Bike Sharing Linear Regression Subjective Questions Answers

Submitted By - Aashna Behl

Assignment-based Subjective Questions

Question 1:

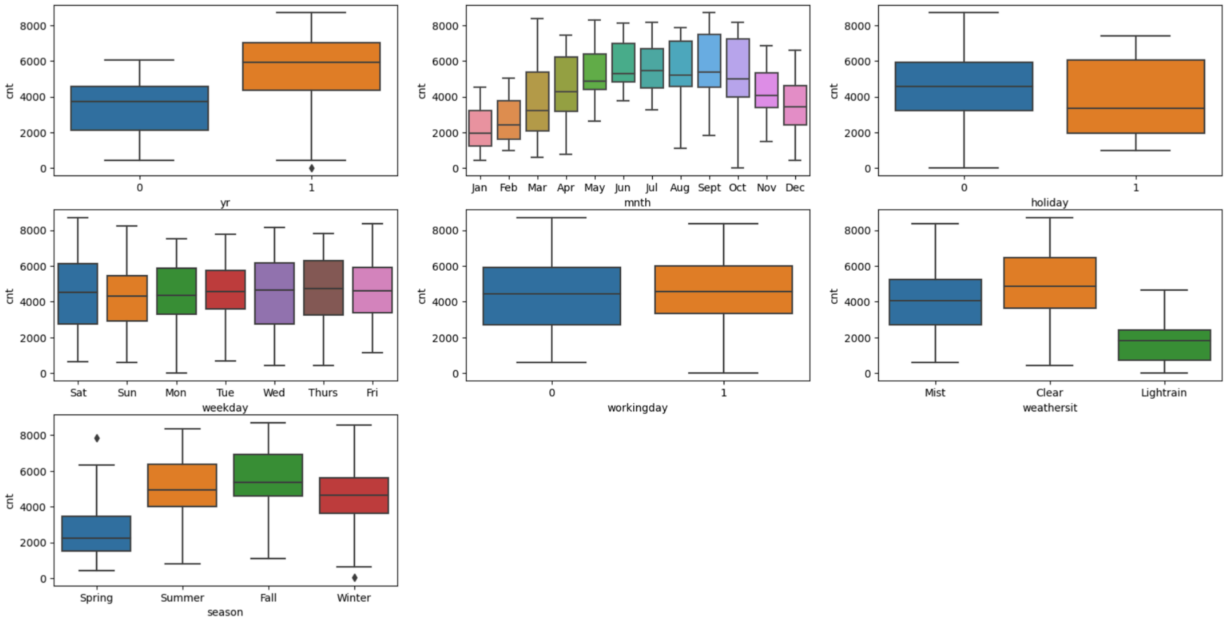
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Response:

The following are the categorical variables from the dataset;

1. Year (yr)
2. Month (mnth)
3. Holiday (holiday)
4. Weekday (weekday)
5. Working day (workingday)
6. Weather Situation (weathersit)
7. Season (season)

The visualisations of categorical against the target variable are as



Observations/ Inference

1. 2019 has had a higher booking count than 2018.
2. The demand increases from Jan to June steadily and then remain high till Oct before it starts to reduce. The demand is lowest in the winter months, the extreme weather conditions might be the cause of it.
3. Nothing conclusive for weekday or working day can be said.
4. Bookings are obviously more on a clear day, and the least on lightly rainy days, no rentals in extreme weather , high rain/ snow.
5. Fall and Summer season have relatively higher rentals than the rest of the seasons.

Question 2:

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

Response:

— When creating dummy variables from categorical variables dropping the first category is used to **avoid multicollinearity** amongst the variables.

If we do not drop first category it increases the chance that the one dummy variable will be highly correlated with another dummy variable. The dropped category can be interpreted from the rest of the dummy variables, it is redundant information in any case.

For eg. For a categorical variable with 3 distinct values (Yes, No , Maybe) we do not need 3 dummy variables. Just having 2 dummy variables for 'Yes' and 'No' will be enough to interpret 'Maybe'.

Yes	No	Maybe
1	0	0
0	1	0
0	0	1

We can easily interpret maybe for the above so we do not need an additional variable.

