EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2023	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:	

ED01/2023

SECTION A

Q1 **1.1** What are the three assumptions underlying the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$?

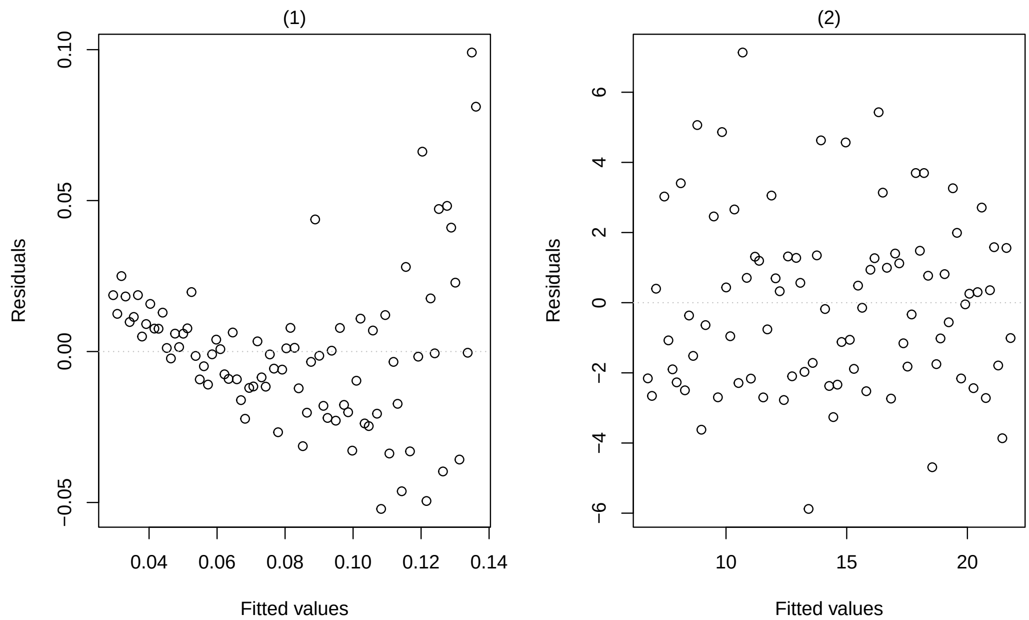
1.2 In the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, show that $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{\epsilon}} = 0$, where $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are the fitted values, and $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ are the residuals. Explain why this result shows that the fitted values and residuals are empirically uncorrelated when there is an intercept term in the model. What are the implications of this result for model diagnostics?

1.3 Fitting the two models:

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

$$Y^{-1} = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

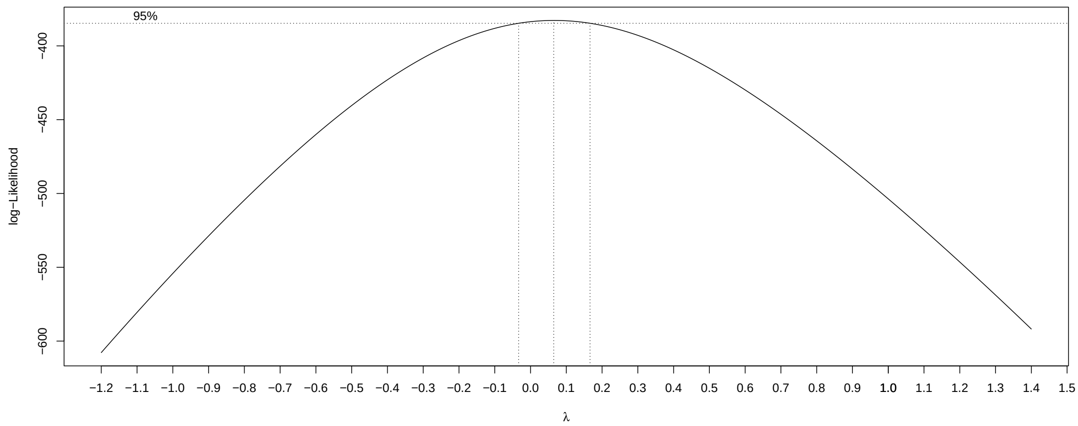
leads to the following residual plots:



Which of the two residual plots indicates the better model fit and why?

