MULTIPLE LINEAR REGRESSION

-SHIVAM BAHUGUNA

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:

Impact of Categorical Variables on Bike Rentals:

Bike rental patterns are influenced by several categorical variables.

1. Season:
    - Majority of bike rentals occur during the fall season. The pleasant weather during fall—neither too hot nor too cold—makes it an ideal time for outdoor activities like biking.
2. Year:
    - In 2019, there was substantial growth in bike rentals. This suggests increasing interest in bike-sharing services over time.
3. Holiday:
    - Bike rentals spike during holidays. People have more leisure time during holidays, leading to increased bike usage for recreational purposes.
4. Weekday:
    - Fridays, Saturdays, and Thursdays witness higher total bike rentals compared to other days. These days likely correspond to weekends and provide more opportunities for bike rides.
5. Working Day:
    - On non-working days (such as weekends), bike rental counts are usually higher. People prefer leisurely activities when not occupied with work.
6. Weather Situation (Weathersit):
    - The largest number of bike rentals occurs when the weather is clear, with few clouds or partly cloudy conditions. Favourable weather encourages biking.
7. Month:
    - August and September stand out as the months with the highest bike rental activity. These months likely coincide with pleasant weather and outdoor events.

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

Answer:

When creating dummy variables from categorical features, the drop_first parameter plays a crucial role.

1. **Multicollinearity Prevention**:
    - Multicollinearity occurs when two or more independent variables are highly correlated. In the context of regression analysis, this can lead to unstable coefficient estimates and difficulties in interpreting model results.
    - By setting drop_first=True, we avoid the "dummy variable trap." This trap arises when we include all levels of a categorical variable as separate columns (including the reference level). Including the reference level can cause perfect multicollinearity, leading to unreliable model estimates.
    - Dropping the first level ensures that each dummy variable is orthogonal (uncorrelated) with the others, preventing multicollinearity.
2. **Interpretability**:
    - When interpreting regression coefficients, it's easier if we have a clear reference category. By dropping the first level, we establish a consistent reference point for comparison.
    - For example, if we create dummy variables for seasons (spring, summer, fall, winter), dropping the first level (spring) allows us to compare the other seasons to spring directly.
3. **Efficiency**:
    - Including unnecessary dummy variables (i.e., all levels) increases the dimensionality of the dataset. This can lead to overfitting and computational inefficiency.
    - By dropping the first level, we reduce the number of columns while retaining the necessary information.

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

Answer:

Based on the pair-plot analysis of numerical variables, the variable **'temp'** exhibits the highest correlation with the target variable **'total_count'**. The line plots depicting the relationship

between these two variables are remarkably similar, suggesting a strong linear association with bike rental counts.

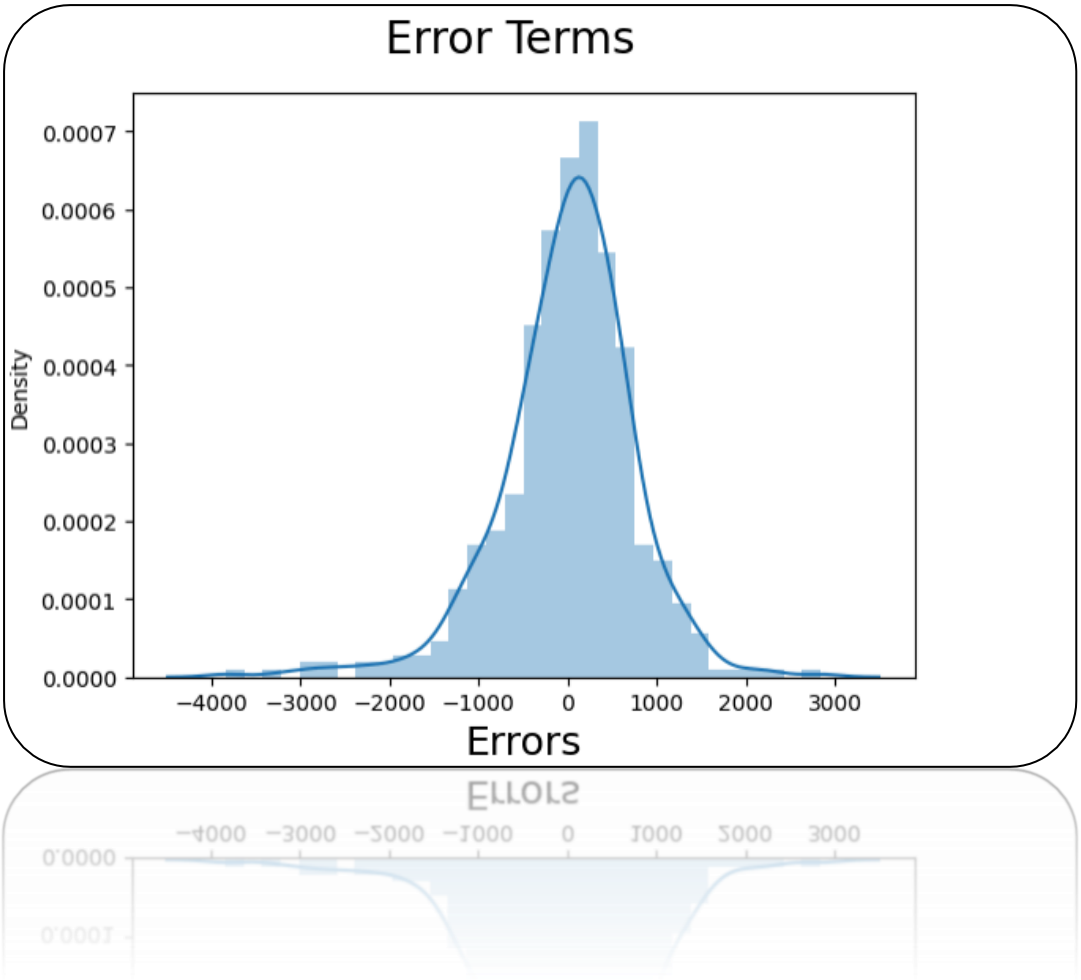
Questions 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

