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-

   Here are the inferences related to the effect of categorical variables (dummy variables) on the dependent variable (cnt - likely representing bike counts/rentals):

   yr (Year): A positive coefficient (0.2469) suggests that as the year increases, there is an increase in bike rentals. This might indicate a general trend of increasing bike rentals over time.

   holiday: A negative coefficient (-0.0836) implies that on holidays, there tends to be a slight decrease in bike rentals compared to non-holidays.

   Seasonal Variables and Month variables (spring, 3, 5, 6, 8, 9, 7, 10): These represent different seasons or months. Negative coefficients for spring and positive coefficients for other months indicate varying effects based on the season or month. For instance, rentals might decrease in spring (spring coefficient: -0.1980) compared to other seasons (e.g., months 3, 5, 6, 8, 9, 7, 10) where rentals show an increase.

   Weather Conditions (Light_Snow_Light_Rain_Thunderstorm, Mist_Cloudy): Both variables have negative coefficients, suggesting that during light snow, light rain, thunderstorms, or misty, cloudy weather, there tends to be a decrease in bike rentals.

   Day of the Week (Sunday): The negative coefficient (-0.0498) indicates a slight decrease in bike rentals on Sundays compared to other days of the week.

   Overall, the coefficients of the categorical variables in the regression model indicate how each category or group within these variables influences the predicted bike rentals. Positive coefficients imply an increase in rentals, while negative coefficients suggest a decrease, relative to the reference category (dropped category due to drop_first=True for dummy variable encoding). These coefficients help in understanding the impact of different categorical factors on the bike rental counts which is “cnt”.

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

   Answer-

   It's important to use drop_first=True during dummy variable creation to mitigate multicollinearity issues and maintain model simplicity in regression analysis.

   Multicollinearity Mitigation:

   - When creating dummy variables to represent categorical variables in regression models, dropping the first category (using drop_first=True) helps prevent multicollinearity.

- Including all dummy variables without dropping one can lead to multicollinearity, where the predictors become highly correlated. This can adversely affect the regression model by inflating standard errors, making coefficients unstable or difficult to interpret.

Simplicity and Efficiency:

- Dropping the first category results in a more parsimonious model, reducing redundancy in the predictors.
- Including fewer dummy variables (by dropping one) maintains model efficiency by simplifying the interpretation of coefficients. The dropped category becomes the reference point, and the coefficients of the remaining dummy variables represent the effect relative to this reference category.

By using drop_first=True, one avoids multicollinearity issues, improves model interpretability, and ensures efficient representation of categorical variables in regression models.

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

   Answer-

   During the analysis found temp and atemp have the highest correlation with the target variable among numerical variables. Both temp and atemp have a correlation coefficient of 0.63 with the target variable cnt. This indicates a relatively strong positive linear relationship between these variables (temp, atemp) and the bike rental count (cnt).
   High positive correlations like these imply that as the temperature (temp or atemp) increases, there tends to be a notable increase in bike rentals. Therefore, among the listed numerical variables, temp and atemp exhibit the highest correlation with the target variable (cnt) in the pair-plot analysis.

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

   Answer-

   I have used validation methods like linearity, Homoscedasticity, no autocorrelation, normality of residuals, no multicollinearity. However, after validating all assumptions, we examine additional diagnostic plots e.g. residual plots.

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-

   Based on the OLS regression results coefficients and their associated p-values:

   - The top three features contributing significantly to explaining the demand for shared bikes (cnt) in the final model lm15 are:
     - yr (Year) with a positive coefficient (0.2469)

- Light_Snow_Light_Rain_Thunderstorm (Weather Condition) with a negative coefficient (-0.3212)
- spring (Season) with a negative coefficient (-0.1980)

- These variables, based on their coefficients and significance levels, have a notable impact on bike demand, highlighting the influence of year, weather conditions (specifically light snow, light rain, or thunderstorms), and season (spring) on bike rentals.