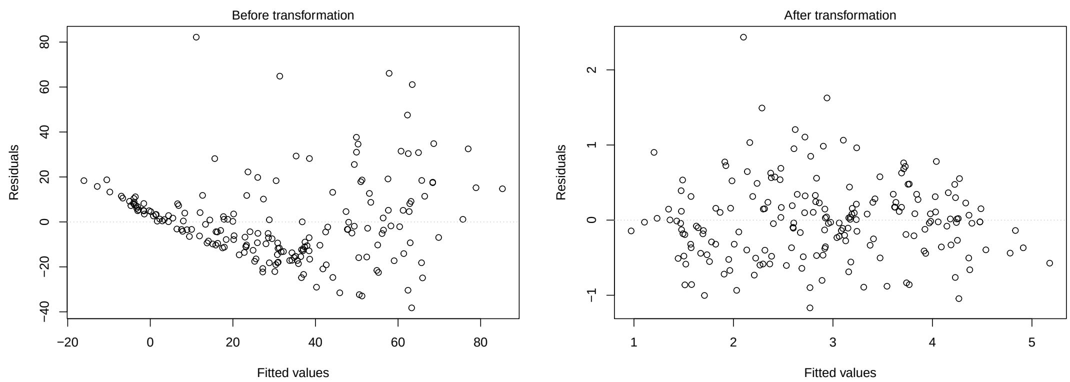
Q2 **2.1** Write down the general expression for the transformed response $y^{(\lambda)}$ used in the Box-Cox transformation of a positive response variable.

2.2 For a particular linear model, the graph of the profile log-likelihood for λ, $L_p(\lambda)$, is provided below.



Read from this graph (approximately) the value of the estimate $\hat{\lambda}$ as well as a 95% confidence interval for λ. Does this suggest a need for a transformation to be applied to the response? Would a logarithmic transformation be appropriate?

2.3 The figures below show two residual plots for the model with untransformed and log-transformed response variable, respectively. Give an interpretation of these plots. Has the log-transformation led to an improvement?



SECTION B

Q3 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]    [A5]       3.7556
```

3.1 Complete the missing entries [A1-A5] 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.

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

3.3 Select the best of the four models D; S; D+S; D+S+D:S according to Mallows' $C_{\mathcal{I}}$. Hint: Mallows' $C_{\mathcal{I}}$ is given by $C_{\mathcal{I}} = \frac{\text{RSS}_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n$, where $\mathcal{I}$ is an index set representing the variables included in the model, and $p_{\mathcal{I}}$ its cardinality.

Q4 From a study investigating the relation of body fat to several possible predictor variables, we are given a sample of $n = 20$ healthy males in their twenties on the following five variables:

`PctBF`	body fat percentage (Response variable)
`Height`	Height (cm)
`Chest`	Chest circumference (cm)
`Waist`	Waist circumference (cm)
`Hip`	Hip circumference (cm)

A linear model is fitted using R. The (edited) R output provided is needed to answer the questions below. In what follows, we denote by β_0 the intercept parameter, and by β_1, β_2, β_3 and β_4 the linear model parameters corresponding to the predictors `Height`, `Chest`, `Waist`, and `Hip`, respectively.

4.1 Provide the missing $SE(\hat{\beta}_2)$ (denoted by A in the R output) and use this result to provide a 95% confidence interval for β_2.

4.2 What is the unbiased estimate s^2 of the assumed common error variance σ^2?

4.3 Provide the missing F-statistic (denoted by B in the R output), and interpret the result of this F-test. Hint: recall that $F = \frac{SSR/(p-1)}{SSE/(n-p)}$ and $R^2 = \frac{SSR}{SSR+SSE}$, and use the result from the previous part to find SSE.

4.4 For a linear model with n observations and p parameters, the hat matrix $\boldsymbol{H}$ 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, $i = 1, \ldots, n$, of $\boldsymbol{H}$ are called leverage values. Show that generally $\text{Tr}(\boldsymbol{H}) = p$, and use this result to compute the missing leverage value h_7 (denoted by C in the R output).

4.5 For the detection of potentially influential observations, work out the numerical value of our rule-of-thumb $(2p/n)$. Which case is detected to be *potentially influential* according to this criterion?

```
> BodyFatmod <- lm(PctBF ~  Height + Chest + Waist + Hip , data = bodyfat)
> summary(BodyFatmod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.2572    25.5133   1.304  0.21205
Height       -0.8324     0.4107  -2.027  0.06084 .
Chest        -0.1752        [A]    -0.703  0.49303
Waist         2.6465     0.6840   3.869  0.00151 **
Hip          -0.3316     0.4001  -0.829  0.42021
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.656 on 15 degrees of freedom
Multiple R-squared:  0.8313,Adjusted R-squared:  0.7863
F-statistic:  [B]   on 4 and 15 DF,  p-value: 1.157e-05
```

```
> fat.infl<- lm.influence(BodyFatmod)
> fat.infl$hat
        1         2         3         4         5         6         7
0.2357360 0.1385595 0.4010487 0.1085318 0.6660972 0.1419920        [C]
        8         9        10        11        12        13        14
0.2703768 0.1887516 0.1208124 0.2483467 0.1461117 0.3252978 0.1183943
       15        16        17        18        19        20
0.2645119 0.3307922 0.1761983 0.3693740 0.2299857 0.2965986
> sum(fat.infl$hat[-7])
[1] 4.777517
```