When validating assumptions after building a Linear Regression model on the training set, one critical factor to consider is the distribution of **error terms**. Here's how we assess this assumption:

Error Terms Should Be Normally Distributed with Mean 0:

- In Linear Regression, we assume that the error terms (residuals) follow a normal distribution with a mean of 0.
- To validate this assumption, we perform a **residual analysis** on the training set.
- Residuals are calculated as the difference between the actual y_train values and the predicted y_train values.
- By plotting a distribution (histogram or density plot) of the residuals, we can check if they exhibit a normal distribution centered around 0.
- Ideally, we want to see a bell-shaped curve with the mean of residuals close to 0.



Questions 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 final model, the top three features that significantly contribute to explaining the demand for shared bikes are:

1. **Atemp (Adjusted Temperature)**:
    - Atemp represents the adjusted feeling temperature, which combines actual temperature with humidity. It plays a crucial role in determining bike rental demand. As the perceived comfort level changes due to temperature, people are more likely to rent bikes when the adjusted temperature is favorable.
2. **Humidity**:
    - Humidity affects human comfort and outdoor activities. High humidity can make biking less appealing, while lower humidity levels encourage more bike rentals. Therefore, humidity is an essential factor in predicting bike demand.
3. **Wind Speed**:

- Wind speed impacts the riding experience. Strong winds can be discouraging for cyclists, leading to decreased bike rentals. Conversely, calm wind conditions promote bike usage. Wind speed is a key feature in understanding demand fluctuations.

General Subjective Questions

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

Answer:

Linear Regression:

Linear regression is a fundamental statistical technique used to model the relationship between a **dependent variable** (also known as the **response variable**) and one or more **independent variables** (also known as **predictors** or **features**). Here are the key points:

1. **Objective**:
   - The primary goal of linear regression is to predict the value of the dependent variable based on the given independent variables.
   - For instance, in predicting house prices, the independent variables could include square footage, number of bedrooms, and location, while the dependent variable would be the price.
2. **Assumption of Linearity**:
   - Linear regression assumes that there exists a linear relationship between the independent variables and the dependent variable.
   - This means that changes in the independent variables result in proportional changes in the dependent variable.
3. **Best-Fit Line**:
   - The algorithm calculates the **best-fit line** that minimizes the difference (residuals) between the predicted values and the actual values within the dataset.
   - The best-fit line represents the estimated relationship between the features and the target variable.
4. **Equation of the Line**:
   - The linear regression equation is typically expressed as: **[y = mx + b]**
     - (y) represents the dependent variable.
     - (x) represents the independent variable.

        - (m) is the slope of the line (how steep it is).
        - (b) is the intercept (the value of (y) when (x) is 0).
5. **Interpreting the Line**:
    - The slope ((m)) indicates how much the dependent variable changes for a unit change in the independent variable.
    - The intercept ((b)) represents the starting point of the line.
6. **Graphical Representation**:
    - When we plot the best-fit line alongside the data points, we visualize the relationship between the independent and dependent variables.
7. **Positive and Negative Regression Lines**:
    - There are two types of regression lines:
        - **Positive line of regression**: Slopes upward (positive correlation).
        - **Negative line of regression**: Slopes downward (negative correlation).

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

Answer:

- Anscombe's quartet is a group of four small datasets.
- These datasets share nearly identical **simple descriptive statistics** (like mean and variance) for both the (x) and (y) variables.
- However, when you plot these datasets on scatter plots, they look very different from one another.

Importance:

1. **Visualizing Data Before Modeling**:
    - Anscombe's quartet teaches us that before analyzing data or building models, we should **plot the data**.
    - Visualizing data helps us identify anomalies (like outliers) and understand its distribution.
2. **Regression Models and Deception**:
    - Even though the datasets have similar statistics, their scatter plots reveal peculiar behaviors.

- Linear regression models can be fooled by these differences.
- Let's look at the four datasets:
- **Data Set 1**: Fits a linear regression model well.
- **Data Set 2**: Nonlinear data; linear regression won't work here.
- **Data Set 3**: Contains outliers; linear regression struggles with outliers.
- **Data Set 4**: More outliers; linear regression still can't handle them.

