7/11

2603N1120

Candidate's Seat No :_____

I.M.Sc. (DS) Sem.-5 Examination

CC-302

Regression Theory

March-2025

Time : 2-30 Hours] **[Max. Marks : 70**

Instructions: All questions are compulsory. Use of non-programmable scientific calculator is allowed.

Q.1 **(a)** The following data are the monthly salaries y and the grade point averages x for the students who obtained a bachelor's degree in business administration with a major in information systems. The estimated regression equation for these data is **(07)**

$\hat{y} = 1790.5 + 581.1x$

GPA	Monthly Salary ($)
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

1. Compute SST, SSR and SSE.
2. Compute the coefficient of determination r^2. Comment on the goodness of fit.
3. What is the value of the sample correlation coefficient?

(b) To test whether the mean time needed to mix a batch of material is the same for machines produced by three manufacturers, the Narayan Chemical Company Ltd. Obtained the following data on the time (in minutes) needed to mix the material. **(07)**

Manufacturer		
1	2	3
20	28	20
26	26	19
24	31	23
11	17	22

1. Use these data to test whether the population mean time for mixing a batch of material differs for the three manufacturers.
2. At the $\alpha = 0.05$ level of significance, use Fisher's LSD procedure to test for the equality of the means for manufacturers 1 and 3. What conclusion can you draw after carrying out this test? ($F_{10.63} = 0.0043, t_{0.025} = 2.262$)

OR

(a) The Scholastic Aptitude Test (SAT) contains three parts: Critical reading, Mathematics and Reading. Each part is scored on 800-point scale. A sample of SAT scores of six students follows. **(07)**

(P.T.O)

Students	Critical Reading	Mathematics	Reading
1	526	534	530
2	594	590	586
3	465	464	445
4	561	566	553
5	463	478	430
6	430	458	420

Using a 0.05 level of significance, do students perform differently on the three portions of SAT? ($p - value = 0.0231$)

(b) A research company has designed three different systems to clean up oil spills. The following table contains the results, measured by how much surface area (in square meters) is cleared in 1 hour. The data were found by testing each method in several trials. Are the three systems equally effective? Use the $\alpha = 0.05$. (Use $Ftab = 3.89$) **(07)**

System A	55	60	63	56	59	55
System B	57	53	64	49	62	
System C	66	52	61	57		

Q.2 **(a)** The following data are from a completely randomized design. **(07)**

	Treatment		
	A	**B**	**C**
	162	142	126
	142	156	122
	165	124	138
	145	142	140
	148	136	150
	174	152	128
Sample mean	156	142	134
Sample variance	164.4	131.2	110.4

1. Compute the sum of squares between treatments.
2. Compute the mean square between treatments.
3. Compute the sum squares due to error.
4. Compute mean squares due to error.

Set up the ANOVA table for this problem and give the appropriate conclusion using the F-test (where for $\alpha = 0.05$, the p-value for given degree of freedom is 3.68)

(b) The following data were collected on the height (inches) and weight (pounds) of women swimmers. **(07)**

Height	68	64	62	65	66
Weight	132	108	102	115	128

1. Develop the estimated regression equation by computing the values of b_0 and b_1.
2. If a swimmer's height is 63 inches, what would you estimate her weight to be?
3. Compute MSE and Standard error of the estimate.

OR

(a) Given are five observations collected in a regression study on two variables, **(07)**

x_i	2	6	9	13	20
y_i	7	18	9	26	23

1. Develop the estimated regression equation for the given data.
2. Use the estimated regression equation to predict the value of y when $x = 6$.

(b) Given are following data for two variables, x and y. **(07)**

x_i	135	110	130	145	175	160	120
y_i	145	100	405	120	130	130	110

Compute the standardized residuals for these data. Do the data include any outliers?

Q.3 **(a)** A 10-year study conducted by the American Heart Association provided data on how blood pressure, and smoking relates. Assume that the following data are from a portion of that study. For the smoking variable, a dummy variable with 1 indicating a smoker and 0 indicating a nonsmoker is already defined **(07)**

Sr. No.	Blood Pressure	Smoker
1	152	0
2	163	0
3	155	0
4	177	1
5	196	0
6	189	1
7	155	1
8	120	0
9	135	1
10	98	0
11	152	0
12	173	1

Find equations to predict smokers and non-smokers.

