# LINEAR REGRESSION SUBJECTIVE QUESTIONS - BOOM BIKE ASSIGNMENT

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

Ans. Categorical variables are **Seasons, Working dats, Weather,Weekdays, Year, Holidays and Months.**

**a. Seasons:** - Biking is most popular during winter and summer , suggesting higher promotional efforts during these periods. Rainy season has negative impact on the business.

**b.Working Days** - Understanding user behavior on different days can inform targeted strategies for increasing bike rentals.

**c**.**Weekdays-** Users specially like to rent bike on Saturday more than other days. Although, no clear pattern emerges when considering overall bike usage, but registered users tend to ride more on working days, while casual users ride more on non-working days.

**d. Yearly Trends** - Bike rentals have increased from 2018 to 2019, based on available data.

**e.Holidays** - Casual users rent bikes more on holidays compared to registered users.

**f.Months** - June, July, August, September, and October see higher bike rental ratios.

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

Using `drop_first=True` during dummy variable creation is important to prevent multicollinearity and avoid the dummy variable trap.

When creating dummy variables from categorical features, each category is represented by a binary column. Without dropping one of these binary columns, we risk multicollinearity, where one dummy variable becomes perfectly predictable from the others. This results in the model becoming unstable and unreliable.

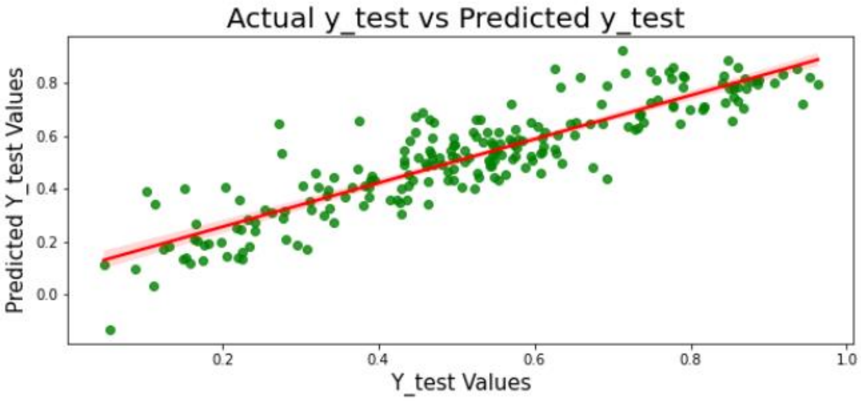
By dropping the first dummy variable, we effectively remove redundancy while still capturing all the necessary information about the categories. This ensures that the model doesn't mistakenly infer false relationships between the features and the target variable. Additionally, dropping one dummy variable aligns with the reference category concept, making the interpretation of coefficients more straightforward. In summary, `drop_first=True` enhances the performance and interpretability of the model by mitigating multicollinearity and preventing the dummy variable trap.

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

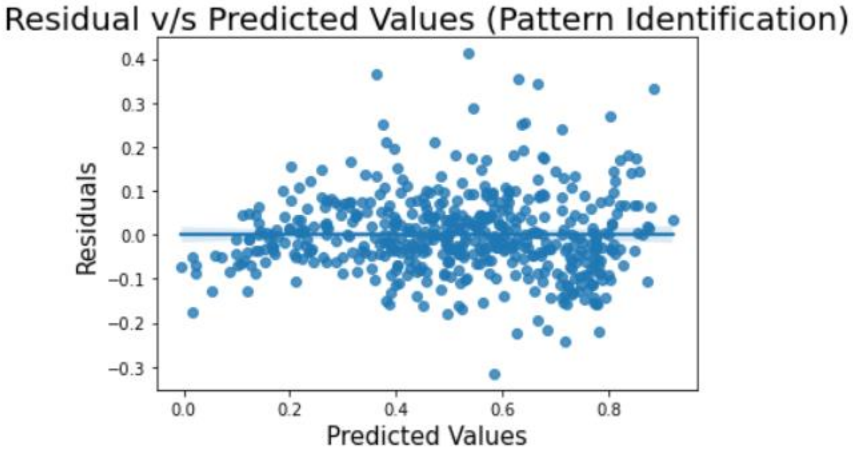
“temp” and “atemp” variables are the numerical variables with highest correlation with target variable “cnt”.

