Bike Sharing:

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? (3 marks)

weekday, mnth, holiday, temp, season and year are categorical variables in the dataset.

2019 has more sales
Working-day and September has more active customers.
Weathersit - more bikes are rent during Clear, Few clouds, Partly cloudy weather.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

To avoid dummary variable correlation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

weathersit(Light_Snow) has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

By error terms correspond to a normal curve in a histogram.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

* weathersit_Light_Snow (negative correlation)
* season_spring (negative correlation)
* yr_2019 (positive correlation)
* mnth_Nov (negative correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression predicts how independent variables are influenced by dependent ones. After cleaning and exploring the data, we split it into training and testing sets. We assess variable collinearity, select relevant variables, and iteratively refine the model by checking R-values and p-values.

Assuming normally distributed errors, we test the model with the test dataset. The final model provides valuable insights and predictions within its range.

2. Explain the Anscombe’s quartet in detail.

A regression model can be misled by cleverly structured data. Anscombe's quartet illustrates this with four different datasets that produce identical regression models despite having distinct characteristics.

3. What is Pearson’s R?

Pearson's correlation coefficient (Pearson's R) measures the strength and direction of the correlation between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with values in between indicating the degree of correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is essential for a model to function correctly by aligning variable ranges. For instance, car sales dependent on price and months would require scaling due to their different ranges. Two types of scaling are:

- Normalized scaling: Adjusts data to a Gaussian distribution without a preset range, commonly used in neural networks.
- Standardized scaling: Compresses variable values to ensure consistent scaling, preventing decimal errors.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

When there's a perfect correlation between the dependent variable and independent variable(s), the R-squared value reaches 1. Consequently, the VIF (Variance Inflation Factor), approaches infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A Q-Q plot assesses if datasets share the same statistical distribution, crucial in linear regression for comparing training and testing datasets.
(base) asj@asj-mac:~/Downloads/bike-sharing/assignment$