LINEAR REGRESSION
SUBJECTIVE QUESTIONS

Deepak Mohanta

SUMMARY SECTION

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

GENERAL SUBJECTIVE QUESTIONS

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOU INFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE?

Regression Model

Demand = 0.374
+ 0.9 X **Casual**
+ 0.03 X **mnth_Aug**
- 0.079 X **mnth_Feb**
- 0.1085 X **mnth_Jan**
- 0.0768 X **mnth_Mar**
+ 0.0203 X **mnth_Nov**
+ 0.0414 X **mnth_Sep**
- 0.2477 X **weekday_Mon**
- 0.2328 X **weekday_Sun**
- 0.1425 X **weathersit_Light Snow**

- Categorical variables have a low to medium impact on the regression model. Though 9 out of 10 independent variables are categorical in our linear regression model, the highest weightage is associated with Casual variable which is not a categorical variable.
- Sunday and Monday have higher impact on the demand as compared to other categorical variables such as different months and weather_light_Snow

WHY IS IT IMPORTANT TO USE DROP_FIRST=TRUE DURING DUMMY VARIABLE CREATION?

Dummy Variable Trap: Without dropping one dummy variable, perfect multicollinearity can occur.
Redundancy Issue: Information about the category can be redundant if all dummy variables are included.
Solution: Use drop_first=True during dummy variable creation.
Benefits: Prevents multicollinearity, ensures model stability, and simplifies interpretation of regression coefficients.

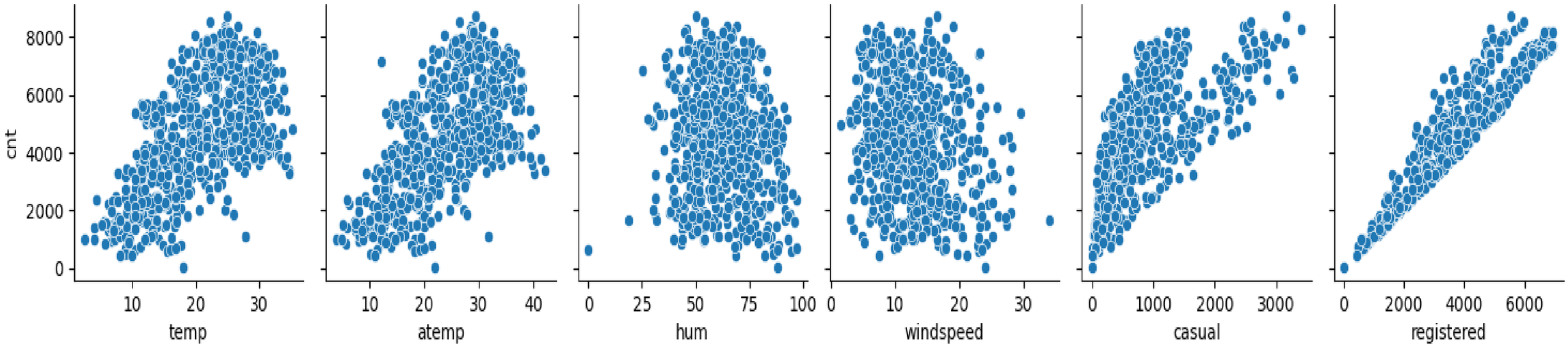
When creating dummy variables, `drop_first=True` is used to avoid the "dummy variable trap." The dummy variable trap occurs when there is a perfect multicollinearity among the dummy variables, meaning one dummy variable can be perfectly predicted from the others.

Here's why the dummy variable trap happens: Suppose you have a categorical variable with two categories (0 and 1), and you create two dummy variables (D1 and D2) to represent these categories. If D1 is 1, it implies the category is 0, and if D2 is 1, it implies the category is 1. So, if D1=0 and D2=0, it means the category is neither 0 nor 1. In other words, the information about the category is redundant.

By setting `drop_first=True`, you exclude one of the dummy variables. This eliminates the perfect multicollinearity issue because the information about the category is still captured by the remaining dummy variable(s). For a binary categorical variable, you only need one dummy variable to represent the two categories effectively.

Not dropping the first dummy variable may lead to the dummy variable trap, causing issues in regression models due to multicollinearity. Multicollinearity can make it challenging to interpret the coefficients of the regression model and can lead to unstable estimates. Therefore, dropping one dummy variable helps to avoid these problems and ensures that the model is well-behaved.

LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?