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

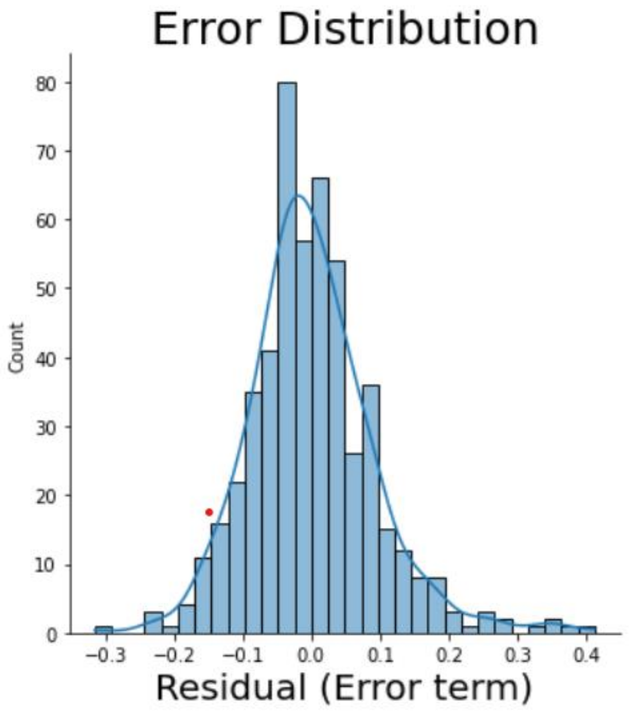
1. Linear Relationship between independent and dependent variables : Symmetrical distribution around the diagonal line in the actual vs predicted plot confirms the linear relationship between independent and dependent variables.



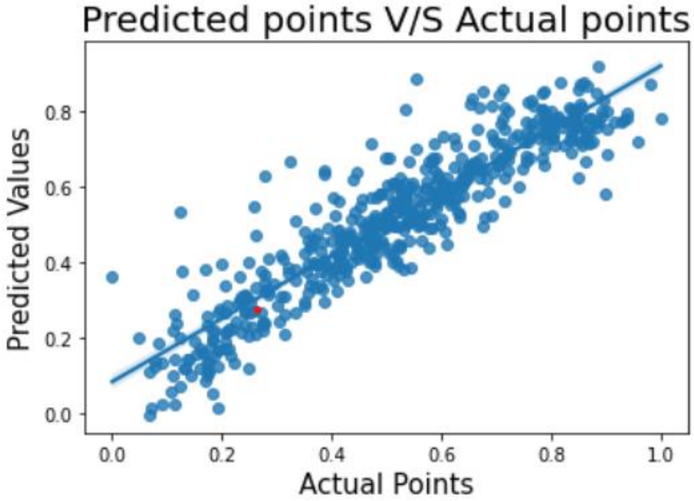
2. Error Terms are independent of each other : Lack of discernible pattern in the Error Terms relative to predictions indicates their independence from each other.



3. Error terms are  Normal Distribution  : Histogram and distribution plot illustrate the normal distribution of error terms with a mean of 0.



4. Error terms have Constant Variance: Consistent variance in Error Terms suggests adherence to the assumption of homoscedasticity.



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

The top 3 variables are:

**weathersit** : Temperature is the Most Significant Feature which affects the Business positively, Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively.

‘**yr’** : The growth year on year seems organic given the geological attributes.

‘**season’** : Winter season is playing the crucial role in the demand of shared bikes.

---

## General Subjective Questions

1. **Explain the linear regression algorithm in detail?**

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.
  - Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.
  - $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$ o $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta 0$, $\beta 1$, $\beta 2$, …, $\beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches

  – Unconstrained and constrained.

  - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
    - The straight-line equation is $Y = \beta 0 + \beta 1X$
    - The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi.

