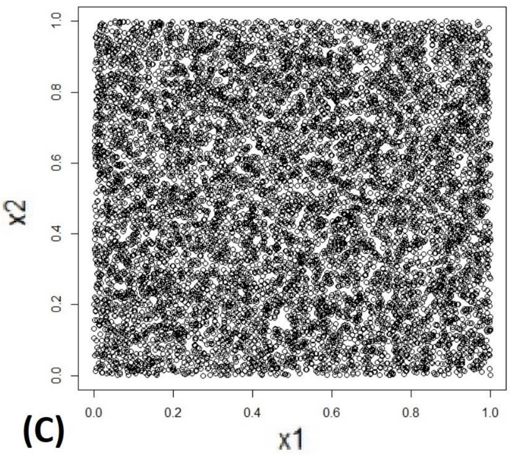
$$f(x_1, x_2) = \begin{cases} 2x_1 & 0 \le x_1 \le 1, 0 \le x_2 \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\boldsymbol{m} = E(X)$ and $\boldsymbol{\Sigma} = \text{Var}(X)$ through explicit calculation. Verify that your result for $\boldsymbol{\Sigma}$ is a valid variance matrix.

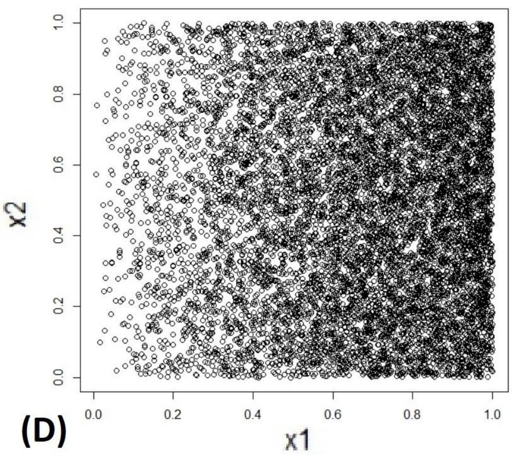
(b) Find the correlation matrix of X.

(c) Below are four data sets (A), (B), (C), and (D), each of size $n = 10000$. Each is generated from a bivariate random vector, but only one of them is generated from X. Which one is it? Explain your answer.

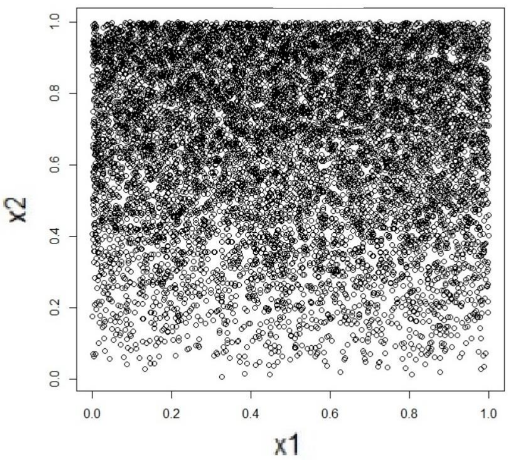
(A)



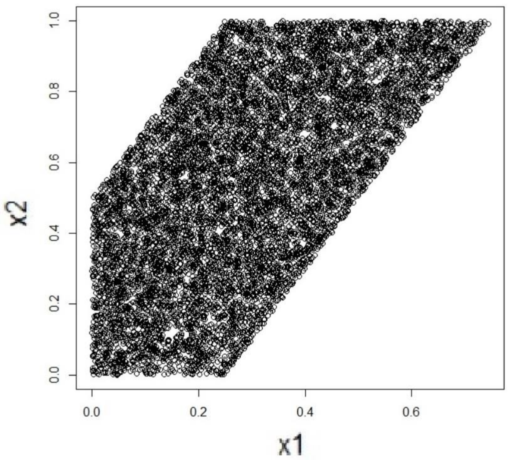
(B)



(C)



(D)



Q2 In 1947, a study was published giving details of an experiment which investigated the effect of vitamin C intake on tooth growth[1]. Guinea pigs were given doses of vitamin C for a fixed period of time, and then the length of each Guinea pig's odontoblasts (the cells responsible for tooth growth) was measured. A total of $n = 60$ Guinea pigs were included in the experiment.

