University of Glasgow

Tuesday, 20th May 2014
2.00 pm – 3.30 pm

EXAMINATION FOR THE DEGREE OF M.SC. (TAUGHT) (SCIENCE)

REGRESSION MODELLING – LEVEL M

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 THREE out of the FOUR questions. If more than three questions are attempted please indicate which questions should be marked; otherwise, the first three questions will be graded.

1. (a) Which of the following are linear models in terms of α and β? You may assume that Y is the response of interest and $\varepsilon \sim N(0, \sigma^2)$.

(i) $Y = \log(x^{\alpha}) + \varepsilon$
(ii) $Y = \alpha + \beta^{-1}(1/x) + \varepsilon$
(iii) $Y = \alpha(x^{\beta}) + \varepsilon$

[3 MARKS]

(b) Write down the standard vector-matrix form for the Normal Linear Model (NLM), $E(Y) = A\theta$, identifying $\underline{Y}$, A and $\underline{\theta}$ for each of the three models below. You may assume $\varepsilon \sim N(0, \sigma^2)$.

(i) $Y_j = \alpha + \beta(x_j - \bar{x}.) + \varepsilon_j$; j=1,2,....n

(ii) $Y_j = \alpha + \beta x_j + \gamma x_j^2 + \varepsilon_j$; j=1,2,....n

(iii) $Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_i.) + \varepsilon_{ij}$; i=1,2; j=1,….$n_i$.

(iv) Derive the ordinary least squares (OLS) estimates of α and β, in model 1b(*i*) stating clearly any formulae you use.

[6, 4 MARKS]

(c) State the Gauss Markov theorem and explain its importance. **[3 MARKS]**

(d) What are the benefits of centring the covariates in the simple linear regression model. **[2 MARKS]**

CONTINUED OVERLEAF/

(e) Describe the R^2 and adjusted R^2 statistics and explain how they are interpreted.

[2 MARKS]

2. (a) In a study of fat production in meat, the effects of the concentrations of two acute phase proteins x_1 and x_2 (in mg per kg) on the fat production Y, in meat samples are being studied. Thirty pigs have had their fat content measured as well as the two acute phase proteins.

$$\text{Model: } Y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \varepsilon_i; \quad i=1,\ldots 30; \text{ where } \varepsilon_i \sim N(0,\sigma^2),$$

$$\varepsilon_i \text{ are independent.}$$

(i) Write down a vector-matrix expression for the model A clearly identifying $\underline{Y}$, A and $\underline{\theta}$.

[2 MARKS]

This model has been fitted to the data, giving the following summary statistics.

$\hat{\alpha}$= -1.408, $\hat{\beta}$ = -0.5322, $\widehat{\square}$ = 3.2535, r= 120.22, $\Sigma(y_i-\bar{y})^2$ = 337.89.

$$(A^TA)^{-1} = \begin{matrix} 0.24746 & -0.05801 & -0.06876 \\ -0.05801 & 0.04165 & -0.00761 \\ -0.06876 & -0.00761 & 0.04863 \end{matrix}$$

(ii) By calculating appropriate interval estimates, show that the term involving x_2 should be retained in the model but that the term involving x_1 should be dropped.

[6 MARKS]

(iii) Calculate R^2 for the fitted model given the output above and comment.

[3 MARKS]

(b) Define a residual and a standardised residual for the normal linear model. Explain how residuals can be used in a graphical way to examine the following two assumptions

(i) the errors are normally distributed and

(ii) the errors have constant variance.

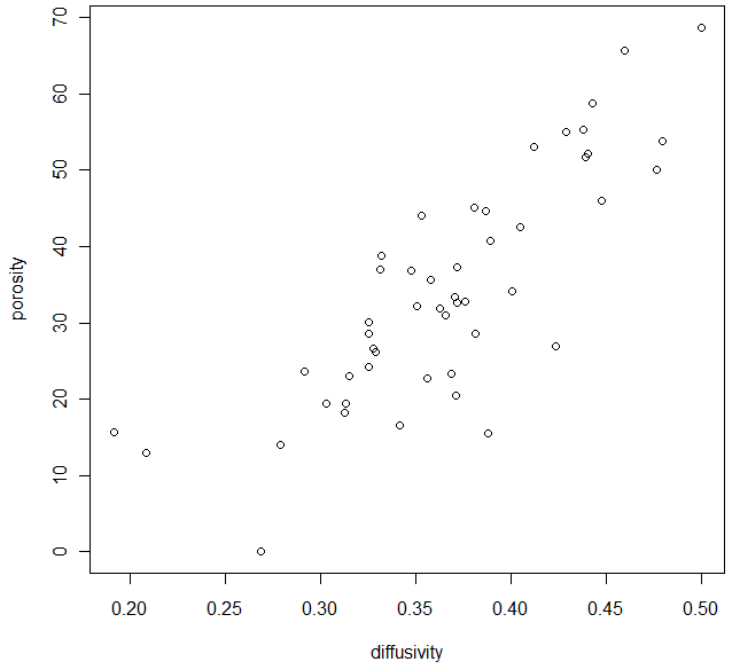
[1, 2, 2 MARKS]

CONTINUED OVERLEAF/

(c) Consider a Normal Linear Model (NLM) for a response variable Y in terms of a (possibly large) number of explanatory variables $X_1,\ldots,X_p$. Describe the algorithm for forward stepwise selection of explanatory variables including methods for determining the size of the model.

[4 MARKS]

3. A geochemist has been studying soil properties and exploring the relationship between diffusivity and porosity. A scatterplot of the results from his soil samples is shown below.



