Indian Statistical Institute, Bangalore
MS (QMS) First Year
Second Semester - Multivariate Data Analysis

Final Exam | Date: May 02, 2018
Maximum marks: 50 | Duration: 3 hours

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

1.

a. Explain multicollinearity? How can it be detected? Briefly, explain three methods to tackle multicollinearity?

b. In predictive modeling, explain the various methods to check the following:

i. Model significance

ii. Model accuracy

iii. Model adequacy

iv. Model generalisability

[10]

2.

a. Describe the step by step procedure for developing a model using regression splines?

b. A model developed to predict the productivity (y) with vintage (x) as the explanatory variable using regression spline is

$$y = 55.63 + 26.89x - 26.52h_1(x) - 8.89h_2(x)$$

Where $h_1(x) = x - 1$ and $h_{12}(x) = x - 2$

Validate the model using the following test data? Compute R^2, mean square error and root mean square error? How good is the model?

x	0.35	0.72	2.23	1.81	1.39	2.45	2.92
y	64	76	80	81	84	79	74

[15]

3.

a. A model needs to be developed to estimate link testing defect density (y) using test environment set up time (x_1), data quality (x_2), team skill (x_3), test cases (x_4) and test coverage (x_5) as explanatory variables. Since the x's are highly correlated, the model is developed using principal component regression. The component importance and loadings are given below. Identify the optimum number of components to be used in the model?

Component Importance	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.90644	0.85579	0.50056	0.36247	0.25697
Proportion of Variance	0.75485	0.15211	0.05204	0.02729	0.01371
Cumulative Proportion	0.75485	0.90696	0.959	0.98629	1

Loadings	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
x_1	-0.47	0.114	0.73		-0.482
x_2	-0.44	0.527		-0.566	0.454
x_3	-0.451	0.389	-0.505	0.588	-0.21
x_4	-0.425	-0.558	-0.417	-0.447	-0.367
x_5	-0.449	-0.497	0.187	0.366	0.618

c. The coefficient of the model developed to predict the y using principal components as explanatory variables is given in the table below. Identify the model with the optimum number of principal components as explanatory variables? (Since the components are orthogonal, the coefficients of intercept and explanatory variables will not change for any combination of explanatory variables).

	Coefficients
(Intercept)	1.14143
PC1	-0.20076
PC2	0.06296
PC3	-0.0272
PC4	0.02203
PC5	0.16249

d. Use the model to predict the response *y* for the following combination of *x*'s (the standardized values of *x*'s are given below)

Project id	x1	x2	x3	x4	x5
1	0.19	-0.27	-1.03	-0.81	-0.44
2	-1.13	-1.54	-1.03	-0.97	-0.68
3	1.51	1	1.19	1.58	2.59
4	1.51	1	1.19	-0.27	0.33
5	1.51	1	1.19	1.58	2.59
6	0.19	-0.27	0.08	-0.27	-0.44
7	-1.13	-1.54	-1.03	-0.97	-0.68
8	-1.13	-1.54	-1.03	-0.74	-0.67
9	-1.13	-0.27	-1.03	-0.67	-0.66
10	0.19	1	1.19	0.94	-0.07

[15]

4.

a. Describe logistic regression and discriminant analysis methods? What are the similarities and dissimilarities between logistic regression and discriminant analysis?

b. The models are developed to classify a binary response variable y as 0 or 1. The models are developed using logistic regression, classification tree and discriminant analysis. The models are validated in 15 cases. The actual *y* value and the probability of *y* = 1 using the aforementioned three methods are given below. Comment on the models and choose the best model? Give reasons?

		Predicted P(*y* = 1)		
SL No	y	Logistic Regression	Classification Tree	Discriminant Analysis
1	0	0.6270	0.4194	0.5916
2	0	0.5879	0.4194	0.5540
3	0	0.8452	0.5574	0.1880
4	1	0.5949	0.4194	0.4696
5	0	0.4330	0.5574	0.4258
6	1	0.5359	0.4194	0.5080
7	0	0.5096	0.4194	0.4828
8	1	0.6751	0.4194	0.6354
9	0	0.0417	0.0139	0.0533
10	1	0.6334	0.4362	0.5994
11	0	0.6294	0.4194	0.5915
12	0	0.5017	0.4194	0.4776
13	0	0.5405	0.4194	0.5119
14	1	0.6046	0.8800	0.5669
15	1	0.6216	0.4362	0.5906

[15]