Computer Lab - Practical Question Bank
FACULTY OF COMMERCE,
OSMANIA UNIVERSITY

B.Com (Business Analytics) Semester IV

Forecasting and Predictive Analytics

Time: 60 Minutes

Tools used: Microsoft Excel, R Programming

Record : 10

Skill Test : 15

Viva - Voce : 10

Total Marks : **35**

1. The data given below explain the attitude towards the city and the duration of residence.

Respondents	1	2	3	4	5	6	7	8	9	10	`11	12
Attitude towards the city	6	9	8	3	10	4	5	2	11	9	10	2
Duration of Residence	10	12	12	4	12	6	8	2	18	9	17	2

   a. Construct a scatter plot.
   b. Draw a the regression line.
   c. Display the equation of the regression line.
   d. Display R^2 for the equation.
   e. Estimate the coefficient of determination.

2. Number of profiled customers (in millions) and Annual Sales in ($Millions) for a Sample of 14 Sunflowers Apparel Stores are given below.

Store	Profiled Customers(millions)	Annual Sales ($ Millions)
1	3.7	5.7
2	3.6	5.9
3	2.8	6.7
4	5.6	9.5
5	3.3	5.4
6	2.2	3.5
7	3.3	6.2
8	3.1	4.7
9	3.2	6.1
10	3.5	4.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3.0	4.1

a. Assuming a linear relationship between the Annual sales and Profiled customers, determine regression equation.
b. Determine the regression coefficients b_0 and b_1
c. Interpret the meaning of the slope, b1 in this problem.
d. Interpret R Square with respect to the data given.
e. Estimate the coefficient of determination.

3. In finance, it is of interest to look at the relationship between Y, a stock's average return and X, the overall market return. The slope coefficient computed by linear regression is called stock's beta by investment analysis. A beta greater than 1 indicates that the stock is relatively sensitive to changes in the market; a beta less than 1 indicates that the stock is relatively insensitive.

Y	10	12	8	15	9	11	8	10	13	11
X	11	15	3	18	10	12	6	7	18	13

For the above data,

a. Compute the beta
b. Compute standard error
c. What is the coefficient of determinant?
d. Explain the significance of F.
e. Explain the significance of t-statistic
f. Test the model to see whether it is significantly less than 1. Use $\alpha = 0.05$.

4. Campus stores has been selling the *Believe It or Not: Wonders of Statistics Study Guide* for 12 semesters and would like to estimate the relationship between sales and number of sections of elementary statistics taught in each semester. The following data have been collected.

Sales	33	38	24	61	52	45	65	82	29	63	50	79
No. Of Sections	3	7	6	6	10	12	12	13	12	13	14	15

a. Develop the estimating equation that best fits the data both from scatter plot and from data analysis tool.
b. Calculate the sample coefficient of determination.
c. Draw Residual plots
d. Draw normal probability plot and explain the output.

5. 1. Realtors are often interested in seeing how the appraised value of a home varies

according to the size of the home. Some data on area (In thousands of square feet) and appraised value (in thousands of dollars) for a sample of 11 homes if given below.

Area	1.1	1.5	1.6	1.6	1.4	1.3	1.1	1.7	1.9	1.5	1.3
Value	75	95	110	102	95	87	82	115	122	98	90

a. Estimate the Least Squares Regression to predict appraised value from area.
b. Predict the value for the size of the home whose area is 1.8.
c. What is the standard error of estimate.
d. Generally, realtors feel that a home's value goes up by $50,000 (50 thousands of dollars) for every additional 1,000 square feet in area. For this sample, does this relationship seem to hold? Use $\alpha = 0.05$
e. Predict the value for the size of the home whose area is 2.3

6. Given the following set of data

Y	25	30	11	22	27	19
X_1	3.5	6.7	1.5	0.3	4.6	2.0
X_2	5.0	4.2	8.5	1.4	3.6	1.3

a. Calculate the multiple regression equation.
b. Give the values of intercept bo, b1 and b2 values.
c. Explain the significance of b0, b1, and b2 values
d. Predict Y when $X_1 = 3.0$ $X_2 = 2.7$ from the regression equation obtained.

7. Find the multiple linear regression of Y on X1 and X2

Y	11	17	26	28	31	35	41	49	63	69
X_1	2	4	6	5	8	7	10	11	13	14
X_2	2	3	4	5	6	7	9	10	11	13

a. Find the Multiple regression equation.
b. What is R^2 for this regression? What is your inference?

c. What is adjusted R^2 for this equation? What is your inference?

d. Explain R^2 and adjusted R^2 with reference to the variables and equation.

e. Why is adjusted R^2different from R^2?