He has fit a simple linear model and an extract of his results from fitting the model in R are shown below.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-36.092	7.211	-5.005	8.26e-06 ***
diffusivity	191.737	19.374	9.897	4.47e-13 ***

Residual standard error: 8.587 on 47 degrees of freedom
Multiple R-squared: 0.6757, Adjusted R-squared: 0.6688
F-statistic: 97.94 on 1 and 47 DF, p-value: 4.475e-13

$(A^TA) = \begin{matrix} 0.70511 & -1.86690 \\ -1.86690 & 5.09026 \end{matrix}$

CONTINUED OVERLEAF/

(a) Calculate a 95% confidence interval for β and comment on its value.

[4 MARKS]

(b) A new soil sample with diffusivity of 0.25 has been collected. Provide a 95% prediction interval for the porosity of this sample. Explain how this differs from a confidence interval for the mean porosity when diffusivity is 0.25.

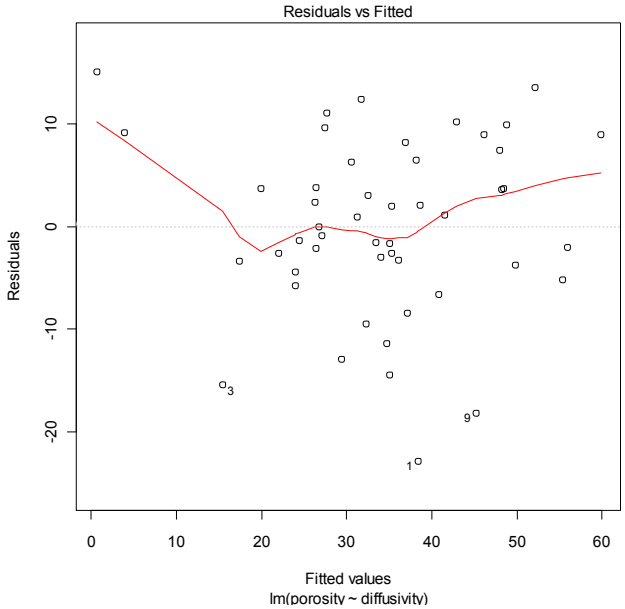
[6 MARKS]

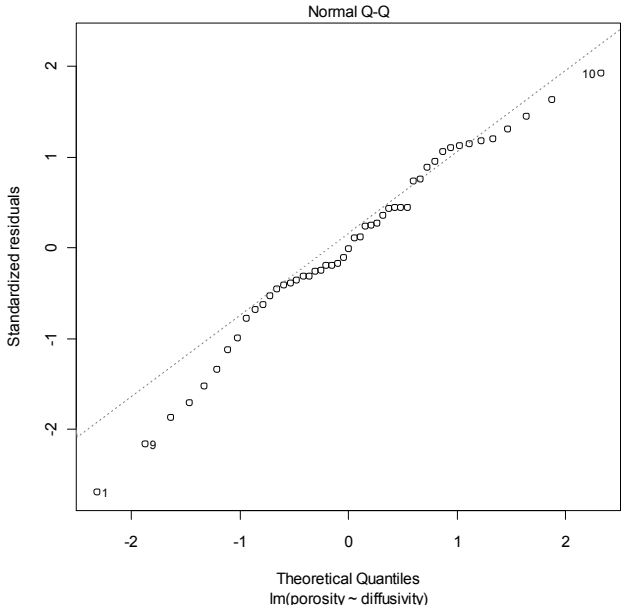
(c) Comment on the R^2 value.

[1 MARK]

(d) Two residual plots are shown below. Comment on their interpretation.

[2 MARKS]





CONTINUED OVERLEAF/

(e) Define the *leverage* of the i^{th} observation and explain how it can be interpreted.

[3 MARKS]

(f) In a first year practical class, the grip strength and hand width of 92 students were measured. T*he estimated correlation coefficient between handwidth and grip strength is **0.579.*** Using statistical tables, construct a 95% confidence interval for ρ, the population correlation coefficient and comment on its value.

[4 MARKS]

4. (a) The One-Way Analysis of Variance model for a continuous response variable, *Y*, that is measured on experimental units that are divided into *3* groups, with n_i observations in the *i*th group, can be written as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; j = 1,...,n_i, i = 1,...,3$$

where the ε_{ij} are independent $N(0, \sigma^2)$ random variables.

(i) Derive the normal equations specific to this problem. **[6 MARKS]**

(ii) Explain why it is not possible to solve the normal equations obtained in (*i*) in their given form, and then outline an approach to making them soluble.

[3 MARKS]

(iii) Using the method outlined in (*ii*), solve the normal equations to obtain a set of parameter estimates for the One-Way ANOVA model.

[3 MARKS]

(b) A vet wishing to study the effect of two different analgesic drugs on heart rate during a surgical procedure has measured the change in heart rate (y), 5 minutes after the introduction of the analgesic as well as the initial heart rate (x) on thirty horses (15 under each analgesic drug).

Consider the model

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}; \text{ i=1, 2; j=1,....15}$$

where i identifies the drug (i=1,2) and j identifies the horse (j=1,…,15). The independent errors ε_{ij} are assumed to have a Normal distribution with variance σ^2.

(i) By writing down the model in the standard vector matrix notation, show that constraints are necessary to ensure that the parameters can be estimated uniquely. Introduce one such constraint and re-write the model in vector matrix form.

[4 MARKS]

(ii) What differences in the relationship between change in heart rate and initial heart rate does the model proposed above allow? Describe one more complex model that includes another difference, and sketch the relationships for the two models on two separate plots.

[4 MARKS]

END OF QUESTION PAPER.