Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	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. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

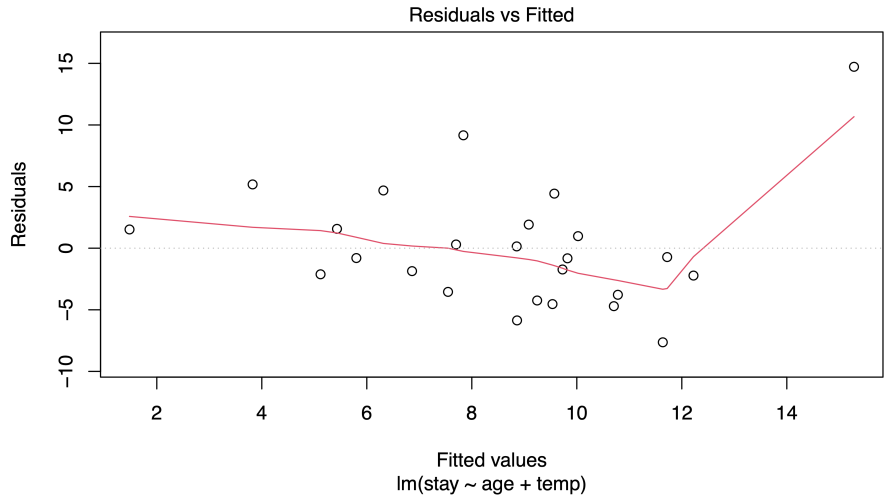
Revision:	

SECTION A

Q1 A random sample of $n = 25$ individuals was selected from a larger dataset of patients discharged from a Pennsylvania hospital. The response variable is the length of hospital stay (`stay`), with the predictors being the patient's age (`age`) and their first recorded temperature following admission (`temp`).

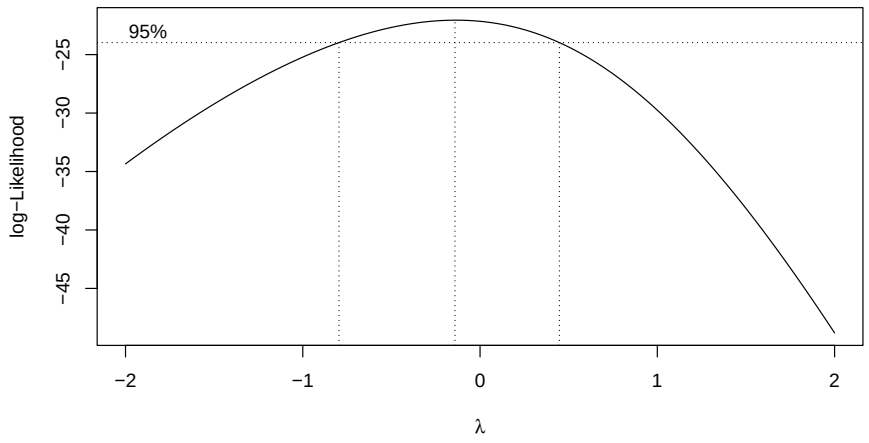
(a) State the usual assumptions underlying a linear regression model.

(b) A linear regression model is fitted with `stay` as the response variable and `age` and `temp` as predictors. A plot of residuals versus fitted values is provided. Interpret this plot to determine whether any violations of the assumptions of linear regression are present. If such violations exist, suggest possible remedies to address these issues and improve the model.



(c) A Box-Cox transformation is applied to the response variable, as shown in the figure below. Discuss the meaning of λ in this context, the purpose of the log-likelihood, and the distributional assumptions of the analysis.

(d) From the figure below, visually deduce a 95% confidence interval for λ. Does this imply a transformation for the response? If so, what transformation would you choose?



Q2 An experiment was designed to study the effects of three different drugs (D) and three types of stressful situations (S) in producing anxiety in adolescent subjects. There were 2 replications of each of the 3×3 factor combinations, resulting in a total of 18 scores.

Stressful situation (factor S)	Drug (factor D) D1	D2	D3
I	4 5	1 3	1 0
II	6 6	6 6	6 3
III	5 4	7 4	4 5

The two (partially edited) R output analysis-of-variance (ANOVA) tables shown below are for the *main effects plus interaction* model (D + S + D:S) and the single *main effect* model (D).

```
> anova(lm(Scores ~ Drug + Stress + Drug:Stress, data = dfStress))

            Df Sum Sq Mean Sq F value   Pr(>F)
Drug         2 10.778  5.3889  3.7308 0.066065 .
Stress       2 33.444 16.7222 11.5769 0.003247 **
Drug:Stress  4  9.889  2.4722  1.7115 0.230886
Residuals    9 13.000  1.4444
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(lm(Scores ~ Drug , data = dfStress))
            Df     Sum Sq     Mean Sq      F value     Pr(>F)
Drug         2      [A1]        [A2]        [A3]       0.269
Residuals   [A4]    56.333       3.7556
```

(a) Explain the statistical term **experiment**, explicitly stating the three main principles of experimental design.

(b) 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.

(c) Complete the missing entries [A1-A4] in the second ANOVA table. Note: You can use the fact that the sum of squares in the ANOVA table will be the same regardless of the order of fitting the main effects, but state explicitly which characteristic of the design of this experiment accounts for this property.

(d) Carry out the partial F-test for model (D + S + D:S) vs. model (D) at the 5% level of significance.