Looking at the pair-plot among the numerical variables,
- We can conclude that the 'registered' has the highest correlation with the target variable

HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

I used the following methods to validate the assumptions of Linear Regression after building the model

- VIF: Checking the VIF values and ensuring that VIF<5
- Residual analysis: Validated whether the model is overall normally distributed or not

BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

Regression Model

Demand = 0.374
+ 0.9 X **Casual**
+ 0.03 X **mnth_Aug**
- 0.079 X **mnth_Feb**
- 0.1085 X **mnth_Jan**
- 0.0768 X **mnth_Mar**
+ 0.0203 X **mnth_Nov**
+ 0.0414 X **mnth_Sep**
- 0.2477 X **weekday_Mon**
- 0.2328 X **weekday_Sun**
- 0.1425 X **weathersit_Light Snow**

The top three features contributing to demand

1. Casual
2. Weekday_Monday
3. Weekday_Sunday

GENERAL SUBJECTIVE QUESTIONS

EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (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 linear equation that predicts the dependent variable based on the independent variables.

Here are the key components and steps involved in the linear regression algorithm:

1. **Assumptions**:
   - **Linearity**: Assumes a linear relationship between independent and dependent variables.
   - **Independence**: Assumes that the residuals (the differences between actual and predicted values) are independent.
   - **Homoscedasticity**: Assumes constant variance of residuals across all levels of independent variables.
   - **Normality of Residuals**: Assumes that the residuals are normally distributed.

EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

2. Simple Linear Regression:

Equation: In simple linear regression with one independent variable, the equation is $y=mx+b$, where y is the dependent variable, x is the independent variable, m is the slope, and b is the y-intercept.

3. Multiple Linear Regression:

Equation: In multiple linear regression with more than one independent variable, the equation is $Y=b0+b1X1+b2X2...$ where b0,b1,b3.. Are the coefficients and X1,X2.. Are the independent variables

4. Objective Function (Cost Function):

Ordinary Least Squares (OLS): The most common method to estimate the coefficients is to minimize the sum of squared differences between the actual and predicted values.

5. Coefficient Estimation:

Slope and Intercept: Coefficients are estimated using methods like OLS. The goal is to minimize the sum of squared residuals.

6. Model Evaluation:

- **R-squared (R^2):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared:** Similar to R-squared but penalizes the inclusion of irrelevant variables.

EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

7. Predictions: Once the model is trained, predictions for the dependent variable can be made using the coefficients and new values for the independent variables.

8. Residual Analysis: Examine residuals to check if the assumptions hold. Residuals should be normally distributed, show no pattern, and have constant variance.

9. Outliers and Influential Points: Identify and handle outliers and influential points that may affect the model.

10. Regularization (Optional): Regularization techniques like Ridge or Lasso regression can be applied to prevent overfitting and handle multicollinearity.

11. Implementation: Linear regression can be implemented using various libraries in Python (e.g., scikit-learn, statsmodels) or other programming languages.

12. Validation and Testing:

Evaluate the model's performance on validation and test datasets to ensure it generalizes well to new data.

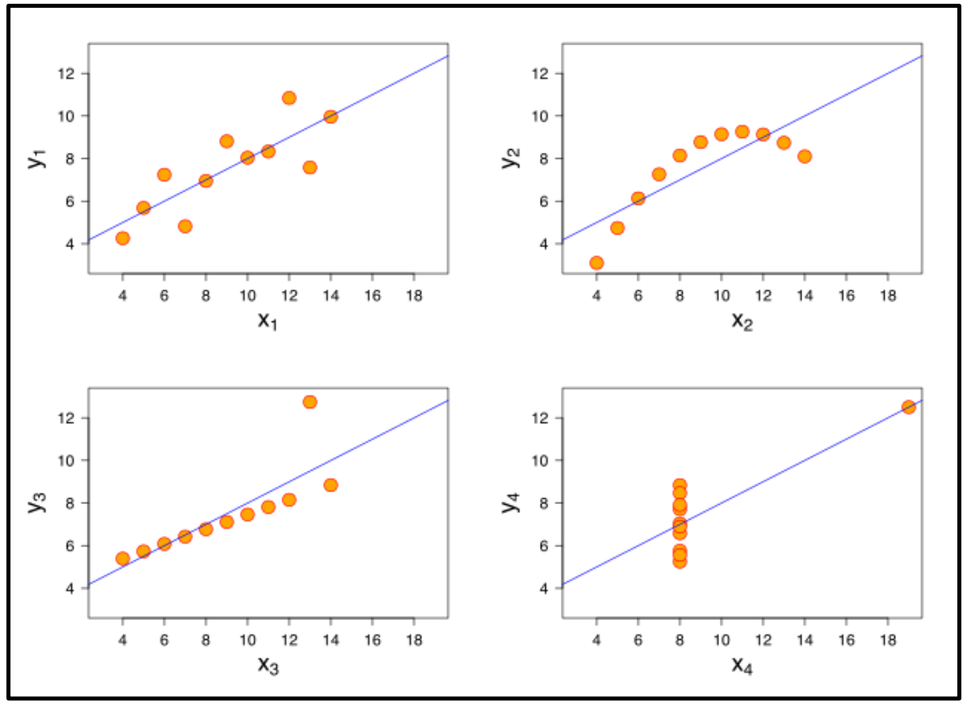
Linear regression is a powerful and interpretable algorithm commonly used for tasks like predicting house prices, sales, and various other continuous outcomes. However, it makes strict assumptions that should be checked and addressed to ensure the model's reliability.

EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but vary widely when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet consists of four datasets, each containing 11 (x, y) pairs:

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places



Key Points:

- Despite having identical summary statistics, the datasets exhibit different patterns when visualized.
- The quartet illustrates the importance of graphical exploration in understanding data.
- It highlights that summary statistics alone may not capture the complexity of the data.
- Different datasets with the same summary statistics can lead to different interpretations and conclusions.

WHAT IS PEARSON'S R?

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is named after Karl Pearson, who introduced the concept.

The correlation coefficient

- r ranges from -1 to 1
  - r=1: Perfect positive linear correlation
  - r=−1: Perfect negative linear correlation
  - r=0: No linear correlation
- The sign of r indicates the direction of the relationship:
  - Positive r: Indicates a positive linear relationship (as one variable increases, the other tends to increase).
  - Negative r: Indicates a negative linear relationship (as one variable increases, the other tends to decrease).

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

It's important to note that Pearson's correlation coefficient measures only linear relationships and may not accurately capture non-linear associations. Additionally, correlation does not imply causation; a high correlation between two variables does not necessarily mean that changes in one variable cause changes in the other.

WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

Scaling:

Scaling is a preprocessing step in data analysis and machine learning that involves transforming the values of variables to a specific range or distribution. The primary goal of scaling is to bring all features or variables to a similar scale, ensuring that no variable dominates others in terms of magnitude. This is particularly important in algorithms that rely on distance measures or gradients, such as k-nearest neighbors, support vector machines, and gradient-based optimization algorithms used in many machine learning models.

Why Scaling is Performed:

- **Equal Weightage**: Scaling ensures that all variables contribute equally to the analysis, preventing variables with larger magnitudes from overshadowing others.
- **Algorithm Sensitivity:** Some machine learning algorithms are sensitive to the scale of input features. Scaling helps improve the performance and convergence of these algorithms.
- **Distance-based Algorithms**: In algorithms like k-nearest neighbors or k-means clustering, where distances between data points matter, scaling ensures that distances are calculated appropriately.

NORMALIZED SCALING VS. STANDARDIZED SCALING:

Normalized Scaling (Min-Max Scaling):

- **Formula**: Xnormalized = (X - Xmin) / (Xmax - Xmin)
- **Range**: Typically scales data between 0 and 1.
- **Advantages**: Simple and intuitive. Preserves the relative differences between data points.
- **Drawbacks**: Sensitive to outliers. The presence of outliers can disproportionately impact the scaling.

Standardized Scaling (Z-score Normalization or Z-score Scaling):

- **Formula**: Xstandardized = (X-mu) / (sigma)
- **Range**: Centers data around 0 with a standard deviation of 1.
- **Advantages**: Less sensitive to outliers. Works well when the data distribution is not necessarily uniform.
- **Drawbacks**: May not preserve the original distribution and relative differences as effectively as normalized scaling.

Key Differences:

- **Range**: Normalized scaling typically scales data between 0 and 1, while standardized scaling centers data around 0 with a standard deviation of 1.
- **Sensitivity to Outliers**: Normalized scaling can be sensitive to outliers, while standardized scaling is more robust in the presence of outliers.
- **Preservation of Original Distribution**: Normalized scaling tends to preserve the original distribution better than standardized scaling.
- **Use Cases**: Normalized scaling is often suitable when the data distribution is relatively uniform and outliers are not a major concern. Standardized scaling is preferred when the data distribution is not necessarily uniform and there is a need for robustness against outliers.

