U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June / July 2025 Semester End Main Examinations

Programme: B.E. **Semester: V**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 24AM5PCSML **Max Marks: 100**

Course: STATISTICAL MODELING

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

<table>
<tr><th></th><th></th><th>UNIT - I</th><th><i>CO</i></th><th><i>PO</i></th><th>Marks</th></tr>
<tr><td>1</td><td>a)</td><td>By stating, the simple linear regression model and its assumptions, Obtain the least square estimates of the parameters.</td><td><i>CO1</i></td><td><i>PO1</i></td><td><b>08</b></td></tr>
<tr><td></td><td>b)</td><td>The relationship between expenditure ($) and income ($) is modelled using simple linear regression, with the following regression output, and its output is given below:
<table>
<tr><th>Coefficients:</th><th>Estimate</th><th>Std. Error</th></tr>
<tr><td>Intercept</td><td>31.42</td><td>0.0565</td></tr>
<tr><td>income ($)</td><td>0.7138</td><td>0.1209</td></tr>
</table>
Obtain the 95% confidence interval for intercept (β_0) and slope (β_1) parameter.</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td>c)</td><td>Define R² and discuss its limitations, especially in the context of comparing models with different numbers of predictors. What advantages does adjust. R² offer in balancing model efficiency and complexity?</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>a)</td><td>As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. Engineers are interested in the following variables. y - moisture control of compressed pellets (%) and x-machine filtration rate (kg-DS/m/hr). Engineers collect observations of (x, y) with a random sample of size 10 sewage specimens.
<table>
<tr><td><b>x</b></td><td>125.3</td><td>98.2</td><td>161.2</td><td>178.5</td><td>165.3</td><td>159.5</td><td>145.8</td><td>159.6</td><td>110.7</td></tr>
<tr><td><b>y</b></td><td>77.9</td><td>76.8</td><td>80.1</td><td>80.2</td><td>79.0</td><td>79.9</td><td>79.0</td><td>79.0</td><td>78.6</td></tr>
</table>
Obtain the fitted values of y using simple regression model and also find the value of y, when x = 180.</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>08</b></td></tr>
</table>

<table>
<tr><td></td><td></td><td>b)</td><td>Summarize the test procedure to test the significance of the simple linear regression model.</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Provide a brief overview of the following topics:
i) the applications of regression analysis, and
ii) how simple regression analysis is used for predicting new observations.</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - II</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>3</td><td>a)</td><td>Provide a brief overview of Q-Q plots, including their purpose and how they are used to assess the normality of data.</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Consider the following dataset related to the performance of sorting algorithms:
<table>
<tr><th>Execution Time in milliseconds (y)</th><th>Number of Elements (X1)</th><th>Input Complexity (X2)</th></tr>
<tr><td>78.5</td><td>7</td><td>26</td></tr>
<tr><td>74.3</td><td>1</td><td>29</td></tr>
<tr><td>104.3</td><td>11</td><td>56</td></tr>
<tr><td>87.6</td><td>11</td><td>31</td></tr>
<tr><td>95.9</td><td>7</td><td>52</td></tr>
<tr><td>109.2</td><td>11</td><td>55</td></tr>
</table>
Provide the least squares estimates for the regression coefficients and summarize the fitted regression model.</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>07</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>When the problem of heteroscedasticity exists in the data? Explain the role of residual graphs to detect the same.</td><td><i>CO2</i></td><td><i>PO1</i></td><td><b>07</b></td></tr>
<tr><td></td><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>4</td><td>a)</td><td>Give a brief note on sources and detection of multicollinearity in multiple regression analysis.</td><td><i>CO2</i></td><td><i>PO1</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Given the following dataset:
<table>
<tr><th>Observation</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th></tr>
<tr><td><b>Rainfall (cms)</b></td><td>30</td><td>23</td><td>34</td><td>31</td><td>17</td><td>36</td></tr>
<tr><td><b>Yield (tons)</b></td><td>65</td><td>62</td><td>70</td><td>64</td><td>52</td><td>68</td></tr>
</table>
Calculate the Durbin-Watson d statistic to test positive autocorrelation and conclude. ($d_L = 0.61$ and $d_U = 1.40$)</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Compare and contrast the regularization techniques of Ridge regression and LASSO regression. What are the key differences in their methodologies and effects on model performance?</td><td><i>CO2</i></td><td><i>PO2</i></td><td><b>08</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - III</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>5</td><td>a)</td><td>“Not all leverage points are outliers” – Justify the statements with an example.</td><td><i>CO1</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>How the DFFITs and DFBETAs are helpful to identify the influential observations.</td><td><i>CO2</i></td><td><i>PO1</i></td><td><b>06</b></td></tr>
</table>

c) Given the table below with columns for Model No, p (number of parameters including the intercept), R-Squared (R²), Adjusted R-Squared (Adj R²), Standard Error (S), and Mallows' Cp: — 08