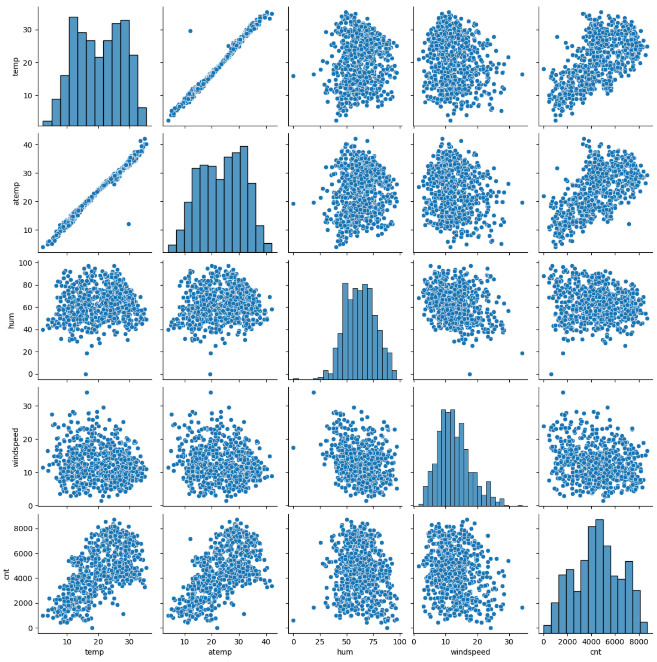
— Since the number of features reduces , it makes it easier to analyse features and increase efficiency of the model.

Therefore for a categorical variable with n values , n-1 dummy variables are required to be created.

Question 3:

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

Response:



As per the pair-plot temp and atemp has the highest correlation with the target variable 'cnt'.

Question 4:

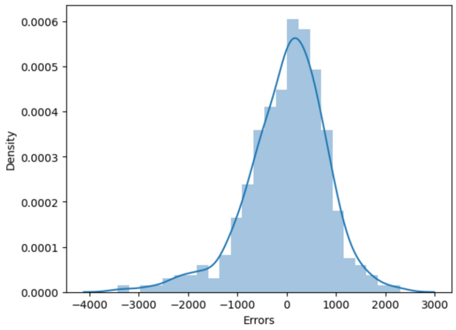
How did you validate the assumptions of Linear Regression after building the model on the training set?

Response:

The following assumptions were validated after the model build:

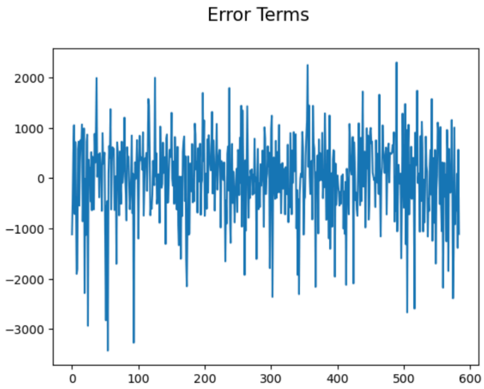
— Linearity : A linear relation has been established between the target variables are the features.

— Normally Distributed Residuals



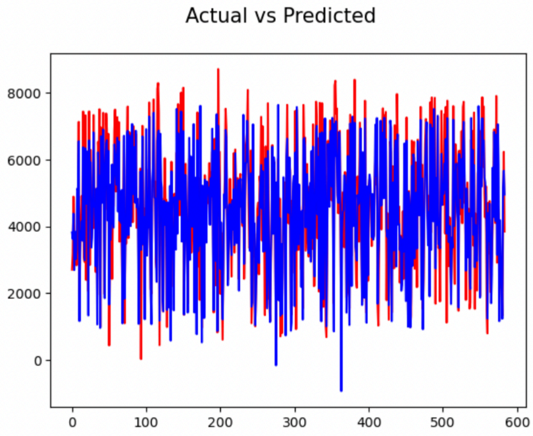
The residuals are normally distributed as visible in the graph.

— No pattern in errors



There is no specific patterns in the error terms

— Actual and Predicted values follow same patterns



Question 5:

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

Response:

Top 3 features

1. Temp
2. Yr
3. weathersit

General Subjective Questions

Question 1:

Explain the linear regression algorithm in detail.

Response:

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation.

The basic concept behind linear regression is to find the best-fitting straight line through the data points as per the mathematical relationship:

Simple Linear Equation (One independent feature)

$$y = mX + c$$

Y — dependent variable to be predicted
X — independent variable
c — Intercept or constant
m — Coefficient of independent variable

Multiple Linear Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

β0 - Intercept or constant
β1, β2…. βp — coefficients for each independent variable
X1,X2…..Xp — independent variables
ϵ — error term of residuals, which is difference between predicted and actual values of Y.

The goal is to find these coefficients is the Ordinary Least Squares (OLS) method, which minimises the sum of the residuals between the observed and predicted values.

The steps involved are:

Data Collection: Gather the data containing the dependent variable and independent variables.

Data Preprocessing (EDA): This step involves handling missing values, dealing with outliers, scaling or normalising variables, and splitting the data into training and testing sets.

Model Training: Use the training data to fit the linear regression model by estimating the coefficients (β's) that best fit the data.

Model Evaluation: Assess the performance of the model using the testing dataset. Common evaluation metrics include Mean Squared Error (MSE), R-squared, Root Mean Squared Error (RMSE), etc.

