

Name: _____

Roll Number: _____

Subject: Data Science Date: _____ Section: _____

Provide justification or calculation in the space available in front of the options. No marks will be given if the justification is incorrect.

1. You are building a Constant Model for the dataset $D = \{2, 4, 4, 10, 30\}$. If you choose to minimize the Mean Absolute Error (MAE), what is your optimal prediction $\hat{\theta}$?
 - A. 10
 - B. 4
 - C. 2
 - D. 30
2. Using the same dataset $D = \{2, 4, 4, 10, 30\}$, if you switch your loss function to Mean Squared Error (MSE), what is the new optimal $\hat{\theta}$?
 - A. 4
 - B. 10
 - C. 12
 - D. 50
3. A dataset has summary statistics: $\bar{x} = 10$, $\bar{y} = 20$, $\sigma_x = 2$, $\sigma_y = 6$, and correlation $r = 0.5$. Calculate the slope $\hat{\theta}_1$ of the least squares regression line. Provide a calculation.
 - A. 0.5
 - B. 1.5
 - C. 3.0
 - D. 6.0
4. Using the results from the previous question ($\bar{x} = 10$, $\bar{y} = 20$, slope, calculate the intercept $\hat{\theta}_0$).
Provide a calculation.
 $\xrightarrow{\text{calculated above}}$
 - A. 5
 - B. 35
 - C. 0
 - D. 15
5. You fit a linear model and plot the residuals against the fitted values. You see a distinct "U-shape" or parabolic curve in the residuals. What does this specifically indicate?
 - A. The noise in the data is random.
 - B. The data has a non-linear structure that the linear model failed to capture.
 - C. The correlation is 0.
 - D. You should use a Constant Model instead.
6. If the correlation coefficient r between two variables is exactly -1 , and you fit a standard regression line, what will the Root Mean Squared Error (RMSE) be? Provide a reason.
 - A. 1

- B. -1
C. 0
D. It depends on the standard deviations.
7. Which of the following scenarios best justifies applying a log transformation to the target variable (y)?
A. When the scatter plot shows a linear relationship but with high variance.
B. When the residuals are perfectly random.
C. When the values of y span several orders of magnitude and the relationship appears multiplicative/exponential.
D. When the inputs x are binary (0 or 1).
8. Which loss function penalizes large errors quadratically? Provide a reason. How?
A. L1 Loss
B. L2 Loss
C. Zero-One Loss
D. Huber Loss
9. Which statistic minimizes the sum of absolute deviations? Provide a reason. How?
A. Median
B. Mean
C. Mode
D. Range
10. Which of the following NumPy snippets correctly calculates the Root Mean Square Error (RMSE) given arrays y_{true} and y_{pred} ? Write the formula in the justification.
A. `np.mean((ytrue - ypred)**2)`
B. `np.sqrt(np.mean((ytrue - ypred)**2))`
C. `np.sqrt(np.sum(ytrue - ypred))`
D. `np.mean(np.sqrt(ytrue - ypred))`
11. Which Python snippet correctly computes the residuals of a model? Write the formula in the justification.
A. `residuals = yobserved - ypredicted`
B. `residuals = yobserved + ypredicted`
C. `residuals = ypredicted / yobserved`
D. `residuals = (yobserved - ypredicted)** 2`