Linear Regression Assignment Questions

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?**

Ans:

1. Fall season has the highest demand for bikes and consistent growth in the month of June and the highest in September. After September, the demand is going downhill.

2. In contrary to my expectations, demand for bikes is decreasing on holidays, maybe most people are using it for daily commute.

3. There is an increase in demand for the coming year

4. Temperature, Humidity and windspeed also has effects on the count (target variable.)

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

Ans:

Using drop_first=True during dummy variable creation helps to avoid multicollinearity issues and reduce the dimensionality of the feature space, which can improve the performance and interpretability of machine learning models.

In my analysis , while creating dummy variable for weathersit, first column was not dropped so as not to lose the info about severe weather situation.

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

Ans: Temp and atemp have the highest correlation with the target variable (cnt).

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

Ans:

R2 value for test predictions: R2 value for predictions on test data (0.812) is almost same as R2 value of train data(0.816). This is a good evaluation metrics to see how well the model performs on unseen data.

Residual Analysis: Errors are normally distributed with a mean of 0. Actual and predicted result follow the same pattern. The error terms are independent of each other.

Test vs Predicted value test: The prediction for test data is very close to actuals.

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

Ans:

The top 3 features are:

1. temp (positive corelation)
2. yr (positive corelation)
3. weathersit_bad (negative corelation)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (also called the response variable) and one or more independent variables (also called predictor variables or features). It assumes that there is a linear relationship between the independent variables and the dependent variable. The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the observed values of the dependent variable and the values predicted by the model.

Steps in Linear Regression:

Model Representation:

- Linear regression models the relationship between the dependent variable y and the independent variables x1,x2,,..,xn using the equation:

$$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn + \varepsilon$$

Model Training:

- The goal of model training is to estimate the coefficients β0,β1,...,βn that best fit the observed data.
- This is typically done using a method such as ordinary least squares (OLS), which minimizes the sum of the squared differences between the observed and predicted values of y.

Model Evaluation:

- Once the model has been trained, it's important to evaluate its performance.
- Common evaluation metrics for linear regression models include:
- Mean Squared Error (MSE): The average of the squared differences between observed and predicted values.

- R-squared (R2): A measure of how well the independent variables explain the variability of the dependent variable. It ranges from 0 to 1, with higher values indicating a better fit.
- Adjusted R-squared: A modified version of R2 that adjusts for the number of predictors in the model.

***Predictions**:*

- Once the model has been evaluated, it can be used to make predictions on new or unseen data.

Linear regression is a versatile and widely used algorithm in various fields such as statistics, economics, finance, and machine learning.

2. Explain the Anscombe's quartet in detail.

Ans:

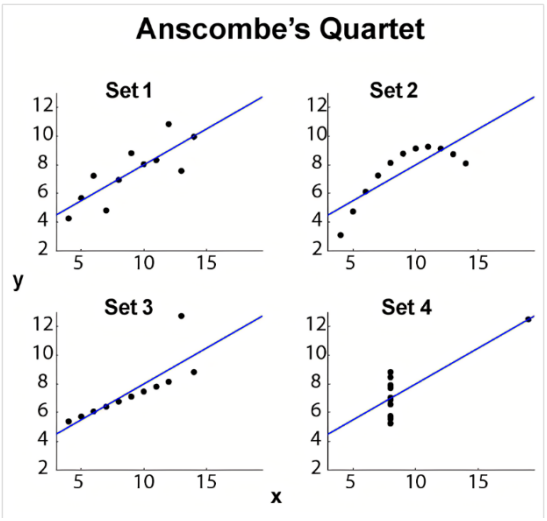
Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including means, variances, correlation coefficients, and linear regression lines. However, visually, these datasets exhibit vastly different patterns when plotted. This demonstration underscores the importance of visualizing data before drawing conclusions and highlights the limitations of relying solely on summary statistics.