8. The following data explains the attitude toward the City of Residence.

Sl.No.	1	2	3	4	5	6	7	8	9	10	11	12
Attitude toward the city	6	9	8	3	10	4	5	2	11	9	10	2
Duration of Residence	10	12	12	4	12	6	8	2	18	9	17	2
Importance attached to weather	3	11	4	1	11	1	7	4	8	10	8	5

a. Find the regression equation. Mention the regression coefficients and their significance.
b. Test the significance of the overall regression equation.
c. What is the standard error of estimate?
d. Explain the attitude toward the city in terms of durations of residence and importance attached to weather.
e. Compare the regression coefficients and explain which attribute – Duration of residence, Importance attached to weather has more influence on Attitude toward the city.

9. Predict the annual demand for widgets (DEMAND) using the following independent vairables. Price = Price of Widgets (in $)
Income = Consumer income (in $) Sub = Price of a substitute commodity (in $) (A substitute commodity is one that can be substituted for another commodity. For example, margarine is a substitute commodity for butter)
The below data is collected form 1982 to 1996.

Year	Demand	Price ($)	Income ($)	Sub ($)
1982	40	9	400	10
1983	45	8	500	14
1984	50	9	600	12
1985	55	8	700	13
1986	60	7	800	11
1987	70	6	900	15
1988	65	6	100	16

1989	65	8	1100	17
1990	75	5	1200	22
1991	75	5	1300	19
1992	80	5	1400	20
1993	100	3	1500	23
1994	90	4	1600	18
1995	95	3	1700	24
1996	85	4	1800	21

a. Using MS-Excel determine the best fitting regression equation .
b. Are the signs (+) and (-) of the regression coefficients of the independent variables as one would expect? Explain briefly.
c. State and interpret the coefficient of multiple determination for this problem
d. State and interpret the standard error of estimate.
e. Using the equation, what would you predict for DEMAND if the price of widgets was $6 consumer income was $1200, and the price of the substitute commodity is $17.

10. Consider a study that examined the business problem facing a concrete supplier of how adding fly-ash affects the strength of concrete. (fly-ash is an inexpensive industrial waste by-product that can be used as a substitute for Portland cement, a more expensive ingredient of concrete.

Batches of concrete were prepared in which the percentage of fly-ash ranges from 0% to 60%. Data were collected from a sample of 18 batches and organized.

Fly-ash %	0	0	0	20	20	20	30	30	30
Strength	4,779	4,706	4,350	5,189	5,140	4,976	5,110	5,685	5,618
Fly-ash%	40	40	40	50	50	50	60	60	60
strength	5,995	5,628	5,897	5,746	5,719	5,782	5,895	5,030	4,648

a. Draw the scatter plot.
b. Draw the regression line.
c. Display the regression equation
d. Display R^2 and interpret the result.
e. Describe the strength of association between the variables.

11. Consider a study that examined the business problem facing a concrete supplier of how adding fly-ash affects the strength of concrete. (fly-ash is an inexpensive industrial waste by-product that can be used as a substitute for Portland cement, a more expensive ingredient of concrete.)

Batches of concrete were prepared in which the percentage of fly-ash ranges from 0% to 60%. Data were collected from a sample of 18 batches and organized.

Fly-ash %	0	0	0	20	20	20	30	30	30
Strength	4,779	4,706	4,350	5,189	5,140	4,976	5,110	5,685	5,618
Fly-ash%	40	40	40	50	50	50	60	60	60
strength	5,995	5,628	5,897	5,746	5,719	5,782	5,895	5,030	4,648

a. Find the quadratic regression model for the concrete strength data.
b. Explain the linear and quadratic effect of fly-ash on strength of the concrete.
c. Explain the test for significance of the quadratic model using f-Stat.
d. Test the significant difference between the quadratic model and linear model for testing the quadratic effect.
e. If you select the 0.05 level significance, then explain the critical values for the t-distribution.

12. A real estate developer studying the business problem of estimating the consumption of heating oil by single family houses has decided to examine the effect of atmospheric temperature and the amount of attic insulation on heating oil consumption. Data are collected from a random sample of 15 single family houses.

