BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: IMSC **SEMESTER : VII**
BRANCH: CQEDS **SESSION : MO/2025**

SUBJECT: ED409 REGRESSION TECHNIQUES

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. F-table will be supplemented during the examination.

			CO	BL
Q.1(a)	(i) For an MLR, estimate the model parameters (OLS estimators) in matrix form. [3] (ii) Prove that the OLS estimator $\hat{\beta}$ is an unbiased estimator of β if the model is correctly specified. [1] (iii) Prove that the $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ [2]	[6]	1	3
Q.1(b)	Discuss White test for heteroskedasticity. Explain the concept, hypotheses, and test statistic. Give all mathematical expressions.	[4]	1	3
Q.2(a)	A researcher is studying the factors associated with monthly household healthcare expenditure (Y). The following multiple linear regression models were fitted using a sample of n = 40 households: Full model: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i$ Reduced model: $Y_i = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \nu_i$ Following regression results are available:	[5]	2	3

Model	SSE (Sum of Squared Errors)	Number of predictors included
Full	8,200	4
Reduced	12,100	2

Using a 5% level of significance, test whether the additional predictors X_3 and X_4 jointly contribute significantly to the prediction of household healthcare expenditure.

Q.2(b)	Explain how the following measures are used for diagnosing the outliers and influential observations in MLR model: (i) Mahalanobis Distance (ii) Cook's Distance	[5]	2	3
Q.3(a)	Mention the functional form of binary logistic regression. Estimate the parameters in a binary logistic regression. Give all mathematical expressions.	[3]	3	3
Q.3(b)	A public health researcher uses binary logistic regression to predict whether a child is malnourished (1 = Yes, 0 = No) based on household socioeconomic indicators. The model was fitted using survey data, and the predicted classification for 12 children was generated at a probability cut-off of 0.50. The table below presents the actual nutritional status and the predicted class produced by the logistic regression model:	[7]	3	4

Child	Actual Malnutrition (Y)	Predicted Probability
1	1	0.83
2	0	0.32
3	1	0.61
4	1	0.47
5	0	0.22
6	1	0.73
7	0	0.55
8	0	0.18
9	1	0.64
10	0	0.41
11	1	0.36
12	1	0.58

PTO

(i) Construct the confusion matrix for the above classification. [1]
(ii) Calculate the following performance measures: [4]
(a) Sensitivity
(b) Specificity
(c) Precision
(d) F1 Score
(iii) Based on the above measures, comment on the quality of classification done by the logistic regression model and whether it is suitable for field-level screening of malnutrition cases. [2]

Q.4(a) A researcher wants to estimate the blood glucose level (Y) of a new patient based on Body Mass Index (BMI) using weighted K-Nearest Neighbour regression. The dataset of 5 patients is given below: , [5] 4 4

Patient	BMI (X)	Glucose Level (Y)
A	18	72
B	22	85
C	25	90
D	28	96
E	30	110

The new patient has BMI = 26. Using K = 3 and inverse distance weights: $w_i = \frac{1}{d_i}$, where $d_i = |X_0 - X_i|$, predict the glucose level using weighted KNN regression.

Q.4(b) For the data given in Q4(a), predict the glucose level for a new patient with BMI=26 using Kernel regression. Use Nadaraya-Watson estimator of weights with the Gaussian kernel and bandwidth h=1.0. [5] 4 4

Q.5 A data scientist is building different regression models for three different real-life cases: [10] 5 4
Case1: A model is to be developed to predict crop yield using 15 climatic and soil variables that are highly correlated with one another. The goal is not to remove variables, but to achieve stable coefficient estimates and reduce multicollinearity, even if all predictors remain in the model.

Case2: A model is to be developed to predict student academic performance using a mixture of demographic factors, socioeconomic factors, attendance, online activity, and parental involvement. The predictors are both very numerous and strongly correlated, and the researcher wants a method that can handle multicollinearity while also performing variable selection, instead of keeping every predictor.

Case3: A model is to be prepared to predict house prices using over 100 property characteristics (e.g., number of rooms, floor area, flooring type, proximity to metro, balcony type, etc.). Many predictors have very weak or no contribution, and the researcher wants the model to automatically select only the most relevant variables, shrinking some coefficients exactly to zero.

Answer the following:
(i) Identify the type of regression used in case1, case2 and case3. Justify your choice in each case.
(ii) For each study, write the corresponding penalty term added to the OLS loss function and explain how it influences the regression coefficients.
(iii) Compare the three approaches in terms of coefficient shrinkage, variable selection ability, and their suitability when predictors are strongly correlated.