YOU MIGHT HAVE OBSERVED THAT SOMETIMES THE VALUE OF VIF IS INFINITE. WHY DOES THIS HAPPEN?

$$VIF_i = \frac{1}{1 - R_i^2}$$

- The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity between predictor variables.
- Perfect multicollinearity occurs when one predictor variable can be precisely predicted by a linear combination of other predictor variables.
- In such cases, the (1 - R_sqr) term in the VIF formula approaches 1, resulting in a denominator close to zero and an infinite VIF.
- This situation requires addressing highly correlated predictor variables through techniques like removing one of the variables, combining them, or using dimensionality reduction methods like PCA.
- Handling multicollinearity is crucial to ensure stable regression coefficients and accurate model interpretation.

WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in linear regression to assess the normality of residuals. It compares the quantiles of observed residuals to those expected in a normal distribution.

Use and Importance:

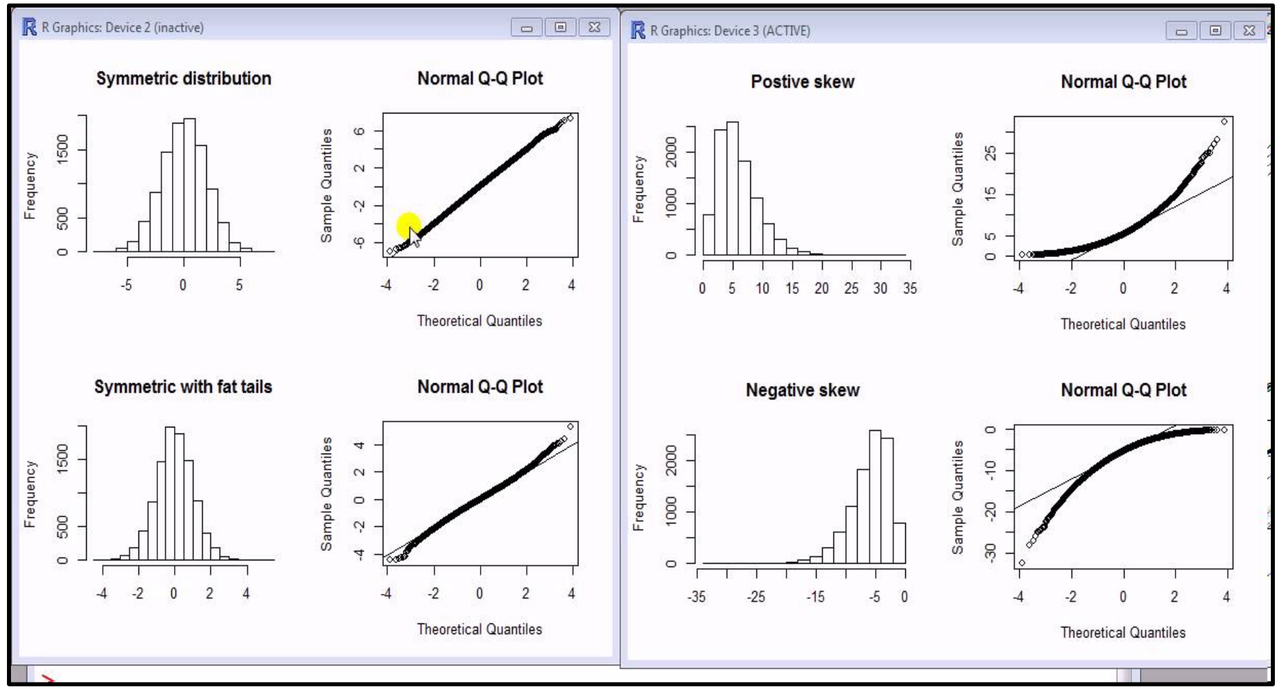
- Checks the assumption of normality in linear regression residuals.
- Identifies skewness, heavy tails, or outliers.
- Deviations from a straight line suggest departures from normality, guiding model diagnostics and refinement.

Interpretation:

- Straight line indicates approximately normal residuals.
- Deviations highlight potential issues that need attention.

Key Points:

- Ensures validity of statistical inferences in linear regression.
- Aids in identifying and addressing issues affecting model reliability.



THANK YOU