Heating oil	Temperature	Insulation
275.3	40	3
363.8	27	3
164.3	40	10
40.8	73	6
94.3	64	6
230.9	34	6
366.7	9	6
300.6	8	10
237.8	23	10
121.4	63	3
31.4	65	10
203.5	41	6
441.1	21	3
323.0	38	3
52.5	58	10

a. Find the Multiple regression equation using two independent variables.
b. For the excel regression model, interpret the results for predicting monthly

consumption of heating oil.

c. Draw the residual plot for attic insulation. Do you find some evidence of quadratic effect.

d. If, then establish the quadratic regression model, by adding a quadratic term for attic insulation to the multiple regression model.

e. Test the significance of quadratic effect.

13. The table gives the data of 30 respondents, 15 of whom are brand loyal (Indicated by 1) and 15 of whom are not (indicated by 0).

Loyalty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Brand	4	6	5	7	6	3	5	5	7	7	6	5	7	5	7
Loyalty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brand	3	4	2	5	4	3	3	3	4	6	3	4	3	5	1

a. Draw a scatter plot.

b. Find the regression equation.

c. Interpret R^2

d. Interpret Adjusted R^2.

e. Mention the type of regression, and why it so called.

14. The table gives the data of 30 respondents, 15 of whom are brand loyal (Indicated by 1) and 15 of whom are not (indicated by 0). Attitude toward the Brand, Attitude toward the Product Category, and Attitude toward the Shopping experience are also measured. The objective is to estimate the probability of a consumer being brand loyal as a function of attitude toward the brand, the product category and shopping.

Loyalty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Brand	4	6	5	7	6	3	5	5	7	7	6	5	7	5	7
Product	3	4	2	5	3	4	5	4	5	6	7	6	3	1	5
shopping	5	4	4	5	4	5	5	2	4	4	2	4	3	4	5
Loyalty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brand	3	4	2	5	4	3	3	3	4	6	3	4	3	5	1
Product	1	6	5	2	1	3	4	6	4	3	6	3	5	5	3
shopping	3	2	2	4	3	4	5	3	2	6	3	2	2	3	2

a. Plot the scatter plot.

b. Find the regression equation.

c. What are the degrees of freedom.

d. Explain R^2. And its significance with each of the regression coefficients.
e. Of the three independent variables, which variable is more significant to predict brand loyalty.

15. One of the investments considered in The Principled Scenario is the entertainment industry. The table given below presents the yearly US and Canada movie attendance (in billions) from 2001 through 2014.

Year	2001	2002	2003	2004	2005	2006	2007
Attendance	1.24	1.14	1.34	1.39	1.45	1.33	1.52
Year	2008	2009	2010	2011	2012	2013	2014
Attendance	1.42	1.34	1.28	1.44	1.56	1.58	1.62

a. Present the Time series plot of this data.
b. Describe the data with the regression equation.
c. Draw the trend line to the data.
d. Explain R^2.
e. Make forecasts for the next 5 years.

16. One of the investments considered in The Principled Scenario is the entertainment industry. The table given below presents the yearly US and Canada movie attendance (in billions) from 2001 through 2014.

Year	2001	2002	2003	2004	2005	2006	2007
Attendance	1.24	1.14	1.34	1.39	1.45	1.33	1.52
Year	2008	2009	2010	2011	2012	2013	2014
Attendance	1.42	1.34	1.28	1.44	1.56	1.58	1.62

For the data given above
a. Obtain a three year moving average.
b. Draw the trend line graph for the 3 year moving average.
c. Obtain the regression line and R^2.
d. Forecast the attendance in the year 2016.

17. For the data given below

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Sales (in billions)	16.6	15.2	13.4	15.5	15.8	16.3	14.5	16.8	19.8	20.5
Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Sales (in billions)	21	19.2	18.3	21.9	23.1	24.5	28.7	31.9	30	35.1

a. Obtain 5 year moving average
b. Draw the trend line graph for the 5 year moving average.
c. Obtain the regression line and normal probability plot
d. Obtain the line fit plot.
e. Forecast the Revenue in the year 2020.

18. According to The Coca-Cola Company's Website, revenues in 2014 were almost $46 billion. The data given below lists The Coca-Cola Company's gross revenues (in $billions) from 1998 to 204 are given.

Year	1998	1999	1200	2001	2002	2003	2004	2005	2006
Revenue (in $ Billions)	18.8	19.8	20.5	20.1	19.6	21.0	21.9	23.1	24.1
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Revenue (in $ Billions)	28.9	31.9	31.0	35.1	46.5	48.0	46.7	45.9	50.1

a. Obtain the linear trend model to forecast the revenues.
b. Obtain the quadratic trend model to forecast the revenues.
c. Using both the equations obtained, forecast the revenues for the year 2020.
d. Do you find any difference in the expected revenues? If explain the reasons and the significance.
e. Of both the models, which model can be considered?