Prediction: Use the trained model to make predictions on new or unseen data.

Model Interpretation: Interpret the coefficients to understand the relationships between the dependent and independent variables.

There are some linear regression assumptions that need to be verified for a linear regression model
— Linearity
— Independence of errors
— Homoscedasticity
— Normality of Residuals
— No multicollinearity

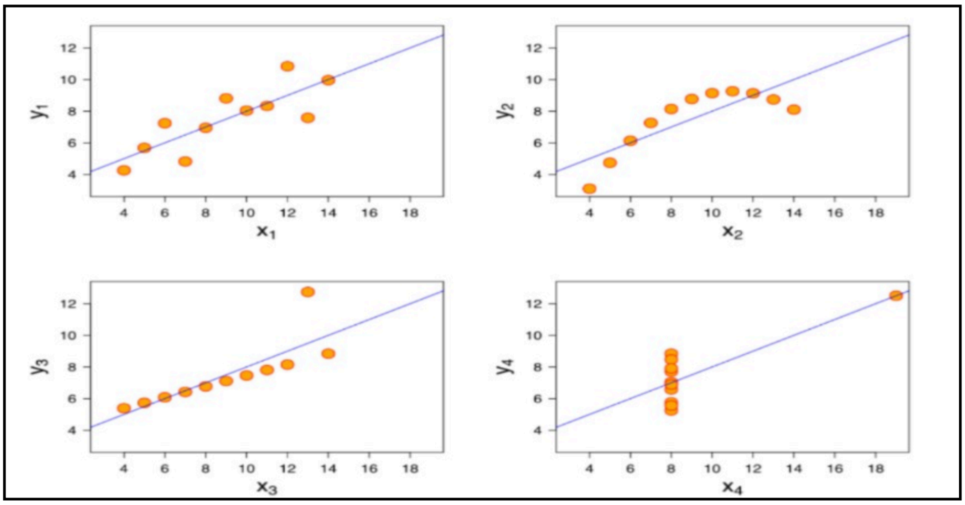
Question 2:

Explain the Anscombe’s quartet in detail.

Response:

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line parameters) but vastly different distributions and relationships when graphed. These datasets were created by the statistician Francis Anscombe in 1973 to emphasise the importance of visualising data and not relying solely on summary statistics.

The quartet comprises four sets of x-y data pairs, each containing 11 observations. When examined individually, these datasets display distinct characteristics:



1. The first dataset shows a linear relationship and fits a linear regression model.
2. The 2nd dataset does not have a linear relationship and hence does not fit the linear regression model.
3. Mostly a linear relationship but an outlier that can't be explained by the line.
4. No linear relationship expect one outlier that may have a high correlation coefficient.

Key takeaways from Anscombe's quartet:

Visualise your data: Visualisation is crucial to understanding data relationships. Simple summary statistics may not reveal the complete story.
Beware of outliers: Outliers can significantly influence summary statistics, regression lines, and correlations.
Consider different models: Different datasets might require different analytical approaches. Linear regression might not always be suitable for capturing relationships accurately.
Anscombe's quartet serves as a reminder to explore and analyse data visually alongside numerical summaries to gain a comprehensive understanding of the underlying patterns and relationships within the data.

Question 3:

What is Pearson's R?

Response:

Pearson's correlation coefficient, often 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 was developed by Karl Pearson in the late 19th century.

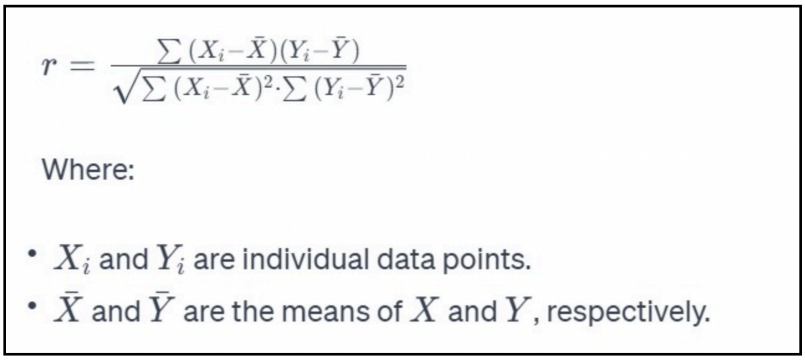
The Pearson correlation coefficient r ranges between -1 and 1:

r=1 indicates a perfect positive linear relationship between the variables. As one variable increases, the other variable increases proportionally.

r=–1 indicates a perfect negative linear relationship between the variables. As one variable increases, the other variable decreases proportionally.

r=0 indicates no linear relationship between the variables.

