Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

UPES

End Semester Examination, December 2023

Course: Predictive Modelling **Semester: I**
Program: M. Tech. (CSE) **Time : 03 hours**
Course Code: CSDA7002P **Max. Marks: 100**

Instructions: Attempt all the questions.

SECTION A
(5Qx4M=20Marks)

S. No.		**Marks**	**CO**
Q1	In a multiple regression analysis involving 12 independent variables and 142 observations, SSR = 884 and SSE = 178. Calculate the coefficient of determination. How much variation in the dependent variable is explained by the independent variables in the model?	**4**	**CO1**
Q2	The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 28 male college students by using least squares regression. The following regression equation was obtained from this study: $y = -0.0234 + 0.0360x$ What does the above equation imply. How does the coefficient 0.0360 in the regression equation reflect the impact of a one-unit change in the number of beers consumed on blood alcohol content?	**4**	**CO2**
Q3	In the context of linear regression, suppose you have fitted a model to predict house prices based on various features such as square footage, number of bedrooms, and location. Upon examining the residuals, you observe the presence of outliers. Discuss the potential implications of these outliers on your regression analysis. Specifically, address how outliers can affect the model's predictive accuracy, the estimates of regression coefficients, and the assumptions underlying linear regression. Furthermore, propose and explain at least two strategies you would consider employing to manage or address the impact of outliers in this regression analysis, taking into account the potential consequences of outlier removal on the overall validity and reliability of the model.	**4**	**CO3**
Q4	Explain the difference between R-squared and adjusted R-squared in regression models.	**4**	**CO4**
Q5	Consider the estimated regression equation: $\hat{y} = 3536 + 1183X_1 - 1208X_2$. Suppose the model is changed to reflect the deletion of X_2 and the resulting estimated simple linear equation becomes $\hat{y} = -10663 + 1386X_1$. a) How should we interpret the meaning of the coefficient on X_1 in the estimated simple linear regression equation $\hat{y} = -10663 + 1386X_1$?	**4**	**CO2**

	b) How should we interpret the meaning of the coefficient on X1 in the estimated multiple regression equation $\hat{y}= 3536 + 1183X_1 – 1208X_2$?		
	SECTION B **(4Qx10M= 40 Marks)**		
Q6	Derive the slope and intercept of linear regression using Ordinary Least Square fitting.	**10**	**CO1**
Q7	Discuss the impact of multicollinearity on regression analysis. In a multiple regression model, you observe high multicollinearity among certain predictors. Discuss two methods or techniques that can be employed to handle or reduce the impact of multicollinearity in regression analysis.	**10**	**CO2**
Q8	Describe the concepts of autocorrelation, normality, and homoscedasticity in the context of regression analysis. How do violations of these assumptions impact the validity of regression models? Given a regression model Y=2X−1, and the data points (1, 1), (2, 3), (3, 5), (4, 8), calculate the residuals. Assess the presence of any patterns or deviations from the assumptions of autocorrelation in the residuals using Durbin Watson statistic.	**10**	**CO4**
Q9	In a regression analysis with 60 observations and 10 independent variables, if SSR = 500 and SSE = 300, perform the F-test to determine whether all regression parameters are simultaneously zero at the 0.05 significance level. If you assume the F-critical value for a 0.05 significance level to be 2.15, calculate the F-statistic and interpret whether the null hypothesis should be rejected. **OR** Prove that slope of linear regression model is equal to $Cov\,(X,Y)/V(X)$. Where X is independent variable and Y is dependent varaible.	**10**	**CO4**
	SECTION-C **(2Qx20M=40 Marks)**		
Q10	Consider a dataset involving housing prices based on the size of the house, number of bedrooms, and the distance to the city center. The dataset includes: House Size (sq. ft) — Bedrooms — House Prices (in Rs.) 1500 — 3 — 350,000 1800 — 4 — 420,000 2000 — 3 — 380,000 1400 — 2 — 300,000 2500 — 5 — 500,000 Perform multiple linear regression by hand using matrix methods to predict house prices based on the house size, number of bedrooms, and distance to the city center.	**20**	**CO3**
Q11	Analyze a dataset that includes study hours and final exam scores: **Study Hours** — **Final Exam Scores** 4 — 92 6 — 88 5 — 75	**20**	**CO2**

3	80
7	85

a) Perform a correlation analysis to determine the relationship strength between study hours and final exam scores. Interpret the obtained correlation coefficient.

b) Apply simple linear regression using matrix methods to predict final exam scores based on study hours.

c) Calculate the residuals in terms of Mean Squared Error (MSE).

OR

In a dataset analyzing the relationship between temperature and ice cream sales, there's a data point with an unusually high temperature (an outlier). Analyze the impact of this outlier on the simple linear regression model built for predicting ice cream sales. Discuss the effect on the regression line, R-squared value, and the model's overall predictive ability.

Temperature (°C)	**Ice Cream Sales**
15	20
20	25
25	30
30	35
35	40
50	60

This dataset represents temperatures in Celsius and the corresponding ice cream sales in some hypothetical context. All values follow a linear relationship except for the last entry, which represents the outlier.