W1-2-60-1-6

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

University Examinations 2016/2017

THIRD YEAR FIRST SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCES

STA 2413: REGRESSION MODELLING

DATE: DECEMBER 2016 **TIME: 2 HOURS**

NOTE: Answer All parts of questions ONE and any other TWO on the answer booklet provided

1. (a) For the regression model

$$Y = X\beta + \epsilon$$

with $\epsilon \sim N(0, \sigma^2 W)$ where W is not an identity matrix, a student argues that the ordinary multiple regression cannot be applied and one should regress $W^{-1}Y$ against $W^{-1}X$ by ordinary multiple regression to estimate β. Do you agree with the student? If so, show how this can be done (6 marks)

(b) An Actuarial Science student is trying to analyze data from a particular stock using an equation of the form $y = f(x) = \frac{a_0}{x} + \frac{a_1}{x^2}$. The measured values of (x, y) are listed in the table below

x	1	2	3	4	5	6	7	8	9	10
y	3	7	0.9	6.8	0.7	0.96	19	37	0.8	0.1

Use nonlinear regression method to determine a_0 and a_1 (7 marks)

(c) Define the term Artificial Neural Network (ANN) (2 marks)

(d) Stock prices (Y, in dollars) are assumed to be affected by the annual rate of dividends of stock (X). A simple linear regression analysis ($Y_i = \beta_o + \beta_1 X_i + \epsilon_i$) was performed on 21 observations where ϵ_i were assumed to follow $\sim iidN(0, \sigma^2)$ and summary statistics were as listed below

$$\bar{X} = 0.4, \bar{Y} = 4, S_X^2 = 0.1, S_Y^2 = 20, S_{XY} = 1.25$$

Perform an ANOVA analysis for the level of significance 5% and give your conclusions (7 marks)

(e) Let $Y = (9, 4, 1, 3)^T$, $X_1 = (9, 3, 9, 4)^T$, $X_2 = (5, 3, 5, 4)^T$

CONTINUED

(i) Find the matrix $(X^TX)^{-1}$ (2 marks)

(ii) Fit a regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Give the fitted coefficients and estimate of error variance (4 marks)

(iii) Construct the 95% confidence interval for $E(Y|X_1 = 6, X_2 = 4)$ (3 marks)

2. In a study of the effect of hormones on the productivity of a certain variety of tomato plant, ten plants were treated at different hormone strengths and their yields, in Kgs, noted. A simple linear regression model is fitted in R, with the depended variable (yield in kgs) stored as the vector y and the independent variable (hormone strength) stored as the vector x. Some of the R commands and edited output are shown below

```
> lm1 = lm(y~x)
> summary(lm1)

Coefficients:
             Estimate    Std. Error
(Intercept) -3.72335     0.28072
      x      0.42999     0.01899

> qt(0.975,8)
[1] 2.306004
```

(a) Write down the equation of the model that has been fitted to the data, defining your notations carefully. State the distribution of any error terms in your model (5 marks)

(b) If X is the design matrix for the model fitted in R, then

$$(X^TX) = \begin{pmatrix} 10.60278 & 154.047 \\ 154.047 & 2320.782 \end{pmatrix}$$

Also,

$$\sum_{i=1}^{10} y_i = 26.76 \qquad \sum_{i=1}^{10} x_i y_i = 424.33$$

where y_i is the ith yield and x_i is the ith hormone strength. Give suitable calculations that show how the parameter estimates have been obtained in the R output (5 marks)

(c) If $x_3 = 11$ and $y_3 = 1.66$, calculate the corresponding fitted value and residual (4 marks)

(d) Test the hypothesis that there is no relationship between hormone strength and yield, stating your conclusion clearly. State the size of your hypothesis test (4 marks)

(e) Calculate the 95% confidence interval for the intercept and gradient in the regression model (2 marks)

CONTINUED

3. (a) Consider data from the simple linear model $y_i = \beta_o + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \cdots, n$ where x_i's are fixed constants, β_o, β_1 are the unknown coefficients and ϵ_i's are unobserved i.i.d random variables from $N(0, \sigma^2)$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimators of β_0 and β_1 respectively. Find the distribution of the vector $(\hat{\beta}_0, \hat{\beta}_1)$ (6 marks)

(b) Suppose the model is modified by including more independent variables so that the model is now written as $Y = X\beta + \epsilon$ and assuming same assumptions for the error model distribution.

(i) Show that the vector covariance of residuals $e = y - X\hat{\beta}$ and y is given by $cov(e, y) = \sigma^2 M$ where M is the idempotent and symmetric matrix (4 marks)

(ii) Suppose the variance of the error term (σ^2) is unknown, demonstrate how such can be estimated from the model given in b(i) above (7 marks)

(c) Suppose we have a data set (x_i, y_i) for $i = 1, 2, \cdots, n$. Consider two different models $y_i = \alpha + \beta x_i^2 + \epsilon_i$ and $y_i = \alpha + \beta x_i^2 + \gamma \exp^{x_i} + \epsilon_i$. Compare the residual sum of squares of the two models. Explain (3 marks)

4. The data given in the following table are the numbers of deaths from AIDS in Kenya for 12 consecutive quarters starting from the second quarter of 1998

Quarter (i)	1	2	3	4	5	6	7	8	9	10	11	12
Number of deaths (n_i)	1	2	3	1	4	9	18	23	31	20	25	37

(a) (i) Draw a scatterplot of the data

(ii) Comment on the nature of the relationship between the number of deaths and the quarter in this early phase of the epidemic (4 marks)

(b) A statistician has suggested that a model of the form

$$E[N_i] = \gamma i^2$$

might be appropriate for these data, where γ is a parameter to be estimated form the above data. She has proposed two methods for estimating γ given in (i) and (ii) below

(i) Show that the least squares estimate of γ, is obtained by minimizing $q = \sum_{i=1}^{12} (n_i - \gamma i^2)^2$ is given by

$$\hat{\gamma} = \frac{\sum_{i=1}^{12} i^2 n_i}{\sum_{i=1}^{12} i^4}$$

(ii) Show that an alternative (weighted) least squares estimate of γ obtained by minimizing $q^* = \sum_{i=1}^{12} \frac{(n_i - \gamma i^2)^2}{i^2}$ is given by

$$\tilde{\gamma} = \frac{\sum_{i=1}^{12} n_i}{\sum_{i=1}^{12} i^2}$$

CONTINUED

(c) Noting that $\sum_{i=1}^{12} i^4 = 60,710$ and $\sum_{i=1}^{12} i^2 = 650$, calculate $\hat{\gamma}$ and $\tilde{\gamma}$ for the above data (8 marks)

(d) To asses whether the single parameter model which was used in part (b) is appropriate for the data, a two parameter model is now considered. The model is of the form $E(N_i) = \gamma i^{\theta}$ for $i = 1, 2, \cdots, 12$

(i) To estimate the parameters γ and θ, a simple linear regression of the form $E(Y_i) = \alpha + \beta x_i$ is used , where $x_i = \log(i)$ and $Y_i = \log(N_i)$ for $i = 1, 2, \cdots, 12$. Relate the parameters γ and θ to the regression parameters α and β

(ii) The least squares estimates of α and β are -0.6112 and 1.6008 with standard errors 0.4556 and 0.2525 respectively. Using the value for the estimate of β, conduct a formal statistical test to asses whether the form of the model suggested in (b) is adequate (8 marks)