

# Linear Regression Assignment Answers

## Assignment-Based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical variables such as season and weather have significant impact on bike demand (cnt). Seasons 3 and 4 have higher demand, while adverse weather (weathersit\_3) decreases demand.

2. Why is it important to use drop\_first=True during dummy variable creation?

It prevents multicollinearity by avoiding the dummy variable trap, ensuring the regression model estimates coefficients accurately.

3. Which numerical variable has the highest correlation with the target variable?

Feels-like temperature (atemp) has the highest positive correlation with bike demand.

4. How did you validate the assumptions of Linear Regression after building the model?

Assumptions were validated through residual plots for homoscedasticity and linearity, Durbin-Watson statistic for autocorrelation, and Q-Q plots for normality.

5. Based on the final model, which are the top 3 features contributing significantly to demand?

Year (yr), feels-like temperature (atemp), and season 4 are the top 3 features significantly explaining demand.

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression models the relationship between a dependent variable and independent variables using a linear equation. The objective is to minimize the difference between observed and predicted

values, often using the least squares method.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets with nearly identical statistical properties but very different distributions when plotted. It shows the importance of visualizing data.

3. What is Pearson's R?

Pearson's R measures the linear correlation between two variables, ranging from -1 to 1, where 1 indicates perfect positive correlation and -1 indicates perfect negative correlation.

4. What is scaling? Why is it performed? What is the difference between normalized scaling and standardized scaling?

Scaling adjusts the range of features for model consistency. Normalized scaling adjusts values between 0 and 1, while standardized scaling centers the data around 0 with a standard deviation of 1.

5. Why does the value of VIF sometimes become infinite?

VIF becomes infinite due to perfect multicollinearity, meaning one variable is an exact linear combination of others.

6. What is a Q-Q plot? Explain its importance in linear regression.

A Q-Q plot is used to check if residuals follow a normal distribution, a key assumption in linear regression. A straight line in the Q-Q plot suggests normality.