Detailed explanation of Anscombe's quartet:

Description: Anscombe's quartet comprises four datasets, each containing 11 (x, y) pairs. Despite having similar statistical properties, the datasets display different patterns when graphically plotted.

The Datasets:

- Dataset I: Consists of linear data with a slight scatter.
- Dataset II: Also linear but with one outlier, which significantly affects the regression line.
- Dataset III: Forms a non-linear pattern, where the data fits a quadratic curve.
- Dataset IV: Appears to have no clear pattern until an outlier is removed, after which it shows a perfect fit for a linear model.



3. What is Pearson's R?

Ans : Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol r and ranges from -1 to 1. The formula for calculating Pearson's r between two variables X and Y is:

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

Where:

- n is the number of data points.
- X_i and Y_i are individual data points for variables X and Y, respectively.
- $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively.

Properties of Pearson's R:

- Range: $-1 \leq r \leq 1$
- Positive r indicates a positive linear relationship; negative r indicates a negative linear relationship.
- The closer r is to 1 or -1, the stronger the linear relationship.
- $r = 0$ suggests no linear relationship.
- Independence of Scale: Unaffected by changes in scale or units of measurement.

Pearson's *r* is widely used in statistics, social sciences, and natural sciences to assess linear associations between variables. However, it only captures linear relationships and may not detect nonlinear or monotonic relationships.

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

Ans :

Scaling is a preprocessing technique used in machine learning to standardize the range of independent variables or features in a dataset. It involves transforming the data so that it falls within a specific range or distribution. The primary goal of scaling is to ensure that all features contribute equally to the analysis, prevent variables with larger scales from dominating those

with smaller scales, and improve the performance and convergence of machine learning algorithms.

Why Scaling is Performed:

- Equalize Variable Influence.
- Enhance Convergence.
- Improve Interpretability.

Normalized Scaling vs. Standardized Scaling:

- Normalized Scaling (Min-Max Scaling): Scales features to a specific range (typically 0 to 1). Useful for non-Gaussian distributions.
- Standardized Scaling (Z-score Scaling): Scales features to have a mean of 0 and standard deviation of 1. Suitable for Gaussian distributions.

Both techniques achieve similar goals but are chosen based on feature distribution characteristics and algorithm requirements.

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

Ans:

VIF (Variance Inflation Factor) can sometimes be infinite, and this typically occurs when one or more independent variables in a regression model are perfectly collinear with each other.

When two or more variables are perfectly collinear, it means that one variable can be expressed exactly as a linear combination of the others. In such cases, the regression model cannot estimate separate coefficients for these variables, resulting in an infinite VIF.

Mathematically, VIF is calculated as $VIF_i = \frac{1}{1-R_i^2}$, where R_i^2 is the coefficient of determination (R-squared) of the regression of the i-th independent variable on all other independent variables. When R_i^2 is 1, meaning perfect collinearity, the denominator becomes zero, resulting in an infinite VIF.

In practical terms, encountering infinite VIF values indicates a severe multicollinearity problem in the dataset, and it usually requires remedial action such as dropping one of the highly correlated variables or applying dimensionality reduction techniques to mitigate multicollinearity and stabilize the regression model.

6. . What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans:

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a given dataset

follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, typically a normal distribution. The plot displays the quantiles of the dataset on the horizontal axis and the quantiles of the theoretical distribution on the vertical axis.

Use and Importance of Q-Q Plot in Linear Regression:

Assumption Checking:

Q-Q plots are commonly used to check the assumption of normality in linear regression residuals. Residuals should ideally follow a normal distribution with a mean of zero and constant variance.

Identifying Outliers:

Outliers can significantly impact the regression model's performance.

Model Fit Evaluation:

A well-fitting linear regression model should yield residuals that closely follow a normal distribution. A Q-Q plot provides a visual assessment of how well the residuals conform to this expectation.

Model Interpretation:

When the residuals are normally distributed, it enhances the interpretability of the regression coefficients.