University of Glasgow

December 2018
1 hour 30 mins

EXAMINATION FOR THE DEGREE OF MASTERS (SCIENCE)

Regression Modelling

This paper consists of 6 pages and contains 3 questions.
Candidates should attempt all questions.

Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Total	60 marks

The following material is made available to you:

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".
"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

NOTE: Candidates should attempt all 3 questions. Questions are equally weighted.

CONTINUED OVERLEAF/

1. (a) For the following linear models, write down the standard vector-matrix form $E(\mathbf{Y}) = X\boldsymbol{\beta}$ clearly identifying $\mathbf{Y}, X$, and $\boldsymbol{\beta}$:

   (i) Data: $y_{ij}, i = 1, 2; j = 1, \ldots, n_i$.

   Model: $E(Y_{ij}) = \mu + \alpha_i$. **[2 MARKS]**

   (ii) Data: $(y_i, x_i), i = 1, \ldots, n$.

   Model: $E(Y_i) = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2$. **[2 MARKS]**

   (iii) Data: $(y_i, x_i), i = 1, \ldots, 2n$.

   Model: $E(Y_i) = \alpha + \beta x_i + \gamma D_i + \delta x_i D_i$, where $D_i = 1$ for $i = 1$ and 0 otherwise. **[3 MARKS]**

   (b) Define the vector of fitted values in vector-matrix notation and show that it is a linear combination of the responses. **[3 MARKS]**

   (c) Write down the formula for the estimator for the variance of the error terms σ^2 for the following linear model:

   Data: $y_{ij}, i = 1, 2; j = 1, \ldots, n_i$.

   Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_i$, where we assume that $\varepsilon_i \sim N(0, \sigma^2 I)$ **[2 MARKS]**

   (d) Explain how residual plots can be used to examine the normal linear model assumptions. Sketch two residual plots where a model assumption has failed and identify which assumption has failed in each case. **[4 MARKS]**

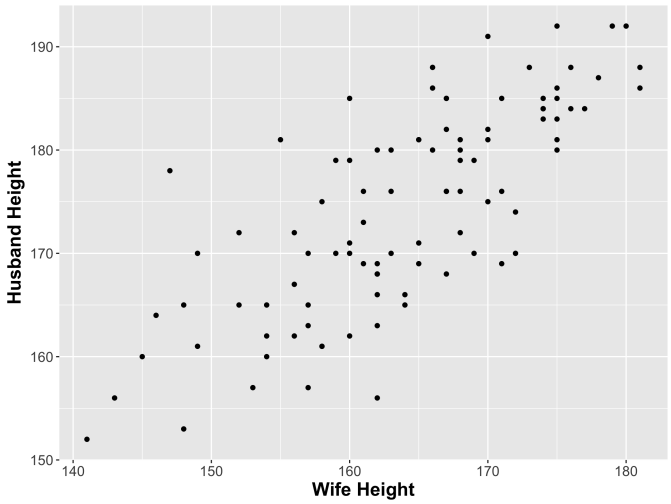
   (e) Explain what you could do if the assumption of linearity in a normal linear model is violated. **[4 MARKS]**

Total: 20 MARKS

CONTINUED OVERLEAF/

2. (a) A medical doctor has collected data for the age, weight, height and body mass index of 90 of her patients and she would like to know whether these variables are related to blood pressure using linear regression modelling. Describe one of the model selection methods seen in the lecture notes to find a best-fitting linear regression model; for example a stepwise regression method or best subset selection. Indicate the starting model, one of the criteria to compare models at each iterative step, and when the iterative method would end. **[6 MARKS]**

(b) We have available a data set that contains the heights in centimetres of 96 newly married wives and husbands. We are interested in knowing if there is a relationship between the wife and the husband's heights and in estimating the husband's height depending on the wife's height. Let X be the wife's height and Y the husband's height. We can see a scatter plot of these data below.



We propose the simple linear regression model:

$$\text{HusbandHeight}_i = \beta_0 + \beta_1 \text{WifeHeight}_i + \varepsilon_i, i = 1, \ldots, 96$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and ε_is are independent for $i = 1, \ldots, 96$.
Some of the results of fitting this model are shown below:

$$\widehat{\text{HusbandHeight}} = 37.81 + 0.83\,\text{WifeHeight}.$$

	Coef.	se(Coef)
Intercept	37.81	11.93231
WifeHeight	0.833	0.07269

CONTINUED OVERLEAF/

Analysis of Variance Table

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
Model	5492.5	1	a	b	< 2.2e-16 ***
Residuals	c	94	41.8		
Total	9425	95			

$$(X^T X)^{-1} = \begin{pmatrix} 3.40336672 & -0.0207018689 \\ -0.0207018689 & 0.0001263111 \end{pmatrix}.$$

where X is the design matrix.

i. Complete the ANOVA table above with the values of `a`, `b` and `c`. **[4 MARKS]**

ii. Calculate and comment on the value of R^2 for this model. **[2 MARKS]**

iii. Calculate a 95% confidence interval for β_1 and comment. **[4 MARKS]**

iv. Calculate a 95% prediction interval for the husband's height when the wife's height is 150 cm. **[4 MARKS]**

CONTINUED OVERLEAF/

3. In an investigation of a new diet to help with blood sugar level, 30 patients were assigned to a control group and 30 patients assigned to the new diet. The final weight of patient j on diet i is denoted by Y_{ij} and their initial blood sugar level is denoted by x_{ij}.

   Consider the Linear Model (Model A).

$$Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i\cdot}) + \varepsilon_{ij}, i = 1, 2; j = 1, \ldots, 30$$

   and where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and are independent.

   (a) Starting from Model A defined above, write down the two simplified models that might also be used in this setting. Describe in words what each model, including Model A, means.

   [5 MARKS]

   (b) Write down a vector-matrix expression for the Model A clearly identifying $\mathbf{Y}$, X and $\boldsymbol{\beta}$. Hence find the least squares estimates for $\boldsymbol{\beta}$. **[3, 4 MARKS]**

   For part (c)-(f) choose the correct answers from the four options. Note that in some cases more than one answer may be correct.

   (c) In a regression study, a 95% confidence interval for β_1 was given as: (-5.65, 2.61). What would a test for $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ conclude? **[2 MARKS]**

   i. Reject the null hypothesis at α=0.05 and all smaller α.

   ii. Fail to reject the null hypothesis at α=0.05 and all smaller α.

   iii. Reject the null hypothesis at α=0.05 and all larger α.

   iv. Fail to reject the null hypothesis at α=0.05 and all larger α.

   (d) If a predictor variable x is found to be highly significant we would conclude that: **[2 MARKS]**

   i. A change in y causes a change in x.

   ii. A change in x causes a change in y.

   iii. Changes in x are not related to changes in y.

   iv. Changes in x are associated with changes in y.

   (e) The following appeared in the magazine Financial Times, March 23, 1995: "When Elvis Presley died in 1977, there were 48 professional Elvis impersonators. Today there are an estimated 7328. If that growth is projected, by the year 2112 one

CONTINUED OVERLEAF/

person in four on the face of the globe will be an Elvis impersonator." This is an example of: **[2 MARKS]**

i. Extrapolation.

ii. Dummy variables.

iii. Misuse of causality.

iv. Multicollinearity.

(f) Which of the following ANOVA components are not additive? **[2 MARKS]**

i. Mean squares.

ii. Sum of squares.

iii. Degrees of freedom.

iv. None of the above.

Total: 60 MARKS

END OF QUESTION PAPER.