19. For the data given above

a. Find the Exponential trend model.
b. Forecast the revenues for the year 2020.
c. Compare the three models Linear Trend Model, Quadratic Trend Model and Exponential Trend Model.
d. Select the best model for the given data.
e. Interpret the R^2.

20. Find the Euclidean Distance, Manhattan Distance and Cosine distance for the following data in Excel.

X	1	3	4	7	8	10	15	18	20	21
Y	3	5	8	11	14	19	25	27	30	35

21. Find the Euclidean Distance, Manhattan Distance and Cosine Distance for the following data in R.

X	1	3	4	7	8	10	15	18	20	21
Y	3	5	8	11	14	19	25	27	30	35

22. One of the investments considered in The Principled Scenario is the entertainment industry. The table given below presents the yearly US and Canada movie attendance (in billions) from 2001 through 2014.

Year	2001	2002	2003	2004	2005	2006	2007
Attendance	1.24	1.14	1.34	1.39	1.45	1.33	1.52
Year	2008	2009	2010	2011	2012	2013	2014
Attendance	1.42	1.34	1.28	1.44	1.56	1.58	1.62

Using R programming solve,

a. Present the Time series plot of this data.
b. Describe the data with the regression equation.
c. Draw the trend line to the data.
d. Explain R^2.
e. Make forecasts for the next 5 years.

23. The data given below explain the attitude towards the city and the duration of residence.

Respondents	1	2	3	4	5	6	7	8	9	10	`11	12
Attitude towards the city	6	9	8	3	10	4	5	2	11	9	10	2
Duration of Residence	10	12	12	4	12	6	8	2	18	9	17	2

Using R,

a. Construct a scatter plot.
b. Draw a the regression line.
c. Display the equation of the regression line.
d. Display R^2 for the equation.
e. Estimate the coefficient of determination.

24. With the built in data set mtcars in R, create a regression model predicting mpg based on weight and hp.

25. With the built in data set mtcars in R, create ggplot of multiple regression.

26. Check the non-linearity of mtcars data between mpg and weight in R.

27. Find the correlation coefficient of mpg and weight of mtcars data in R.

28. With mtcars in R, create multiple regression of weight, hp, cyl on mpg.

29. Make predictions of mtcars with weight (2.5, 3.0), hp (110,150), cyl(4,6) on mpg

30. According to The Coca-Cola Company's Website, revenues in 2014 were almost $46 billion. The data given below lists The Coca-Cola Company's gross revenues (in $billions) from 1998 to 204 are given.

Year	1998	1999	1200	2001	2002	2003	2004	2005	2006

Revenue (in $ Billions)	18.8	19.8	20.5	20.1	19.6	21.0	21.9	23.1	24.1
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Revenue (in $ Billions)	28.9	31.9	31.0	35.1	46.5	48.0	46.7	45.9	50.1

Using R, solve,

a. Obtain the linear trend model to forecast the revenues.
b. Obtain the quadratic trend model to forecast the revenues.
c. Using both the equations obtained, forecast the revenues for the year 2020.
d. Do you find any difference in the expected revenues? If explain the reasons and the significance.
e. Of both the models, which model can be considered?

31. Classify the following data in Excel.

Person	Age	Weight (kg)	Exercise Frequency (days/week)	Fitness Level (Target)
1	25	70	5	Fit
2	30	80	3	Fit
3	35	90	1	Unfit
4	40	95	2	Unfit
5	45	85	4	Fit
6	50	100	0	Unfit

32. Classify the above data in R.

33. Classify the given data in Excel.

Person	Age	Income	Buy (Target)
1	25	20000	No
2	30	25000	No
3	35	30000	Yes
4	40	35000	No
5	45	40000	Yes
6	50	45000	Yes
7	55	50000	Yes
8	60	55000	Yes

34. Classify the above data set in R with KNN means.

35. Classify the following data based on the Naïve Bayes Classification in Excel.

Email ID	Contains 'Free'	Contains 'Win'	Contains 'Money'	Spam? (Yes/No)
1	Yes	Yes	Yes	Yes
2	No	Yes	Yes	Yes
3	Yes	No	No	No
4	No	No	Yes	No