The predictor variables used were the `delivery` method for vitamin C (a factor with two levels: orange juice and absorbic acid), and the `dose` of vitamin C (a continuous measurement, equal here to 0.5, 1, or 2 milligrams per day). Each combination of delivery method and dose was given to ten Guinea pigs.

A sequence of linear models is fitted to the study data, giving the following residual sums of squares (RSS):

Model	Variables included	RSS
(1)	1	3452.209
(2)	1 + delivery	3246.859
(3)	1 + delivery + dose	1022.555
(4)	1 + delivery + dose + delivery:dose	933.6349

(a) Complete the sequential Analysis of Variance table shown below (note that we assume normally distributed error terms).

	DoF	Sum of squares	Mean squares	F value
Delivery				
Dose				
Delivery:Dose				
Residuals				-

(b) At the 5% level of significance, and choosing the most appropriate F-distribution possible from statistical tables, carry out the following hypothesis tests:

(i) Model (1) versus Model (4);

(ii) Model (2) versus Model (4).

(c) The data is relabelled so that the `dose` variable takes values "low", "medium", and "high", replacing measurements of 0.5, 1, and 2 milligrams, respectively. This variable is now treated as a factor with three levels, rather than as a continuous measurement. Models containing the dose variable are recalculated accordingly.

What is the new value of the number of predictor variables, p, for the full model? What effect will the change have upon

(i) SST?

(ii) SSR?

Give a justification for your answer in each case.

[1] Crampton, E. W. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the guinea pig. The Journal of Nutrition, **33**(5), 491504

Q3 (a) What are the two assumptions underlying the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$? Which assumption does not underly the linear regression model, but is commonly assumed to aid with model analysis?

(b) In the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, prove that $\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$.

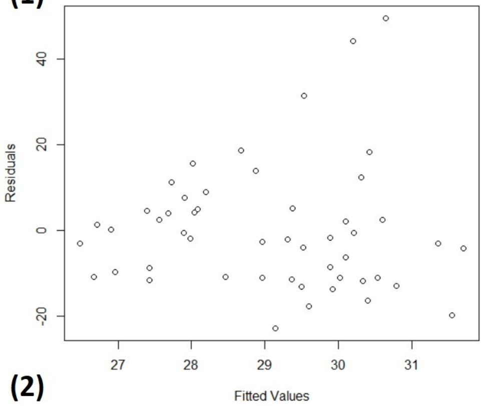
Describe what this result implies for the sum of the residuals if an intercept term is included in the model.

(c) The yield (in tonnes) of fifty fields of corn in 2020 was recorded. The average width and average length (both in metres) of each field was also recorded. Fitting the two models

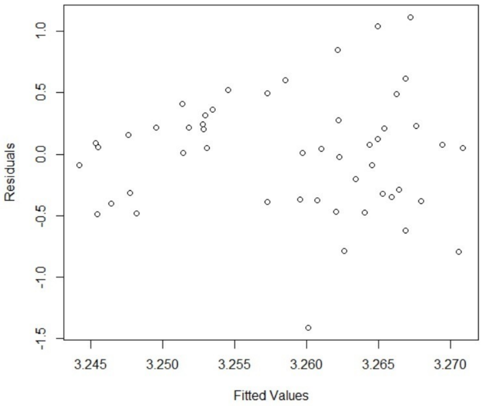
$$\begin{aligned} \text{yield} &= \beta_1 + \beta_2\text{length} + \beta_3\text{width} + \epsilon & (1) \\ \log(\text{yield}) &= \beta_1 + \beta_2\log(\text{length}) + \beta_3\log(\text{width}) + \epsilon & (2) \end{aligned}$$

leads to the following residual plots:

(1)



(2)



(i) Which of the two residual plots indicates the better model fit? Give a reason for your answer.

(ii) In terms of the relationship between field dimensions and the amount of corn which can be grown, which of two model specifications is the more natural one? Give a reason for your answer.

Q4 The World Almanac and Book of Facts, 1975 reports the height (in inches) and weight (in pounds) of multiple American women aged 30 to 39. Six of these data points are given in the table below, with height denoted (x) and weight denoted (y).

i	1	2	3	4	5	6
x_i	58	59	60	61	62	63
y_i	115	117	120	123	126	129

We describe the relationship between height and weight with a simple linear regression model $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$. Let $\boldsymbol{Y} = (y_1, \ldots, y_6)^T$.

(a) Give the design matrix $\boldsymbol{X}$, and, hence, compute $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{Y}$.

