Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2024	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 Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

1.1 Show that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric and positive semidefinite, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$.

1.2 For a given data set, assume that $\boldsymbol{X}^T\boldsymbol{X}$ 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 $\boldsymbol{\beta}$.

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

Q2 The following linear regression model was fitted using the least squares approach, where $x_1 = 1$.

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \tag{1}$$

There are eight possible submodels of (1) that include an intercept. Each submodel is represented using the index set $\mathcal{I}$, for example, $\mathcal{I} = \{1, 3\}$ represents a model that includes the intercept and the predictor x_3, where $p_{\mathcal{I}}$ refers to the number of terms in that submodel. These submodels have been fitted, resulting in the following table of Mallows' $C_{\mathcal{I}}$ values:

$\mathcal{I}$	$\{1\}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$
$C_{\mathcal{I}}$	64.525	60.290	66.312	10.862
$\mathcal{I}$	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,3,4\}$	$\{1,2,3,4\}$
$C_{\mathcal{I}}$	62.252	9.973	3.567	??

2.1 Give the missing value ?? of $C_{\{1,2,3,4\}}$.

2.2 Carry out forward selection to select a suitable model.

2.3 Explain why we expect $C_{\mathcal{I}} \leq p_{\mathcal{I}}$ for "good" submodels and critically discuss your solution to (**2.2**) in this context.

SECTION B

Q3 For a linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with n observations and p parameters, the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is defined as the $n \times n$ matrix, which maps the vector of responses, $\boldsymbol{Y}$, to the vector of fitted values, $\hat{\boldsymbol{Y}}$. The diagonal values $h_i \equiv h_{ii}$, $i = 1, \ldots, n$, of $\boldsymbol{H}$ are called leverage values.

3.1 Show that, for the i-th residual $\hat{\epsilon}_i$ under the usual linear model assumptions, one has $\hat{\epsilon}_i/(\sigma\sqrt{1-h_i}) \sim N(0,1)$.
Hint: Show first that the residual vector $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ can be written as $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$.

3.2 *Cook's Distance* can be written as

$$D_i = \frac{1}{p}\, r_i^2\, \frac{h_i}{1-h_i}\,, i = 1, \ldots, n$$

where $r_i = \hat{\epsilon}_i/(s\sqrt{1-h_i})$ is the standardised residual corresponding to the ordinary residual $\hat{\epsilon}_i$, and s^2 is the usual unbiased estimate of the assumed common error variance. Describe briefly for what Cook's distance is used, and which information h_i contributes to it.

3.3 Suppose that $n = 54$, $p = 5$, $s^2 = 4$ and the following statistics for four of the cases were obtained:

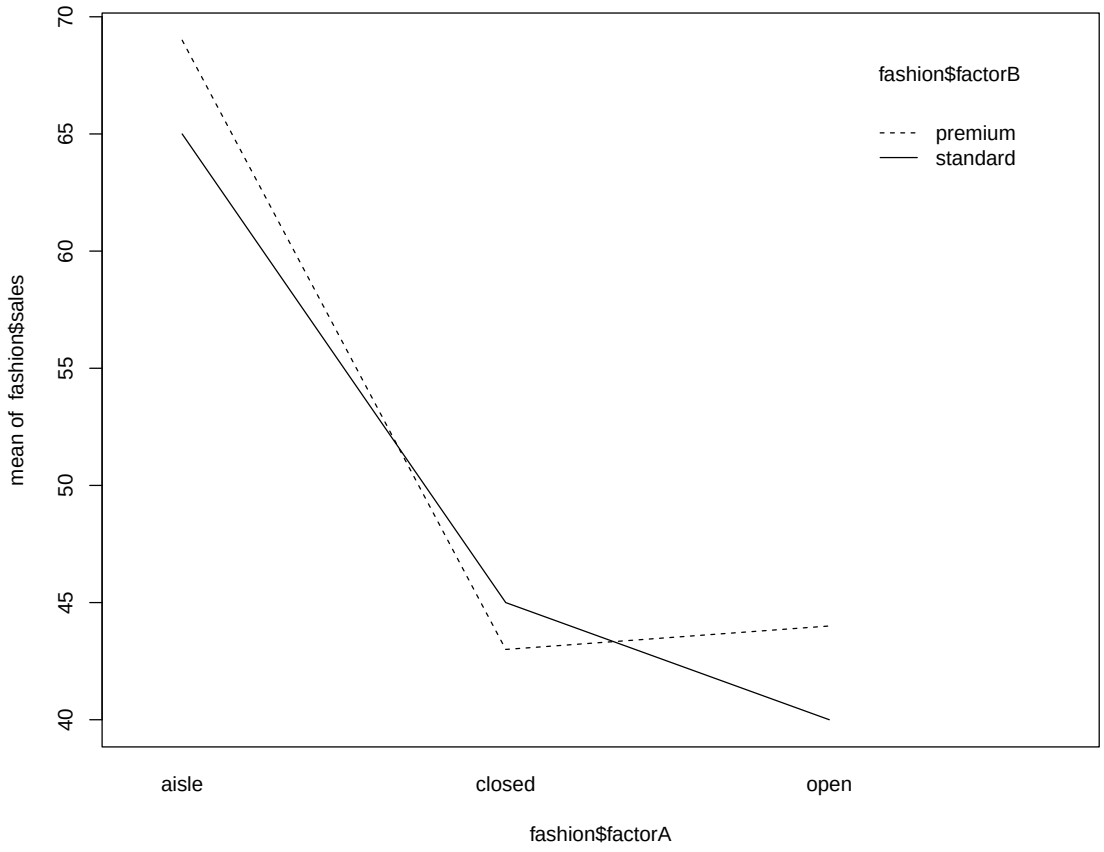
$\hat{\epsilon}_i$	h_i
0.6325	0.9000
1.7320	0.7500
9.0000	0.2500
10.2950	0.0185