(P.T.O)

N1120-4

(b) Consider the following data for a dependent variable y and two independent variables, x_1 and x_2. **(07)**

x_1	x_2	y
10	5	0
20	7	12
30	10	20
40	14	24
50	15	40
60	21	36
70	20	60

1. Develop an estimated regression equation relating y to x_1. Estimate y if $x_1 = 90$
2. Develop an estimated regression equation relating y to x_2. Estimate y if $x_2 = 25$
3. Find SSE, SST and SSR for both x_1 and x_2.
4. Compute R^2 and R_a^2 for both x_1 and x_2.

OR

(a) Data for two variables, x and y, follow: **(07)**

x_i	22	24	26	28	40
y_i	12	21	31	35	70

1. Develop the estimated regression equation for these data.
2. Compute the leverage values for these data. Do there appear to be any influential observations in this data? Explain.

(b) Consider the following data for a dependent variable y and two independent variables, x_1 and x_2. **(07)**

x_1	x_2	y
30	10	90
50	20	110
40	20	120
20	30	180
60	10	90
80	5	170
70	40	220

1. Develop an estimated regression equation relating y to x_1. Estimate y if $x_1 = 45$
2. Develop an estimated regression equation relating y to x_2. Estimate y if $x_2 = 15$
3. Find SSE, SST and SSR for both x_1 and x_2.

Q.4 **(a)** Consider the following time series data to forecast using three months moving or four months moving average and compute the following measures of forecast accuracy for given data. **(07)**

Week	1	2	3	4	5	6	7	8	9	10	11	12
Value	9.5	9.3	9.4	9.6	9.8	9.7	9.8	10.5	9.9	9.7	9.6	9.6

1. Mean Absolute error
2. Mean Squared error
3. Mean Absolute Percentage error
4. Forecast for 13th Month

(b) The following time series shows the sales of a particular product over the past 12 months. **(07)**

Month	1	2	3	4	5	6	7	8	9	10	11	12
Sales	17	21	19	23	18	16	20	18	22	20	15	22

1. Construct a time series plot. What type of pattern exists in the data?
2. Use $\alpha = 0.2$ to compute the exponential smoothing forecasts for the time series.

OR

(a) Collected the following information on the number of tips he has collected from parking cars the last seven nights. **(07)**

Day	Tips
1	2
2	4
3	5
4	9
5	12
6	10
7	15

1. Compute the 2-day moving averages for the time series.
2. Compute the mean square error for the forecasts.

(b) The quarterly sales data (number of copies sold) for a college textbook over the past three years follow. **(07)**

Quarter	1	2	3	4
Year1	1690	940	2625	96
Year2	1800	900	2900	2360
Year3	1850	1100	2930	2615

1. Construct a time series plot. What type of pattern exists in the data?
2. Show the four-quarter and centered moving average values for this time series.

(P.T.O)

Q.5 Attempt any **SEVEN** out of **TWELVE** (Each carries **TWO** marks): **(14)**

(1) In a regression analysis, the regression equation is given by $y = 12 - 6x$. If $SSE = 510$ and $SST = 1000$, then what will be the coefficient of correlation?

(2) Write down the formula of Holt's Exponential moving average.

(3) The required condition for using an ANOVA procedure on data from several populations is that the

a. The selected samples are dependent on each other.
b. Sampled populations are all uniform
c. Sampled populations have equal variances
d. Sampled populations have equal means

(4) Which of the following is the cyclic behavior of time series?

a. Level
b. Trend
c. Seasonality
d. Noise

(5) If $r = 1$, the angle between two regression lines is _____

(6) A regression analysis is inappropriate when _____

a. you have two variables that are measured on an interval or ratio scale.
b. you want to make predictions for one variable based on information about another variable.
c. the pattern of data points forms a reasonably straight line.
d. there is heteroscedasticity in the scatter plot.

(7) When an analysis of variance is performed on samples drawn from K populations, the mean square between treatments (MSTR) is _____

(8) For the ANOVA table

Source of variations	Sum of squares	Degree of freedom
Between treatment	45	3
Error	32	16
Total	99	19

The F-statistics is _____.

(9) Write an equation to predict dependent variable for logistic regression.

(10) If $t_{\frac{\alpha}{2}} = 2.305$, $s_{pred} = 14.69$ and $\hat{y}^* = 110$ then find the value for prediction interval.

(11) Regression analysis was applied between demand for a product (Y) and the price of the product (X), and the following estimated regression equation was obtained.

$$\hat{Y} = 120 - 10X$$

Based on the above estimated regression equation, if price is increased by 2 units, then demand is expected to

a. increase by 120 units
b. increase by 100 units
c. increase by 20 units
d. decease by 20 units

(12) If the demand is 100 during October 2016, 200 in November 2016, 300 in December 2016, 400 in January 2017. What is the 5-month simple moving average for February 2017?

_____ **** _____ **** _____