36. Classify the above data based on Naïve Bayes Classification in R.

37. Classify the data in Excel.

Email ID	Contains 'Free'	Contains 'Win'	Contains 'Money'	Spam? (Yes/No)
1	Yes	Yes	Yes	Yes
2	No	Yes	Yes	Yes
3	Yes	Yes	No	No
4	Yes	No	Yes	No

38. Classify the above data in R.

39. Predict whether the person has Flu or no Flu based on the symptoms from the given data set in Excel.

Dataset:

Person ID	Fever	Cough	Fatigue	Flu?
1	Yes	Yes	Yes	Yes
2	Yes	Yes	No	Yes
3	Yes	No	Yes	No
4	No	Yes	Yes	No

40. Perform Market Basket Analysis in Excel.

Transaction ID	Transaction Details
1	[Bread, Milk]
2	[Bread, Milk,Eggs,Diapers, Beer]
3	[Bread,Milk, Beer, Diapers]
4	[Diapers, Beer]
5	[Milk, Bread, Diapers, Eggs]
6	[Milk, Bread, Diapers, Beer]

41. Perform Market Basket analysis or Association Mining Rule for the following data in R.

Transaction ID	Transaction Details
1	[Bread, Milk]
2	[Bread, Milk,Eggs,Diapers, Beer]
3	[Bread,Milk, Beer, Diapers]
4	[Diapers, Beer]
5	[Milk, Bread, Diapers, Eggs]
6	[Milk, Bread, Diapers, Beer]

42. Form two clusters for the given 9 elements in Excel.

2	3	4	10	11	12	20	25	30

a. How many combinations of clusters are possible.

b. Which cluster is more suitable and Why?

c. How many elements are in each of the cluster which is suitable for modelling.

43. Form two clusters for the given 9 elements in R.

2	3	4	10	11	12	20	25	30

a. How many combinations of clusters are possible.

b. Which cluster is more suitable and Why?

c. How many elements are in each of the cluster which is suitable for modelling.

44. Perform K means clustering in Excel for the following data.

Age	25	30	35	40	45	50	55	60
Income	20000	25000	30000	35000	40000	45000	50000	55000

45. Perform K means clustering in R for the following data.

Age	25	30	35	40	45	50	55	60
Income	20000	25000	30000	35000	40000	45000	50000	55000

46. Ascent Construction builds four different types of residential spaces: Flats, Condominiums, Single Story houses, and double story houses. Each type requires basic development in five different units.

Particulars	Flat	Condo	Single Story	Double story	Days Available
Plumbing	3	5	8	15	69
Electrical	8	10	15	24	90
Partition	34	40	60	120	520
Painting	14	20	25	39	190
Installation	3	14	29	42	70
Cost (in $)	160	380	550	760	

a. How many of each type of residential space should be developed to minimize cost?

b. If marketing requires at least 20 units of each to be developed, what is the optimal development plan and cost?

47. A company produces two products X and Y. Each product requires a certain amount of resources (machine hours and labour hours) and each product

contributes a profit. The factory has limited resources available. Maximize the total profit if the profit per unit of X is $40 and and Y is $30. The resource requirements per unit are

Product	Machine Hours Required	Labor Hours Required
Product X	2	3
Product Y	4	2

And the machine hours available is 60 hours and labour hours 40 hours

48. Using Solver, achieve the target return of 9% for the investment options

Option	Expected Return (%)	Risk (Volatility, %)
Stocks	12	15
Bonds	5	3
Mutual Funds	8	7

And total investment is Rs. 100000 and the stock should not exceed Rs. 40,000. The total risk (weighted by allocation) must not exceed 8%.

49. A company wants to allocate its advertising budget across three platforms (TV, Online, and Print) to generate a specific target revenue while adhering to budget constraints.

The advertising platforms are

Platform	Revenue Generated per $1,000	Cost ($)
TV	15,000	20,000
Online	10,000	10,000
Print	5,000	5,000

a. Achieve a target revenue of $250,000.

With Constraints:

- Total advertising budget is $60,000.
- No more than $30,000 can be spent on TV.
- At least $10,000 must be spent on Online advertising. Decision Variables:
- Budget allocation for TV, Online, and Print.

50. X Company Ltd produces a product and to sell. But the demand for the product is uncertain. Simulate 1,000 trials to estimate the potential profit, assuming:

- Selling price per unit: $50
- Cost per unit: $30

- Demand follows a uniform distribution between 100 and 200 units.