Submitted By,

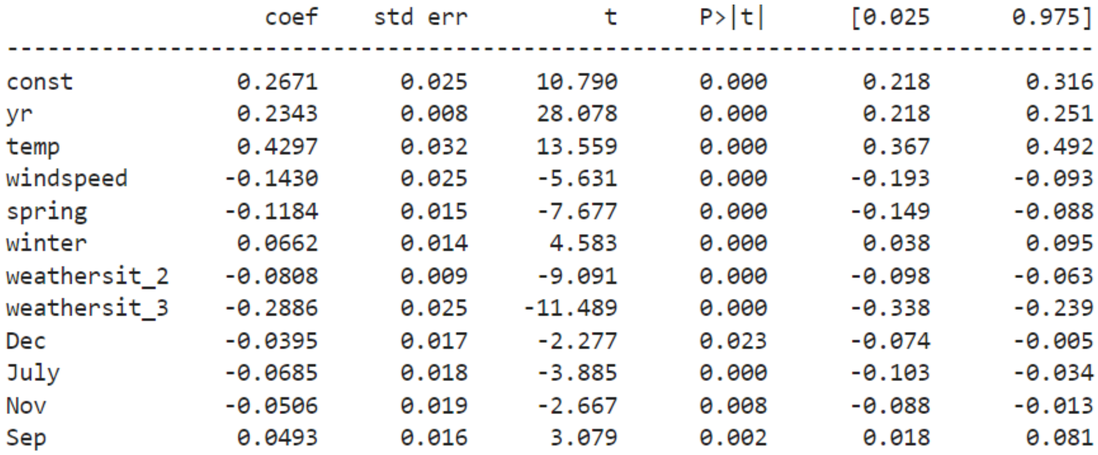
Midhun M Thomas

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:

   The final Multiple Linear Regression model comprises several predictor variables that are categorical, with certain variables transformed into dummy variables for encoding.



```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2671      0.025     10.790      0.000       0.218       0.316
yr             0.2343      0.008     28.078      0.000       0.218       0.251
temp           0.4297      0.032     13.559      0.000       0.367       0.492
windspeed     -0.1430      0.025     -5.631      0.000      -0.193      -0.093
spring        -0.1184      0.015     -7.677      0.000      -0.149      -0.088
winter         0.0662      0.014      4.583      0.000       0.038       0.095
weathersit_2  -0.0808      0.009     -9.091      0.000      -0.098      -0.063
weathersit_3  -0.2886      0.025    -11.489      0.000      -0.338      -0.239
Dec           -0.0395      0.017     -2.277      0.023      -0.074      -0.005
July          -0.0685      0.018     -3.885      0.000      -0.103      -0.034
Nov           -0.0506      0.019     -2.667      0.008      -0.088      -0.013
Sep            0.0493      0.016      3.079      0.002       0.018       0.081
```

   The variables spring and winter fall under the season category and have been dummy encoded. Similarly, weathersit_2 and weathersit_3 are part of the weathersit category, and the “Sep,Nov,July”variables fall under the mnth category, both of which have also been dummy encoded. The image above indicates that these variables are statistically significant and effectively explain the variance in the model.

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

   Answer:

   Reason 1:

   To prevent the dummy variable trap, which can cause multicollinearity among dummy variables, potentially violating the assumptions of linear regression. When encoding categorical variables with k levels, where k≥3, only k−1 levels are retained in the model. The omitted level is accounted for by the intercept as the reference category.

   Reason 2:

Consider a categorical variable representing colours, such as Red, Green, and Blue. As a nominal categorical variable, there is no inherent order or relationship among these colours. Simply label encoding them as 1, 2, and 3 would introduce an artificial order, leading to bias in the model (e.g., Red < Green < Blue). To avoid this issue, dummy encoding is used. Additionally, since the model cannot interpret string or text data directly, converting them into numerical form is essential.