(b) Write down the normal equations, and solve them for $\hat{\beta}_1$ and $\hat{\beta}_2$.

(c) The estimate s of the error standard deviation takes the value 0.345. Construct 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$.

(d) The simple linear model for this data has an R^2 value of 0.9967. A friend tells you the value of R^2 will get even higher if more data is collected, as

$$R^2 = \frac{\text{Sum of squares for regression}}{\text{sum of squares in total}} = \frac{SSR}{SST}$$

and SST will tend to zero as more data is collected.

Is your friend correct? Give a reason for your answer.

SECTION B

Q5 In 1973, the New York State Department of Conservation (NYDoC) and the National Weather Service (NWS) published data on air quality in New York City. A statistician decides to use this data to model maximum daily temperature (degrees Fahrenheit). The predictor variables available are average daily ozone level (parts per billion), average daily solar radiation level (log(Angstroms)), and average daily wind speed (mph).

As part of the process of choosing the best model possible, the statistician calculates all models possible using the three predictor variables. They include an intercept in all models. They then calculate $C_{\mathcal{I}} = \frac{RSS_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n$ for each model, where for each model $\mathcal{I} \subseteq \{1, \text{ozone}, \text{radiation}, \text{wind}\}$ represents the predictor values included in the model, $p_{\mathcal{I}} = |\mathcal{I}|$, and $RSS_{\mathcal{I}}$ represents the RSS of the model which includes only predictor variables $\mathcal{I}$.

The results of these calculations are given below.

Model number	Model formula	$C_{\mathcal{I}}$
(1)	1	105
(2)	$1 + \text{ozone}$	2.5
(3)	$1 + \text{radiation}$	88
(4)	$1 + \text{wind}$	54
(5)	$1 + \text{ozone} + \text{radiation}$	3.9
(6)	$1 + \text{ozone} + \text{wind}$	2.9
(7)	$1 + \text{radiation} + \text{wind}$	44
(8)	$1 + \text{ozone} + \text{radiation} + \text{wind}$	4

(a) Prove that

$$C_{\mathcal{I}} = \frac{RSS_{\mathcal{I}} - RSS}{s^2} + p_{\mathcal{I}} - p_{\mathcal{D}}$$

where $\mathcal{D} = \{1, \text{ozone}, \text{radiation}, \text{wind}\} \setminus \mathcal{I}$ for a given model. You may state without proof the definition of s^2.

(b) Which of the eight models would you recommend as a way of explaining the variance in maximum daily temperature? Give a reason for your answer.

(c) Further investigation of the NYDoC and NWS archives reveals an additional 12 continuous measurements that could be used as predictor variables in a model with max daily temperature as the response variable.

(i) State the number of models (including those that do not include an intercept) that can be specified using a total of p parameters, assuming no model includes interaction terms. Hence calculate the number of possible models (including those that do not include an intercept) that can be specified using the 15 measurements now available, assuming no model includes interaction terms.

(ii) Define the interaction term (measure 1:measure 2), and hence show it fulfils identical role in a linear model to the interaction term (measure 2:measure 1), where measure 1 and measure 2 are continuous measurements. Your answer should contain no more than 50 words.

(iii) Show that the number of possible models that can be specified using the 15 measurements now available, in which any or all interaction terms between two (and no more than two) different measurements are permitted, is $2^{121}-1$.

(iv) Give a brief summary of a method by which this number of models can be reduced when searching for a model which performs well. Your answer should contain no more than 50 words.

Q6 (a) Consider three individual price indexes, $\boldsymbol{P} = (P_1, P_2, P_3)^T$, recorded monthly in the UK from 1990-2009 in the categories P_1 = "Health", P_2 = "Recreation & Culture", and P_3 = "Hotels, Cafés and Restaurants". These three indexes are to be combined into a single index with name W ("Wellbeing"), through an appropriate linear combination

$$W = a_1P_1 + a_2P_2 + a_3P_3.$$

The unit vector $\boldsymbol{a} = (a_1, a_2, a_3)^T$ is to be chosen such that the projection $\boldsymbol{a}^T\boldsymbol{P}$ has maximal variance among all linear combinations of P_1, P_2 and P_3. A principal component analysis of $\boldsymbol{P}$ is carried out, with results summarised in the usual notation as

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.610 & -0.215 & -0.763 \\ 0.278 & 0.959 & -0.048 \\ 0.742 & -0.183 & 0.645 \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} 578.013 & 0 & 0 \\ 0 & 14.775 & 0 \\ 0 & 0 & 1.454 \end{pmatrix}.$$

(i) Give the vector $\boldsymbol{a}$.