Model	p	R²	Adj R²	S	Mallows' Cp
1	1	0.64	0.62	0.50	9.4
2	1	0.62	0.60	0.52	10.7
3	2	0.80	0.77	0.39	1.5
4	2	0.79	0.76	0.40	3.3
5	3	0.81	0.76	0.40	3.2
6	3	0.80	0.76	0.40	3.4
7	4	0.81	0.74	0.41	5.0

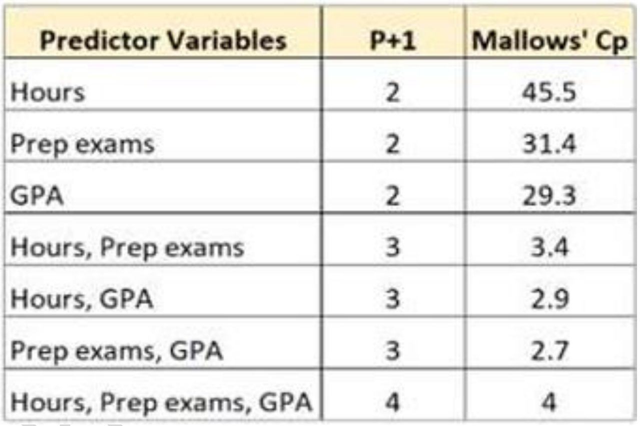
Select the model that best fits the data and justify your answer.
Justify why other models are not suitable.

OR

6 a) Explain added-variable plots. — *CO2* *PO1* **06**

b) Describe the Cook’s distance to detect the influential observations. — *CO2* *PO1* **06**

c) Define Mallow’s C_p and select the best model for the below given output. — *CO2* *PO1* **08**



Predictor Variables	P+1	Mallows' Cp
Hours	2	45.5
Prep exams	2	31.4
GPA	2	29.3
Hours, Prep exams	3	3.4
Hours, GPA	3	2.9
Prep exams, GPA	3	2.7
Hours, Prep exams, GPA	4	4

UNIT - IV

7 a) Identify the type of categorical scale (nominal or ordinal) used in each of the following scenarios and explain your reasoning: — *CO1* *PO2* **06**

1. **Customer satisfaction ratings**: "Very Unsatisfied," "Unsatisfied," "Neutral," "Satisfied," "Very Satisfied."
2. **Type of pet owned**: "Dog," "Cat," "Bird," "Fish," "Other."
3. **Highest level of education:** "High School," "Associate Degree," "Bachelor's," "Master's," "Doctorate."

b) Write down the test procedure to test the independence of attributes. — *CO2* *PO1* **06**

c) Elucidate the following. — *CO3* *PO1* **08**

i. Sensitivity
ii. Specificity
iii. ROC curve

OR

<table>
<tr><td></td><td>8</td><td>a)</td><td>In a study of 263 adolescents evaluated for suicidal behaviour, 186 were classified as non-suicidal (NS) at a six-month follow-up. Among them, 86 were assessed as having depression at baseline. Of the 77 adolescents with persistent suicidal behaviour (SB) at follow-up, 45 had been assessed for depression at baseline.
i. Construct a contingency table based on this data.
ii. Calculate the odds ratio and provide an interpretation.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>08</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>The following is the data regarding family condition and examination result of 100 students test whether family conditions and results are independent (critical value is 6.63).
<table>
<tr><td rowspan="2">Family condition</td><td colspan="2">Examination results</td></tr>
<tr><td>Pass</td><td>Fail</td></tr>
<tr><td>Good</td><td>30</td><td>10</td></tr>
<tr><td>Bad</td><td>20</td><td>40</td></tr>
</table></td><td><i>CO3</i></td><td><i>PO3</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Explain how logistic regression is used for binary classification and discuss its advantages. Additionally, mention other machine learning methods that can also be applied for binary classification tasks.</td><td><i>CO1</i></td><td><i>PO1</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - V</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>9</td><td>a)</td><td>Discuss the different types of Markov models and their applications in various fields.</td><td><i>CO3</i></td><td><i>PO1</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Explain the forward and backward algorithm of hidden Markov models.</td><td><i>CO3</i></td><td><i>PO1</i></td><td><b>08</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Describe the smoothing and filtering techniques of classification using hidden Markov models.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>10</td><td>a)</td><td>State the procedure of choosing number of hidden states in hidden Markov models.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>08</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Provide pseudocode for the working of Viterbi Algorithm.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
<tr><td></td><td></td><td>c)</td><td>Discuss the Gaussian mixture models with hidden Markov models.</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>06</b></td></tr>
</table>

******