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

Answer:

Before model building and training, the pair plot revealed that the variable "registered" exhibited the highest correlation with a value of 0.945. However, we excluded both "casual" and "registered" from the pre-processed training data to avoid model overfitting, as these variables directly sum to "cnt," which could inadvertently introduce information leakage. Among the remaining variables, "atemp" shows the strongest correlation with the target variable "cnt," followed closely by "temp.

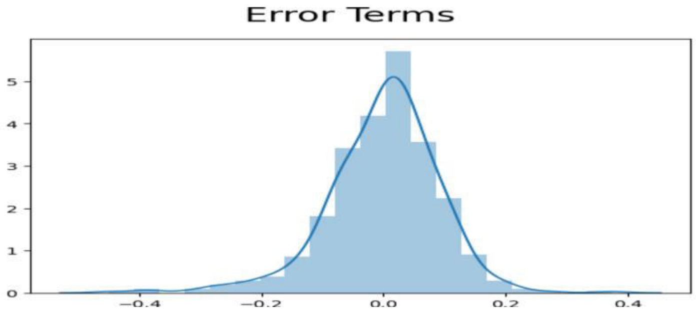
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

To ensure the model's assumptions are met and to validate the reliability of its inferences, the following procedures were undertaken:

Residual Analysis:

It is essential to confirm that the error terms are normally distributed, as this is a key assumption of linear regression. A histogram of the residuals was plotted, revealing that the residuals closely follow a normal distribution with a mean of 0, indicating that this assumption is satisfied.



Linear Relationship Between Predictor Variables and the Target Variable:

All predictor variables demonstrate a statistically significant relationship with the target variable, as evidenced by p-values below 0.05. The model's R-squared value on the training set is 0.831, with an adjusted R-squared value of 0.828, suggesting that the variance in the data is well-explained by the included predictor variables.

Independence of Error Terms:
The model ensures that the error terms are independent, as the predictor variables are uncorrelated with one another. Multicollinearity is not a concern, as the Variance Inflation Factor (VIF) for all predictor variables is below the threshold of 5.

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:

The top three features that significantly influence the demand for shared bikes are as follows:

1. Temperature (temp): With a coefficient of 0.4297, temperature emerges as the most influential factor, indicating a strong positive relationship with bike demand.

2. Year (yr): The year variable, with a coefficient of 0.2341, also plays a substantial role, reflecting the impact of temporal trends on bike-sharing demand.

3. September The month of September, represented by the dummy variable with a coefficient of 0.0493, contributes notably to the demand, highlighting seasonal variations in bike usage.

General Subjective Questions

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

Answer:
Linear Regression identifies the optimal linear relationship between independent and dependent variables, serving as a method for fitting the best straight line to a given dataset. Technically, it is a machine learning algorithm designed to establish the best linear-fit relationship between independent and dependent variables, typically using the Sum of Squared Residuals method.

The key assumptions underlying linear regression include:

a. Model Assumptions:
  - Linear Relationship: It is presumed that a linear relationship exists between the dependent and independent variables.

b. Assumptions Regarding Residuals:

1. Normality Assumption: The error terms ε(i) are assumed to be normally distributed.
2. Zero Mean Assumption: The residuals are expected to have a mean value of zero, indicating that the error terms are normally distributed around zero.
3. Constant Variance Assumption: The residual terms are assumed to exhibit constant variance (denoted as $\sigma 2$ - sigma square), a concept also known as homogeneity or homoscedasticity.
4. Independence Assumption: The residual terms are presumed to be independent of one another, with pairwise covariances equal to zero.

c. Assumptions About Estimators:
1. Measurement Accuracy: The independent variables are assumed to be measured without error.
2. Linear Independence: The independent variables are expected to be linearly independent, ensuring that multicollinearity is absent in the data.