General Subjective Questions

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

   Answer

   Linear regression is a fundamental and widely used supervised learning algorithm in machine learning and statistics, particularly for predictive modeling and understanding the relationship between a dependent variable and one or more independent variables. Let's dive into a detailed explanation of the linear regression algorithm:

   Objective of Linear Regression:

   The primary goal of linear regression is to model the relationship between a dependent variable (target or response variable denoted as

   y) and one or more independent variables (predictors denoted as

   x1,x2,...,xn).

   Assumptions of Linear Regression:

   Linearity: Assumes a linear relationship between predictors and the target variable.

   Independence: Assumes independence between observations.

   Homoscedasticity: Assumes constant variance of errors (residuals) across the range of predictors.

   Normality: Assumes residuals are normally distributed.

   No multicollinearity: Assumes little or no multicollinearity among predictors.

   Working of Linear Regression:

   Simple Linear Regression: In its simplest form, it deals with a single independent variable (x) to predict the dependent variable (y) using a linear equation:

$y = \beta 0 + \beta 1.x + \varepsilon$

- $\beta 0$ - is the intercept (constant term).
- $\beta 1$ - is the coefficient for the independent variable (x).
- ε - represents the error term.

Multiple Linear Regression: Extends to multiple predictors (x1,x2,...,xn) to predict

$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + ... + \beta n.xn + \varepsilon$

- $\beta 0$ - is the intercept.
- $\beta 1, \beta 2, ..., \beta n$ - are the coefficients for each independent variable.
- ε represents the error term.

Model Training:

Objective: To estimate the coefficients ($\beta 0, \beta 1, ..., \beta n$) that minimize the difference between actual and predicted values.

Cost Function: Typically, the Ordinary Least Squares (OLS) method minimizes the sum of squared differences between observed and predicted values.

Gradient Descent or Analytical Methods: Algorithms like gradient descent or analytical methods (matrix operations) can be used to find the optimal coefficients.

Model Evaluation:

R-squared (Coefficient of Determination): Measures the proportion of variance explained by the model. Higher R-squared indicates a better fit.

Residual Analysis: Checking residuals (errors) to validate assumptions like normality, homoscedasticity, and independence.

Adjusted R-squared: Adjusts R-squared for the number of predictors to penalize excessive variables.

Prediction:

Once the model is trained and validated, it can be used to predict the dependent variable (y) for new or unseen data based on the learned coefficients.

Conclusion:

Linear regression is a simple, yet powerful algorithm used for modeling relationships between variables and making predictions. Its simplicity, interpretability, and effectiveness make it a widely used technique in various fields, serving as a foundation for more complex regression-based models.

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

Answer-

Anscombe's quartet is a famous example in statistics that comprises four datasets with nearly identical statistical properties, yet vastly different when visualized. These datasets were created by Francis Anscombe in 1973 to emphasize the importance of visualizing data and the limitations of relying solely on summary statistics.

Characteristics of Anscombe's Quartet:

- Four Distinct Datasets: Anscombe's quartet consists of four sets of x-y data pairs.
- Statistical Properties: Despite having different distributions, means, variances, and correlation coefficients, when looking at summary statistics, these datasets are strikingly similar.
- Diverse Relationships: Each dataset displays a unique relationship between the independent (x) and dependent (y) variables, showcasing examples of linear, non-linear, or even outliers.
- Importance of Visualization: While summary statistics might appear identical, when plotted, the datasets show significant differences, highlighting the necessity of visual exploration.

Description of Anscombe's Quartet Datasets:

- Dataset I: Linear relationship; simple and clear. Fits well with a linear regression model.
- Dataset II: Also a linear relationship but with an outlier that influences the regression line and correlation.
- Dataset III: Appears to be non-linear with a strong relationship, forming a curve. Challenging for linear regression models.
- Dataset IV: Appears to have no relationship when looking at summary statistics, but actually has a perfect relationship when plotted, except for an outlier.