(ii) What proportion of the total variance of $\boldsymbol{P}$ is explained by W?

(iii) Compute the components $\Sigma_{1,1}$, $\Sigma_{3,1}$ and $\Sigma_{3,3}$ of matrix $\boldsymbol{\Sigma}$ where

$$\boldsymbol{\Sigma} \equiv (\Sigma_{i,j})_{1\le i,j\le 3} = \text{Var}(P)$$

(b) We are given a q-dimensional random vector X with mean $\boldsymbol{0} = (0, \ldots, 0)^T$ and variance $\boldsymbol{\Delta}$. Derive theoretically the first principal component; that is, find the unit vector $\boldsymbol{\gamma}$ such that $\text{Var}(\boldsymbol{\gamma}^T X)$ is maximal.

Q7 (a) For a q-variate random vector $X = (X_1, \dots, X_q)^T \sim N_q(\boldsymbol{m}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ positive definite, we know that the density is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right\}. \qquad (1)$$

(i) Consider the special case of (1) where $\mathrm{Var}(X_j) = \sigma_j^2$ for $j = 1, \dots, q$ and $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, with $i, j = 1, \dots, q$. Write down the density for this case and simplify as far as possible. Hence, show that $X_1, \dots, X_q$ are multivariate normal and uncorrelated if and only if they are marginally normal and independent.

(ii) Consider the special case of (1) where $q = 2$, $\boldsymbol{m} = (0, 0)^T$ and $\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Assume we know that the density of X is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-(x_1^2 + x_2^2 - \sqrt{3}\, x_1 x_2)\right\}.$$

Give the values of ρ and σ.

(b) For a random vector $X \sim (\boldsymbol{m}, \boldsymbol{\Sigma})$, we define the Mahalanobis distance to the mean as:

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

(i) Explain why the Cartesian distance function is not appropriate in general when judging how far an instance $\boldsymbol{x}$ of X is from $\boldsymbol{m}$. Your answer should contain no more than 100 words.

(ii) Consider a random vector $X = (X_1, X_2)^T$ where $\mathrm{Var}(X_1) = \sigma_1^2$, $\mathrm{Var}(X_2) = \sigma_2^2$, and $\mathrm{Cov}(X_1, X_2) = 0$. We transform X as follows:

$$X = (X_1, X_2)^T \rightarrow (X_1/\sigma_1, X_2/\sigma_2) = \tilde{X}$$

Explain the effect this transformation has upon the Mahalanobis distance for $\tilde{X}$, in terms of your answer to section (b)(i).

Q8 A sandwich bar is considering making updates to its ranges for those with specific dietary requirements. The sandwich bar provides three such ranges at present - gluten free, vegetarian, and vegan. Customers can choose to either eat their sandwich at a table in the bar ("eat in"), or take the sandwich away to eat outside the bar ("take away").

The sandwich bar decides to run a factorial experiment to explore the relationship between range, eating location, and sales. The ranges are used as the three levels for a factor $\mathcal{A}$, and the choice of where to eat is used as the two levels for a factor $\mathcal{B}$. Over two time periods, each of one week, each sandwich in one of the three ranges of $\mathcal{A}$, and the choice of where to eat it $\mathcal{B}$, is recorded. The resulting sales numbers, y, are provided in the table below.

$\mathcal{A}$ (range)	$\mathcal{B}$ (eating location)	
	eat in	take away
gluten free	41, 39	42, 46
vegetarian	62, 68	67, 71
vegan	47, 43	46, 40

(a) Define the terms "complete" and "balanced" for a factorial experiment, and state whether either term applies to the sandwich bar's experiment.

(b) Produce and interpret an interaction plot for this data.

(c) The dependence of y on $\mathcal{A}$ and $\mathcal{B}$ can be described by an interaction model. Note that you do **not** need to estimate the regression parameters of the interaction model in order to complete the tasks below.

(i) Justify the fact that the interaction model for this situation will include five dummy variables, despite there being six combinations of levels.

(ii) Write down explicitly the numerical values of the twelve residuals of this model.

(iii) Hence, calculate the unbiased estimator s^2 of the error variance under this model.