Q3 We analyse daily air quality data from New York City, collected between May and September 1973. After removing days with missing values, the dataset consists of $n = 111$ observations. The response variable, ozone concentration (in ppb), is modelled using the transformed form `Ozone^(1/3)`, as suggested by previous studies. The predictors are solar radiation (`Solar.R`, in Langleys), average wind speed (`Wind`, in mph), and maximum daily temperature (`Temp`, in F), all continuous variables. Consider the full linear model based on these variables.

```
Ozone^(1/3)= 1+ Solar.R + Wind + Temp
```

Fitting this model leads to the following Analysis of Variance (ANOVA) table:

```
           Df Sum Sq Mean Sq F value    Pr(>F)
Solar.R     1 15.531 15.5314  59.645 6.435e-12 ***
Wind        1 26.326 26.3260 101.100 < 2.2e-16 ***
Temp        1 17.489 17.4890  67.163 5.860e-13 ***
Residuals 107 27.862  0.2604
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) How many submodels does this model possess (when always including the intercept)?

(b) A commonly used criterion for model selection is Mallows' $C_{\mathcal{I}}$, defined as:

$$C_{\mathcal{I}} = \frac{\mathrm{RSS}_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n, \tag{1}$$

where s^2 is the estimate of the error variance σ^2 under the full model, and $\mathcal{I}$ denotes the indices of the predictors retained in the submodel. Here, $p_{\mathcal{I}}$ is the number of predictors in the submodel, and $\mathrm{RSS}_{\mathcal{I}}$ is the residual sum of squares for the submodel. Explain how Mallows' $C_{\mathcal{I}}$ criterion addresses the trade-off between model parsimony and goodness-of-fit.

(c) Show that Mallows' $C_{\mathcal{I}}$ in (1) can be expressed as $C_{\mathcal{I}} = p_{\mathcal{D}}(F_{\mathcal{D}} - 1) + p_{\mathcal{I}}$, where $F_{\mathcal{D}}$ is the F-statistic for testing the null hypothesis H_0 that $\boldsymbol{\beta}_{\mathcal{D}} = \mathbf{0}$, with $\mathcal{D} = \{1, 2, \ldots, p\} \setminus \mathcal{I}$, and $p_{\mathcal{D}} = p - p_{\mathcal{I}}$, where p is the total number of predictors in the full model.

(d) Using Equation (1) and the ANOVA table, compute the values of $C_{\mathcal{I}}$ for the following submodels, and based on these values, determine the most appropriate model:

(i) `1`

(ii) `1 + Solar.R`

(iii) `1 + Solar.R + Wind`

(iv) `1 + Solar.R + Wind + Temp`

(e) Is it possible to compute the values of $C_{\mathcal{I}}$ for the remaining $x - 4$ submodels (where x is your result from part (a)) based on the information provided in the ANOVA table? If so, briefly explain how this can be done. If not, provide an explanation as to why this is not possible.

Q4 A hospital administrator wishes to study the relationship between patient satisfaction and patient's age, severity of illness, and anxiety level. The administrator selected $n = 46$ patients at random and collected data on the following four variables:

`sat`:	an index for patient satisfaction
`age`:	the patient's age (in years)
`sev`:	an index for severity of illness
`anx`:	an index for the anxiety level

A linear model for `sat` versus the three predictor variables is fitted. The corresponding R output is provided below. We denote the model parameters for the intercept, `age`, `sev`, and `anx` by β_1, β_2, β_3, and β_4, respectively.

```
> fit <- lm(sat ~ age + sev + anx, data = patients)
> fit

Call:
lm(formula = sat ~ age + sev + anx, data = patients)

Coefficients:
(Intercept)          age          sev          anx
    158.491       -1.142       -0.442      -13.470

> s<- summary(fit)$sigma
> s
[1] 10.05798
> summary(fit)$cov.unscaled
             (Intercept)           age           sev           anx
(Intercept)  3.247711653  0.0092211391 -0.0679307897 -0.067298817
age          0.009221139  0.0004560816 -0.0003185955 -0.004662271
sev         -0.067930790 -0.0003185955  0.0023924814 -0.017710085
anx         -0.067298817 -0.0046622713 -0.0177100848  0.498257730
```

(a) State the estimated regression function. What is the predicted satisfaction score of a patient aged 54, with a severity of illness of 40 and an anxiety level of 2.5?

(b) Show that the standard error of the parameter estimate for age, $SE(\hat{\beta}_2)$, is given as 0.2148.

(c) Using 4(b), provide the t-value corresponding to $\hat{\beta}_2$. Decide whether the age parameter is significantly different from 0 at the 5% significance level. *Hint:* If the provided statistical tables do not list the exact degrees of freedom, use the critical t-value for the nearest available degrees of freedom.

(d) Compute a 95% confidence interval for β_2, and provide an interpretation in light of your answer to 4(c).

(e) For the detection of potentially influential observations, work out the numerical value of our rule of thumb $(2p/n)$. Which cases are detected as potentially influential according to this criterion? The R output required to complete this task is provided on the following page.

```
> infl<- lm.influence(fit)
> round(infl$hat,3)
    1     2     3     4     5     6     7     8     9    10
0.078 0.067 0.037 0.154 0.097 0.129 0.034 0.075 0.184 0.058
   11    12    13    14    15    16    17    18    19    20
0.088 0.031 0.090 0.033 0.143 0.047 0.120 0.062 0.034 0.129
   21    22    23    24    25    26    27    28    29    30
0.078 0.137 0.033 0.136 0.043 0.103 0.087 0.186 0.059 0.090
   31    32    33    34    35    36    37    38    39    40
0.117 0.110 0.045 0.037 0.103 0.027 0.121 0.071 0.181 0.087
   41    42    43    44    45    46
0.038 0.154 0.061 0.051 0.073 0.083
```