

Indian Statistical Institute, Bangalore

M.S. (QMS) First Year

Second Semester - Multivariate Data Analysis

Final Exam Date: 18 April, 2016 Time: 3 hours Maximum Marks: 50

Answer as many questions as you can. Maximum you can score is 50 marks

1. Explain and interpret the following graphical tools and its usefulness in predictive modelling? [10]

- a. Scatter plot
- b. Box plot
- c. Conditional density plot
- d. Mosaic plot

2. Explain multicollinearity? How it is measured? Describe the following techniques and how they can be useful when multi-co-linearity exists? [15]

- a. Stepwise regression
- b. Principal component regression
- c. Ridge regression

3. Write short note on the following: [12]

- a. LOOCV (leave one out cross validation)
- b. k - fold cross validation
- c. Component + residual plots
- d. Random forest

4. Describe step by step details of carrying out linear discriminate analysis. [5]

5. The quality professionals in an IT company suspects that the sprint velocity of the agile projects depends on the test coverage and code reuse. The data collected on sprint velocity, test coverage and code reuse are given below. Develop a model for sprint velocity using k nearest neighbour method and predict the sprint velocity for a test coverage of 60 and code reuse of 25. Take $k = 3$? [18]

Sprint velocity	Test Coverage	Reuse	Sprint velocity	Test Coverage	Reuse
0.46	100	30	1.60	70	70
0.73	70	50	0.57	50	50
0.50	100	50	0.26	70	15
0.35	100	25	0.54	50	50
0.19	70	25	0.18	100	10
0.23	70	10	0.47	100	20
0.16	70	20	1.26	100	70
0.73	50	50	0.32	100	10