Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2026	MATH2697-WE01

Title: Statistical Modelling II

Time:	2 hours	
Additional Material provided:	Tables for chi square, F, Normal, and t distributions	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. The indicative marks shown in brackets for the main parts of each question are given as a guide to the weighting the markers expect to apply. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2026

SECTION A

1. (a) What are the three assumptions underlying the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$? **[3]**

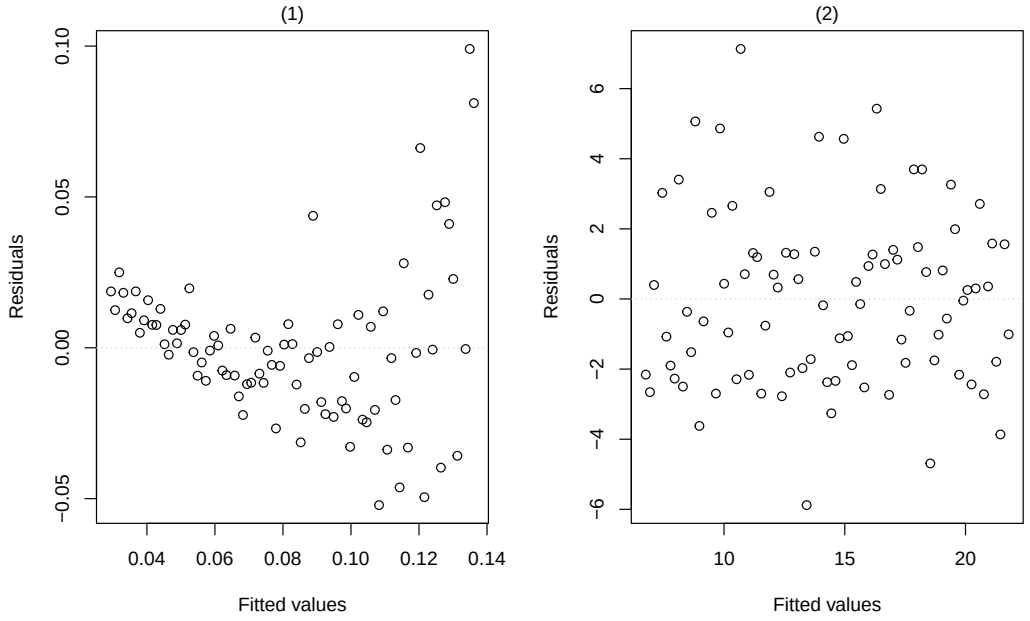
(b) In the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, show that $\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = 0$, where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are the fitted values, and $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$ 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? **[4]**

(c) 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? **[3]**

---

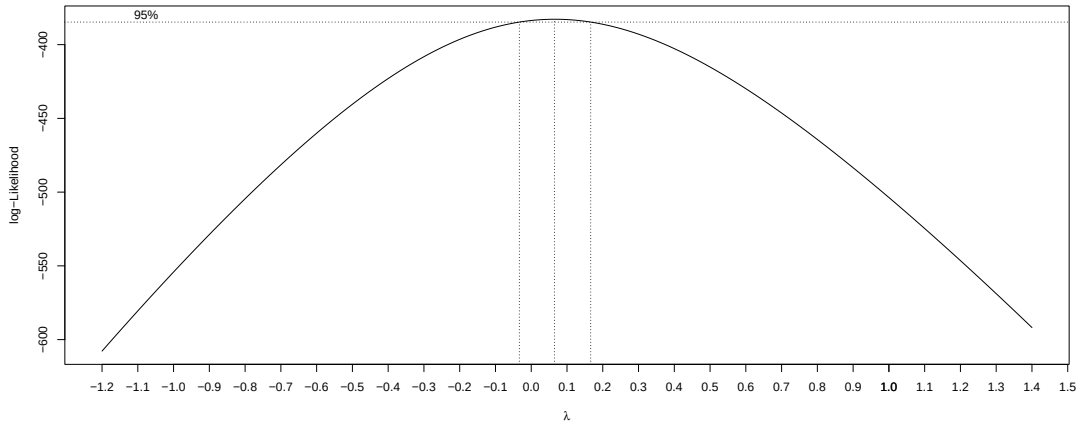
2. (a) Write down the general expression for the transformed response $y^{(\lambda)}$ used in the Box-Cox transformation of a positive response variable, and state any assumptions the transformed response needs to satisfy. **[3]**

(b) Assuming that $y_i^{(\lambda)}$ meets the standard linear model assumptions. Write down the probability density function of $y_i^{(\lambda)}|\mathbf{x}_i$. From this, find the probability density function of $y_i|\mathbf{x}_i$. Then, derive the log likelihood $L(\boldsymbol{\beta}, \sigma^2, \lambda)$ based on independent observations $y_1, \ldots, y_n$.

Hint: the probability density function of a normal distribution with mean μ and variance σ^2 is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$ **[3]**

(c) 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? **[4]**

SECTION B

3. 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 symbolises the intercept term and `variety:density` symbolises 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.

(a) 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]**

(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. **[3]**

(c) 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 $F_{0.05}(df_3, df_{res}) = 2.51$. **[6]**

(d) 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]**

4. Data on violent crime rates (`y`) and three predictor variables, poverty rate (`x1`), percentage living in urban areas (`x2`), and percentage of single-parent families (`x3`), were collected for 30 U.S. states. A linear model was fitted; its summary and supplementary information are shown in the R output below.

(a) In the summary table, provide the missing values of A, B, and C, and a meaningful lower bound for D. **[4]**

(b) Which of the four regression parameters are significantly different from zero at the 5% level of significance? **[2]**

(c) Without carrying out calculations, specify which of the 95% confidence intervals for the four parameters contains zero and which do not. Then, explicitly calculate the confidence interval for the parameter associated with the `x1` variable. **[3]**

(d) We are interested in predicting the violent crime rate for a state, with a poverty rate of 12, a percentage of those living in urban areas of 70% and a 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. **[6]**

```
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,5)
            (Intercept)       x1       x2       x3
(Intercept)     1.42117 -0.04759 -0.00842 -0.00868
x1             -0.04759  0.00607  0.00047 -0.00245
x2             -0.00842  0.00047  0.00013 -0.00027
x3             -0.00868 -0.00245 -0.00027  0.00236
```