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

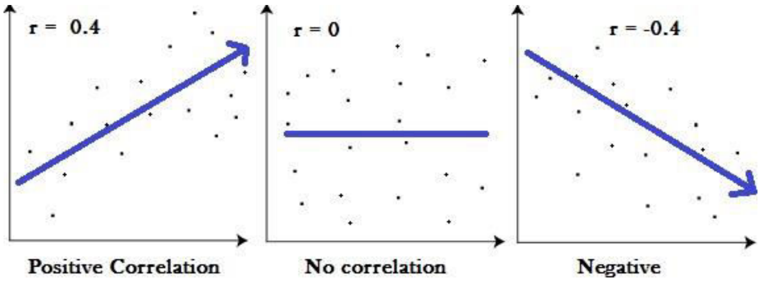
Answer:

Anscombe's quartet consists of four distinct datasets that share nearly identical descriptive statistics, yet exhibit markedly different distributions and appearances when visualized through graphs. Although the summary statistics of all four datasets are virtually the same, their graphical representations reveal significant disparities.

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

Answer:

Pearson's correlation coefficient, often denoted as Pearson's R, quantifies the degree of linear correlation between two datasets. It is calculated as the ratio of the covariance of the two variables to the product of their standard deviations, effectively normalizing the covariance. The resulting value always falls between -1 and 1, making it a standardized measure of linear association. Importantly, this metric is specific to linear relationships and does not capture other forms of association or correlation.



1. Correlation Coefficient of 1:

A value of 1 indicates a perfect positive linear relationship, where an increase in one variable corresponds to a proportional increase in the other. For example, shoe sizes typically increase in near-perfect correlation with foot length.

2. Correlation Coefficient of -1:

A value of -1 signifies a perfect negative linear relationship, where an increase in one variable corresponds to a proportional decrease in the other. For instance, the amount of gas in a tank decreases in almost perfect inverse correlation with speed.

3. Correlation Coefficient of 0:

A value of 0 suggests no linear relationship between the variables; changes in one variable do not predict changes in the other, indicating they are uncorrelated.

The magnitude of the correlation coefficient reflects the strength of the relationship between the variables. Larger absolute values indicate stronger linear relationships. For instance, a correlation coefficient of 0.95 (or -0.95) represents a stronger relationship than a coefficient of 0.55.

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

Answer:

Scaling is a technique employed to normalize the range of independent variables or features in a dataset. This process is crucial because raw data often exhibit widely varying ranges, which can affect the performance of some machine learning algorithms. For example, many classifiers use Euclidean distance to measure similarity between data points. If one feature has a significantly broader range than others, it will disproportionately influence the distance calculations. Thus, normalizing the range of all features ensures that each feature contributes approximately equally to the final distance computation.

Moreover, feature scaling can substantially improve the efficiency of gradient descent algorithms by ensuring faster convergence.

Normalization:

Also known as min-max scaling, normalization rescales feature values to a specified range, typically [0, 1]. The general formula for normalization is:

$$\mathbf{X' = x - min(X)/max(x) - min(x)}$$

where max(x) and min(x) are the maximum and minimum values of the feature, respectively.

Standardization:
Feature standardization transforms the feature values to have a mean of zero and a unit variance. This method is computed using the formula: where ¯x represents the mean of the feature values and σ is the standard deviation of the feature values.

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

Answer:
In the case of perfect multicollinearity, the Variance Inflation Factor (VIF) becomes infinite. This occurs when two independent variables exhibit a perfect correlation, leading to an \(R^2 \) value of 1 for the regression of one variable on the other. Consequently, the VIF formula \(\frac{1}{1 - R^2}\) results in an infinite value.

To address this issue, it is necessary to remove one of the variables contributing to the perfect multicollinearity from the dataset. An infinite VIF indicates that the variable in question can be expressed as an exact linear combination of other variables, which also exhibit infinite VIF values. This situation highlights the need for variable selection to resolve the problem of perfect multicollinearity.

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 Plots (Quantile-Quantile plots) are graphical tools used to compare the quantiles of two datasets. A quantile represents a fraction of the data falling below a specific value. For instance, the median is a quantile where 50% of the data points fall below and 50% fall above it. The Q-Q plot aims to determine if two datasets originate from the same distribution by plotting their quantiles against each other.

In a Q-Q plot, a 45-degree reference line is drawn. If the datasets are from the same distribution, the points should align closely with this line. This method is particularly useful in linear regression when verifying whether training and test datasets are drawn from populations with identical distributions.