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)

Answer:

a. Most of the bikes are rented in fall.
b. Bikes are rented more on a holiday as compared to non holiday days
c. Users prefer to take bikes in clear or partly cloudy weather.
d. There is no entry in the database for weathersit heavy rainfall and snow.

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

Answer: Using drop_first=True during dummy variable creation is important to avoid multicollinearity issues and maintain the correct number of degrees of freedom in regression models. It's important to avoid multicollinearity

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

Answer: "registered" users has the highest correlation with cnt.

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

Answer: Checking the R2 value for y_predicted_test and actual y_test. The R2 value is 82.9 which is a good score.

5. **Based on the final model, which are the top 3 features contributing significantly towards
explaining the demand of the shared bikes?** (2 marks)

Answer: atemp, yr and weather

General Subjective Questions

1. **Explain the linear regression algorithm in detail.** (4 marks)

Answer:

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as y) and one or more independent variables (often denoted as (x_1, x_2, ..., x_n)). The fundamental assumption of linear regression is that there exists a linear relationship between the independent variables and the dependent variable. The goal of linear regression is to find the best-fitting linear equation that describes this relationship.

Here's a detailed explanation of the linear regression algorithm-

- **Model Representation:**

In linear regression, the relationship is represented by the following linear equation:

[y = beta_0 + \beta_1 x_1 + beta_2 x_2 + ... + beta_n x_n]

where:

- (y) is the dependent variable.

- (x_1, x_2, ..., x_n) are the independent variables.

- (beta_0, beta_1, ..., beta_n) are the coefficients (parameters) of the linear equation.

- **Coefficient Estimation:**

To estimate the coefficient (beta_0, beta_1, ..., beta_n), various optimization techniques can be used. The most common method is to use calculus to minimize the cost function (the sum of squared residuals) with respect to the parameters. This results in a closed-form solution for the parameters.

- **Model Evaluation:**

Once the model parameters are estimated, we evaluate the goodness of fit using various metrics such as the coefficient of determination ((R^2)), mean squared error (MSE), or root mean squared error (RMSE). These metrics help assess how well the linear model fits the data.

- **Predictions:**

Once the model is trained and evaluated, it can be used to make predictions on new or unseen data. Given new values of the independent variables, we can use the learned parameters to predict the corresponding values of the dependent variable.

Linear regression is widely used in various fields including economics, finance, social sciences, and machine learning due to its simplicity, interpretability, and effectiveness in modeling linear relationships between variables. However, it's important to note that linear regression assumes a linear relationship between the independent and dependent variables, and the presence of non-linear relationships may require more complex modeling techniques.

2. **Explain the Anscombe's quartet in detail.** (3 marks)

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation, and regression coefficients), yet have very different distributions and appear very different when graphed. This illustrates the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the limitations of summary statistics and the importance of visualizing data. Here's an overview of the four datasets in Anscombe's quartet:

1. Dataset I:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

2. Dataset II:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

- y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

3. Dataset III:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

- y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

4. Dataset IV:

- x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8

- y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

Despite having the same mean, variance, correlation coefficient, and linear regression line for each dataset, the data points in each dataset have distinct distributions and relationships when graphed.

- Dataset I: Forms a fairly linear relationship.

- Dataset II: Forms a curve.

- Dataset III: Appears to have an outlier which affects the regression line.

- Dataset IV: Appears to have a high leverage point which also affects the regression line.

The key lesson from Anscombe's quartet is that summary statistics alone may not fully capture the characteristics of a dataset. Visual inspection and exploratory data analysis (EDA) are essential for understanding the underlying patterns and relationships within the data. It also underscores the importance of robust statistical techniques that are less sensitive to outliers and other anomalies.

3. **What is Pearson's R?** (3 marks)

Answer:

Pearson's correlation coefficient, often denoted as r or Pearson's r, is a measure of the strength and direction of the linear relationship between two variables. It quantifies the degree to which two variables are linearly related to each other. Pearson's r ranges from -1 to 1, where

(r = 1) indicates a perfect positive linear relationship (as one variable increases, the other also increases by a consistent amount).(r = -1) indicates a perfect negative linear relationship (as one variable increases, the other decreases by a consistent amount).(r = 0) indicates no linear relationship between the variables.

Formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

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

Answer:

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

Answer:

The Variance Inflation Factor (VIF) measures the extent of multicollinearity in a regression analysis. Specifically, it quantifies how much the variance of a coefficient estimate is inflated due to multicollinearity with other predictor variables.

The formula for VIF for a given predictor variable is:

VIF = 1/(1 - R^2)

where   R^2 value obtained by regressing the predictor variable in question against all other predictor variables.

When the value of VIF is infinite, it indicates perfect multicollinearity between the predictor variable and the other variables in the model. Perfect multicollinearity means that one or more of the predictor variables can be expressed as a perfect linear combination of the other predictor variables. This situation typically arises due to one of the following reasons:

- **Duplicated or Linearly Dependent Variables**: One or more predictor variables are identical or can be expressed as a perfect linear combination of the other predictor

variables. For example, having two variables that are exactly the same or one variable that is the sum or difference of two others.
- **Too Many Predictor Variables Relative to Sample Size**: When there are too many predictor variables relative to the sample size, it increases the risk of multicollinearity, especially if the variables are highly correlated with each other.
- **Data Transformation**: Sometimes, data transformation techniques such as creating dummy variables or polynomial features can inadvertently introduce multicollinearity, especially if not handled carefully.

When VIF is infinite, it poses serious problems for regression analysis. It means that the affected predictor variable cannot be independently assessed or interpreted, and it may lead to unstable coefficient estimates. In such cases, it's essential to identify and address the root cause of multicollinearity, which may involve removing redundant variables, collecting more data, or using regularization techniques to stabilize the regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** (3 marks)

Answer:

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a particular probability distribution, suchs as the normal distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, usually the normal distribution.

Here's how a Q-Q plot works:
- The observed data is sorted in ascending order.
- The quantiles of the observed data are plotted against the quantiles of the theoretical distribution.
- If the data closely follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.

The use and importance of a Q-Q plot in linear regression are as follows:

- **Assumption Checking**: Linear regression relies on certain assumptions, including the assumption of normally distributed errors. Q-Q plots are used to visually inspect whether the residuals (the differences between observed and predicted values) are normally distributed. If the residuals follow a normal distribution, the points on the Q-Q plot will approximately lie on a straight line.
- **Detection of Departures from Normality**: Departures from normality in the residuals can indicate violations of linear regression assumptions, such as heteroscedasticity or model misspecification. Q-Q plots allow for the detection of departures from normality, such as skewness or heavy tails, which may warrant further investigation or model refinement.

- **Model Improvement**: If the Q-Q plot reveals significant departures from normality, corrective actions can be taken to improve the linear regression model. This may involve transforming the response variable or using robust regression techniques that are less sensitive to non-normality in the residuals.

In summary, Q-Q plots are a valuable diagnostic tool in linear regression analysis. They help assess the assumption of normally distributed errors, detect departures from normality, and guide model improvement efforts to ensure the validity and reliability of regression results.