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

Answer:

Pearson's Correlation Coefficient

- It's a statistical measure that quantifies the **linear relationship** between two variables.
- The value of (r) ranges from **-1.0 to +1.0**.
- It tells us how closely the data points follow a straight line (positive or negative) when plotted on a scatter plot.

- **Interpreting (r)**:
  - If (r) is close to **+1**, it indicates a **strong positive correlation**:
    - When one variable increases, the other tends to increase as well.
  - If (r) is close to **-1**, it indicates a **strong negative correlation**:
    - When one variable increases, the other tends to decrease.
  - If (r) is close to **0**, there's **no linear relationship**:
    - The variables don't move together in a predictable way.
- **Limitations of Pearson's (r)**:
  - It only captures **linear relationships**; nonlinear patterns are missed.
  - It treats both variables equally (doesn't differentiate between dependent and independent).
  - For more complex relationships, other correlation measures (like Spearman's rank correlation) may be better.

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

Answer:

1. **Scaling**:
    - **What Is Scaling?**
        - Scaling refers to adjusting the numerical feature values to a common range.
        - It ensures that all features have similar scales, which is crucial for many machine learning algorithms.
    - **Why Is Scaling Performed?**
        - Scaling is essential because some variables may have values on vastly different scales (e.g., one feature ranging from 0 to 100, while another ranges from 0 to 10,000).
        - Algorithms can be sensitive to these differences, leading them to prioritize features with larger scales.
        - Scaling also speeds up algorithm convergence.
    - **Benefits of Scaling**:
        - Improves model performance by making features comparable.
        - Helps maintain proportionality in predictions.
        - Commonly used in linear regression, k-nearest neighbors, and neural networks.
2. **Normalized Scaling (Min-Max Scaling)**:
    - **What Is It?**
        - Transforms data into a common range between 0 and 1.
        - Formula: [\text{Normalized Value} = \frac{{\text{Original Value} - \text{Min Value}}}{{\text{Max Value} - \text{Min Value}}}]
    - **Pros**:
        - Simple and intuitive.
        - Preserves the original distribution.
    - **Cons**:
        - Sensitive to outliers.
3. **Standardized Scaling (Z-Score Normalization)**:

- **What Is It?**
    - Scales data to have a mean of 0 and a standard deviation of 1.
    - Formula: [\text{Standardized Value} = \frac{{\text{Original Value} - \text{Mean}}}{{\text{Standard Deviation}}}]
- **Pros**:
    - Handles outliers better.
    - Suitable for algorithms that assume normally distributed data.
- **Cons**:
    - Alters the original distribution.

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

Answer:

Variance Inflation Factor (VIF):

- VIF measures the extent of **multicollinearity** (correlation) between predictor variables in regression models.
- When two or more predictors are highly correlated, the VIF increases.

Infinite VIF (Variance Inflation Factor):

- **Why Does This Happen?**
    - VIF measures multicollinearity (correlation) between predictor variables in regression models.
    - When two or more predictors are perfectly correlated (linearly dependent), the VIF becomes infinite.
    - Perfect multicollinearity means one predictor can be expressed as a linear combination of others.
    - For example, if you have two identical columns in your dataset, their VIF will be infinite.

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

Answer:

Q-Q Plot (Quantile-Quantile Plot):

- A Q-Q plot is a graphical tool used to assess whether two datasets come from populations with a common distribution.
- It helps us compare the quantiles of one dataset against the quantiles of another dataset.

How Does It Work?:

1. **Quantiles:**
    - A quantile represents the fraction (or percentage) of data points below a given value.
    - For example, the 0.3 (or 30%) quantile is the point where 30% of the data fall below, and 70% fall above that value.
2. **Creating a Q-Q Plot:**
    - We sort the data points in both datasets.
    - For each quantile in the first dataset, we find the corresponding quantile in the second dataset.
    - We plot these pairs of quantiles on a graph.
3. **Reference Line:**
    - A 45-degree reference line (also called the identity line) is plotted.
    - If the two datasets come from populations with the same distribution, the points on the Q-Q plot should approximately follow this reference line.
4. **Interpretation:**
    - Departure from the Reference Line:
        - If the points deviate significantly from the reference line, it suggests that the datasets have different distributions.
        - The greater the departure, the stronger the evidence for distinct distributions.
    - Nature of Differences:

        - Q-Q plots provide more insight into the nature of differences than analytical tests (e.g., chi-square or Kolmogorov-Smirnov tests).
        - They help us understand how the distributions diverge.

Importance of Q-Q Plots in Linear Regression:

- Assumption Checking:
    - In linear regression, we assume that the residuals (errors) follow a normal distribution.
    - Q-Q plots allow us to check if this assumption holds.
- Detecting Non-Normality:
    - If the residuals deviate significantly from the reference line, it indicates non-normality.
    - Non-normality affects the validity of statistical tests and confidence intervals.
- Guidance for Model Selection:
    - Q-Q plots help us choose appropriate regression models by revealing deviations from normality.
    - **If the residuals are not Gaussian, alternative models or transformations may be needed.**