Importance and Implications:

- Visualizing Data: Emphasizes the importance of data visualization in understanding relationships that summary statistics might not reveal.
- Caution in Analysis: Shows that reliance on summary statistics alone can be misleading, especially in complex datasets.
- Model Assumptions: Highlights the significance of checking assumptions (linearity, homoscedasticity) before applying regression models.

Conclusion:

Anscombe's quartet serves as a cautionary example, demonstrating that datasets with identical statistical properties can exhibit vastly different patterns when plotted visually. It underscores the necessity of not solely relying on summary statistics but incorporating visualizations to comprehend and interpret data effectively. This quartet remains an essential teaching tool in statistics to emphasize the importance of exploring and understanding data through visualization.

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

Answer-

Pearson's correlation coefficient, denoted as r or Pearson's r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses how closely the data points in a scatter plot cluster around a straight line, indicating the degree of association between the variables.

Characteristics of Pearson's Correlation Coefficient:
Range: The value of r ranges between -1 and +1.
Direction: The sign of r (+ or -) indicates the direction of the relationship.

r=+1 indicates a perfect positive linear relationship, where variables increase together.
r=−1 indicates a perfect negative linear relationship, where one variable decreases as the other increases.
r=0 indicates no linear relationship between variables.

Strength: The closer
r is to +1 or -1, the stronger the linear relationship. The closer it is to 0, the weaker the relationship.
Formula: Pearson's r is calculated as the covariance of the variables divided by the product of their standard deviations.
Formula for Pearson's Correlation Coefficient (r):
Given two variables X and Y with observations (xi,yi) for i=1,2,...,n:

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \cdot \sum (y_i - \bar{Y})^2}}$$

Where:
$\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.
∑ denotes the sum across all observations.

Interpretation of r:
r=1 or r=−1: Perfect linear relationship.
0.7<|r|≤1: Strong positive or negative linear relationship.
0.3<|r|≤0.7: Moderate linear relationship.
|r|≤0.3: Weak linear relationship or no linear relationship.

Use of Pearson's Correlation:
Pearson's r is widely used in various fields, including statistics, social sciences, economics, and machine learning, to quantify and understand the linear relationship between variables. It assists in feature selection, understanding associations between variables, and assessing the strength of predictive relationships in regression analysis, among other applications.

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

Answer-

It is the process of transforming data to a standard range or distribution. It is performed to bring all the features or variables of a dataset onto a similar scale, ensuring fair comparisons and improving the performance of certain machine learning algorithms.

Why Scaling is Performed:

- Algorithm Performance: Some machine learning algorithms are sensitive to the scale of features. Scaling helps algorithms converge faster and perform better by treating all features equally.
- Distance-Based Algorithms: Algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) that rely on distance measures between data points can be affected by feature scales. Scaling prevents features with larger scales from dominating.

Gradient Descent: In optimization algorithms like gradient descent, scaling helps in reaching the minimum faster by ensuring a more symmetrical and narrow shape of the cost function.

Types of Scaling:

Normalized Scaling:

- Normalization scales the features between 0 and 1, bringing the values within a specific range.
- Formula: X-Xmin/Xmax - Xmin

- Pros: Preserves the relationship between values; useful when features have varying ranges.
- Cons: Sensitive to outliers; may not handle extreme values well.

Standardized Scaling (Standardization):

- Standardization transforms features to have a mean of 0 and a standard deviation of 1.
- Formula: $X-\mu/\sigma$
- Pros: Less affected by outliers; maintains the shape of the distribution.
- Cons: Does not bound values to a specific range.

Differences between Normalized Scaling and Standardized Scaling:

Range:

- Normalized scaling brings values within the range of 0 to 1, maintaining relative proportions between them.

- Standardized scaling centers the data around the mean of 0 and adjusts by the standard deviation, without bounding values to a specific range.

