Question Paper

MANIPAL UNIVERSITY

SCHOOL OF INFORMATION SCIENCES
SECOND SEMESTER Master of Engineering - ME (Big Data and Data Analytics)
DEGREE EXAMINATION - NOVEMBER 2017
DATE : Saturday, November 18, 2017
TIME : 10:00AM - 1:00PM

Multiple Linear Regression and Logistic Regression [BDA 606]

Marks: 100 **Duration: 180 mins.**

Answer all the questions.

1) a) State the simple linear regression equation. What are the assumptions? (10)

b) The following partial ANOVA refers to the simple regression model where the response variable is the 'weight' and the explanatory variable is 'height' and the number of observations is 30. Complete the ANOVA table.

ANOVA Table

Source	d.f	S.S.	M.S.S.	F-Ratio
Regression	?	21.81	?	3.27
Residual	?	?	?	
Total	?	?		

(3+7)

2) a) Establish the relation between the regression coefficient $\beta 1$ and the correlation coefficient. (10)

b) If the correlation coefficient between 'mileage' and 'horsepower' is -0.71, test for the significance of correlation coefficient. (n=20, critical value= 2.1) (5+5)

3) a) With reference to multiple linear regression, derive the least square estimators of the regression coefficients. (10)

b) Write the variance-covariance matrix of the regression coefficients. How do you estimate σ^2?

(6+4)

4) a) With reference to multiple linear regression, distinguish (10)

between exploratory data analysis and confirmatory data analysis.
b) Describe backward elimination procedure and forward selection procedure in multiple linear regression. In handling big data, which one of these procedures is preferable? Justify your answer.

(2+8)

5) a) With the help of an example, explain why we do not interpret the intercept term in linear regression. (10)
b) Define the following: (i) residual (ii) standardized (normalized) residual (iii) studentized residual. Which one of these is commonly used in residual analysis? (5+5)

6) a) Explain any two causes of Multicollinearity. (10)
b) Explain the role of correlation matrix in detecting Multicollinearity.
c) Explain what you understand by ridge regression. Write the expression for the ridge estimator.

(3+3+4)

7) a) With the help of examples, distinguish between linear model and generalized linear model. (10)
b) What is link function? When the response variable is binary, describe the various link functions that are commonly used. What is the justification for these link functions? (5+5)

8) Consider the data given below. (10)

Estimated p_i	0.10	0.15	0.20	0.35	0.40	0.45	0.57	0.61	0.68	0.79	0.82	0.89
Observed Y_i	0	0	1	1	0	0	1	1	0	1	1	1

a) Compute the values of sensitivity and specificity in the following cases.

(i) When the estimated value of Y_i is taken as 1 if the estimated value of $p_i \geq 0.5$

(ii) When the estimated value of Y_i is taken as 1 if the estimated value of $p_i \geq 0.6$.

b) Define ROC curve.
(8+2)

9) a) What is the salient resemblance in the output of multiple linear regression and logistic regression? (10)
b) With reference to one explanatory variable which is dichotomous in nature, explain the Pearson Chi-square goodness of fit test for testing the overall fit of the logistic regression. (5+5)

10) a) Describe Hosmer-Lemeshow test for testing the overall fit in the logistic regression. (10)
b) How do you test the regression coefficients in logistic regression? (8+2)

-----End-----