The formula for calculating Pearson's correlation coefficient between two variables X and Y with n data points is:



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

Where:

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

Key points about Pearson's correlation coefficient:

Measures Linear Relationship: Pearson's r quantifies only linear relationships between variables. It assumes a straight-line relationship between the variables.

Sensitivity to Outliers: Pearson's r can be sensitive to outliers, as outliers can disproportionately affect the mean and standard deviation, impacting the correlation coefficient.

Assumes Normality: While Pearson's r does not require strict normality in the data, it performs best when the variables follow a roughly normal distribution.

Does Not Imply Causation: A high correlation does not imply causation. It indicates association or relationship, but it doesn't determine a cause-and-effect relationship between variables.

Pearson's correlation coefficient is widely used in various fields such as statistics, economics, social sciences, and more, to assess the strength and direction of relationships between continuous variables. However, it's essential to consider other factors, perform visual inspections, and analyse the context of the data before drawing conclusions based solely on correlation values.

Question 4:

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

Response:

Scaling is transforming the numerical or continuous variables so they fit within a specific scale.

Advantages:

Improves Model Performance: Many machine learning algorithms, such as gradient descent-based methods (e.g., linear regression, logistic regression) and distance-based algorithms (e.g., K-nearest neighbours, support vector machines), perform better when features are on a similar scale. Scaling helps algorithms converge faster and prevents features with larger scales from having a disproportionate impact on the model.

Equalizes Variable Impact: Without scaling, variables with larger magnitudes or ranges can dominate the learning process or influence the outcome more than others. Scaling ensures that each feature contributes proportionally to the analysis or modeling.

Normalised Vs Standardised Scaling

1. Min Max scaling in Normalised Scaling vs Mean and Standard Deviation used for Standardized scaling.
2. Normalised scaling used when features are of different scales vs Standardized scaling is used when we want to ensure 0 mean and unit Std Deviation
3. Normalised scales between (0,1) and (-1, 1) , while standardization transforms features to have a mean of 0 and a standard deviation of 1
4. Standardization is less affected by outliers compared to normalization.

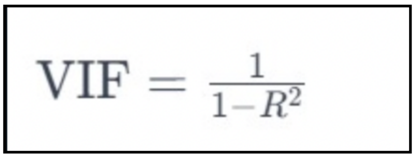
The choice between normalization and standardization depends on the specific requirements of the problem, the characteristics of the data, and the algorithms being used for analysis or modeling.

Question 5:

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

Response:

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when predictor variables in a regression model are highly correlated with each other. VIF measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity.



$$\mathrm{VIF} = \frac{1}{1-R^2}$$

R2 is the coefficient of determination obtained by regressing the predictor variable in question against the other predictor variables.

If VIF is infinite (or extremely high), it indicates an extremely high degree of multicollinearity. This is when one or more predictor variables can be perfectly predicted from a linear combination of other predictors.

Perfect multicollinearity causes numerical issues in regression analysis because it becomes impossible to obtain unique estimates for the coefficients of the correlated variables. In such cases, the standard errors of the coefficients become extremely large or infinite, making it impossible to obtain reliable estimates.

Perfect multicollinearity might occur due to various reasons such as including redundant variables in the model, creating dummy variables without dropping a reference category, or mathematical manipulations that lead to perfect linear relationships among predictor variables.

Question 6:

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

Response:

A Q-Q plot, or a quantile-quantile plot, is a graphical method used to assess if a dataset follows a particular probability distribution.
It compares the quantiles of the dataset against the quantiles of a theoretical distribution (e.g., normal distribution), allowing us to visually inspect whether the data deviates from the expected distribution.

Identification of Departure from Normality: Q-Q plots help identify departures from normality. If the points on the Q-Q plot deviate significantly from the diagonal line, it indicates a departure from the assumed distribution.

Model Reliability Assessment: Assessing the normality of residuals is crucial for the reliability of the linear regression model. Departure from normality might affect the accuracy of the coefficient estimates, the validity of statistical tests, and confidence intervals.

Decision-Making for Model Improvement: If the Q-Q plot indicates non-normality in the residuals, it suggests that the model assumptions are violated. In such cases, corrective actions might be needed, such as transforming the variables, using robust regression techniques, or considering alternative models.

Interpretation of Q-Q Plot:
Points on the Diagonal Line: If the points in the Q-Q plot fall closely along the diagonal line, it suggests that the data approximately follows the assumed distribution (e.g., normal distribution).

Deviation from the Diagonal Line: Significant deviations from the diagonal line indicate departures from the assumed distribution. Skewedness, heavy tails, or other non-normal patterns might be observed, indicating potential issues with the normality assumption.

Q-Q plots serve as a valuable diagnostic tool to assess the normality assumption and other distributional assumptions in linear regression. They provide visual insights into the distributional characteristics of residuals, guiding researchers or analysts in making informed decisions about the model's validity and potential improvements.