For each case, compute the standardised residual r_i and Cook's distance D_i and state whether or not it appears to be (i) potentially influential, (ii) an outlier, (iii) influential.

Q4 A fashion retailer conducted an experiment to assess the influence of store layout (factor $\mathcal{A}$ with three levels: *aisle*, *closed*, and *open*) and promotional strategy (factor $\mathcal{B}$ with two levels: *premium* and *standard*) on the sales of their new clothing line over a specific time frame. A total of twelve boutique stores, similar in size, participated in the study, with the six experimental conditions assigned randomly across these stores. The recorded sales, y, are presented in the table below.

Factor $\mathcal{A}$	Factor $\mathcal{B}$	
	premium	standard
aisle	67, 71	62, 68
closed	40, 46	43, 47
open	42, 46	39, 41

4.1 Give an interpretation of the interaction plot below.



4.2 An interaction model can be used to describe how factor $\mathcal{A}$ and factor $\mathcal{B}$ affect the value of y. How many parameters would a full unconstrained interaction model have for this dataset? And how many parameters would a constrained model have? Describe an appropriate constraint which is commonly used.

Hint: A full *interaction* model can be written in the form,

$$y_{jk\ell} = \mu + \tau_j^{\mathcal{A}} + \tau_k^{\mathcal{B}} + \tau_{jk}^{\mathcal{AB}} + \epsilon_{jk\ell}$$

where $\tau_j^{\mathcal{A}}$ refers to the main effect of level j of factor $\mathcal{A}$ ($j = 1, \ldots, a$), $\tau_k^{\mathcal{B}}$ to the main effect of level k of factor $\mathcal{B}$ ($k = 1, \ldots, b$), $\tau_{jk}^{\mathcal{AB}}$ to the interaction effect of level j of factor $\mathcal{A}$ with level k of factor $\mathcal{B}$, and $\epsilon_{jk\ell}$ refers to the error term of replicate ℓ for the j, k factor combination ($\ell = 1, \ldots, r$).

4.3 Below is the output of fitting the model with the main effects and interaction to this data set. Comment on the significance of the inclusion of the interaction terms, then find the expected sales, y, for a store using a layout *aisle* and a *standard* promotional strategy.

```
> interfit<-lm(sales ~ factorA + factorB + factorA:factorB, data = fashion)
> summary(interfit)
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.900e+01  2.273e+00  30.356 8.48e-08 ***
factorAclosed                 -2.600e+01  3.215e+00  -8.088 0.000191 ***
factorAopen                   -2.500e+01  3.215e+00  -7.777 0.000238 ***
factorBstandard               -4.000e+00  3.215e+00  -1.244 0.259777
factorAclosed:factorBstandard  6.000e+00  4.546e+00   1.320 0.235013
factorAopen:factorBstandard    2.755e-14  4.546e+00   0.000 1.000000
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4.4 The ANOVA table is also provided for this interaction model. Comment on the significance of the inclusion of the interaction terms, and then calculate the sum of squares of the residuals for the non-interaction model.

```
> anova(interfit)
Analysis of Variance Table

Response: sales
               Df Sum Sq Mean Sq F value    Pr(>F)
factorA         2   1544  772.00 74.7097 5.754e-05 ***
factorB         1     12   12.00  1.1613    0.3226
factorA:factorB 2     24   12.00  1.1613    0.3747
Residuals       6     62   10.33
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4.5 Carry out the partial F-test for the model (`factorA + factorB + factorA:factorB`) vs. the model (`factorA`) at the 5% level of significance.