Sensitivity to Outliers:

- Normalized scaling can be highly sensitive to outliers, affecting the scale range.
- Standardized scaling is less affected by outliers due to its use of mean and standard deviation.

Preservation of Distribution:

- Normalization preserves the original distribution and proportions between values.
- Standardization maintains the shape of the distribution but centers the data around 0.

Conclusion:

Scaling is essential in preprocessing data to ensure that different features contribute equally to the analysis and modeling process. Normalized scaling and standardized scaling offer different approaches to transforming data, each with its own advantages and considerations depending on the specific requirements of the dataset and the machine learning algorithm being used.

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

Answer-

VIF is a measure used to detect multicollinearity in regression analysis by quantifying how much the variance of an estimated regression coefficient is inflated due to collinearity among predictors. It measures how much the variance of the estimated regression coefficients increases if your predictors are correlated.

In the context of VIF becoming infinite, this occurs when one or more predictors in your regression model are perfectly correlated (i.e., exhibit a perfect linear relationship) with other predictors. When perfect multicollinearity exists, it means one predictor can be expressed as an exact linear combination of other predictors, leading to a situation where the regression matrix becomes singular or non-invertible.

In practical terms, perfect multicollinearity occurs when:

- One variable is a constant multiple of another (e.g., X1 = 2 x X2).
- One variable is a mathematical combination of others (e.g., X1= X2 + X3).

When such perfect multicollinearity exists, the computation of VIF involves taking the inverse of a matrix that is not full rank, resulting in an infinite VIF value for the correlated predictor(s). This situation typically leads to issues in the regression analysis, rendering the estimates of coefficients unreliable or impossible to calculate due to the non-invertibility of the matrix.

To address infinite VIF values and multicollinearity issues:

- Identify and investigate the highly correlated predictors.
- Consider excluding one of the correlated variables from the model.
- Use domain knowledge to decide which variable(s) to retain or transform.
- Apply regularization techniques like ridge regression or LASSO to handle multicollinearity.

Resolving multicollinearity issues is crucial for obtaining stable and reliable estimates in regression analysis. Identifying the source of perfect multicollinearity and taking appropriate actions to mitigate it helps ensure the validity of the regression model.

5. **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 plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a given dataset follows a particular probability distribution. It visually compares the quantiles of the dataset against the quantiles of a theoretical distribution (such as the normal distribution) to determine if they align.

Construction of Q-Q Plot:

- Quantiles: Both the dataset and the theoretical distribution are sorted in ascending order.
- Plotting: The quantiles of the dataset are plotted against the quantiles of the theoretical distribution.
- Diagonal Line: If the data perfectly fits the theoretical distribution, the points on the plot will fall along a diagonal line (the 45-degree line).

Use and Importance in Linear Regression:

- Normality Assessment: In linear regression, Q-Q plots are used to check if the residuals (errors) follow a normal distribution. Normality of residuals is an important assumption in linear regression analysis.
- Identifying Departures from Normality: Deviations from the diagonal line in a Q-Q plot indicate departures from the assumed distribution. If the points deviate substantially from the diagonal, it suggests non-normality in the residuals.
- Model Assumption Checking: Q-Q plots help validate the assumption of normally distributed errors, which is essential for accurate estimation of regression coefficients, confidence intervals, and hypothesis tests.
- Residual Analysis: By examining the spread of points in the plot, Q-Q plots assist in identifying if the residuals exhibit skewness, heavy tails, or outliers.

Interpretation:

- Points along the diagonal line suggest a good fit between the data and the assumed distribution (e.g., normal distribution).
- Deviations from the line suggest departures from normality. Upward or downward bending indicates skewness or heavy tails, respectively.
- Points clustering at the extremes or outliers indicate potential issues with the residuals.

Q-Q plots are powerful visual tools used in linear regression to assess the assumption of normally distributed residuals. They provide insights into the distributional properties of residuals and assist in identifying departures from normality, helping analysts make informed decisions about the appropriateness of regression models and the validity of statistical inferences.