▪ $Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0\ ) = \sum(yi - \beta 1xi - \beta 0\ )\ 2$

o The unconstrained minimization are solved using 2 methods

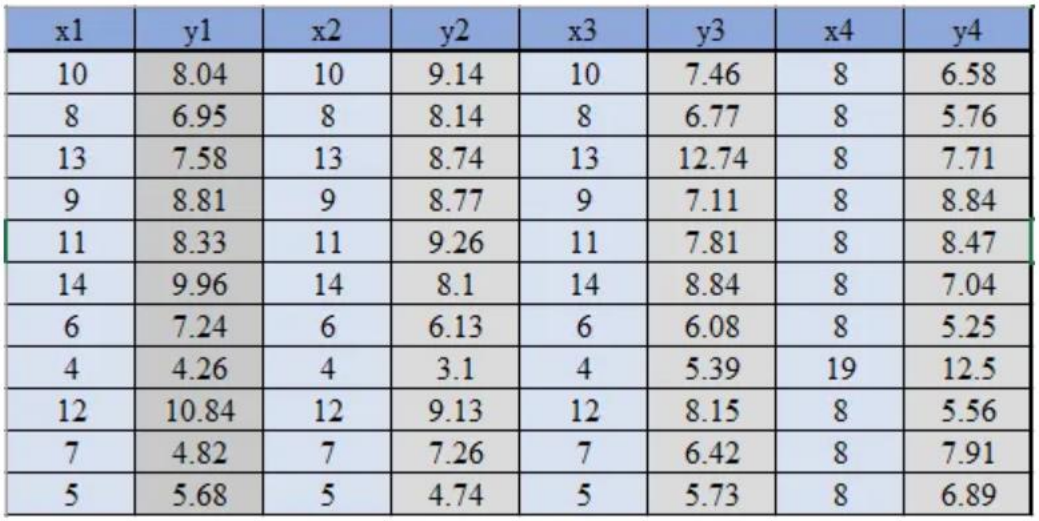
▪ Closed form ▪ Gradient descent

• While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used. o $ei = yi - ypred$ is provides the error for each of the data point. o OLS is used to minimize the total e2 which is called as Residual sum of squares. o RSS = = $\sum\ (yi - ypred)\ 2\ n\ i$=1

• Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. **Explain the Anscombe's quartet in detail.**

***Anscombe's Quartet*** is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



| x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

Apply the statistical formula on the above data-set, Average Value of x = 9

Average Value of y = 7.50

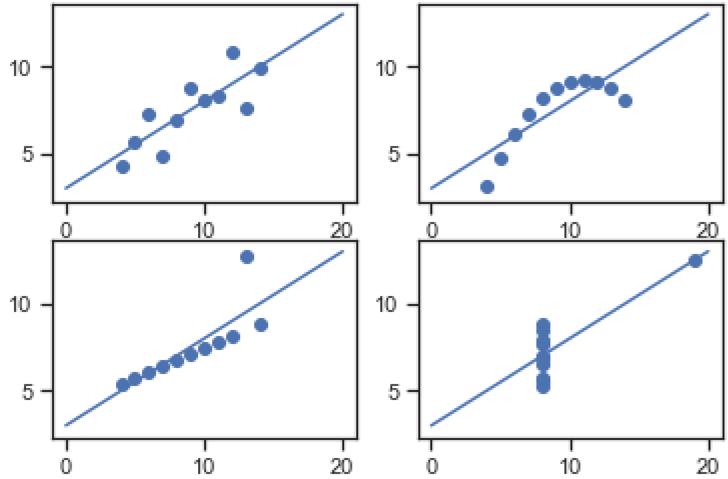
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



*Graphical Representation of Anscombe's Quartet*

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).

- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.

- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

3. **What is Pearson's R?**

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$r = n(\Sigma x * y) - (\Sigma x) * (\Sigma y) \surd[n\Sigma x\ 2 - (\Sigma x)\ 2] * [n\Sigma y\ 2 - (\Sigma y)\ 2]$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y score

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

Scaling is the process of transforming numerical features of a dataset to a standard range or distribution. It's done to ensure that all features have similar scales and magnitudes, which can improve the performance and stability of machine learning models. Here's why scaling is performed and the difference between normalized scaling and standardized scaling:

1. **Why Scaling is Performed**:
    - **Equal Weight**: Scaling ensures that all features contribute equally to the analysis, preventing features with larger scales from dominating the learning process.
    - **Faster Convergence**: Many machine learning algorithms, like gradient descent, converge faster when features are on similar scales.
    - **Regularization**: Some regularization techniques, like L1 and L2 regularization, assume that features are centered around zero and have similar scales.
2. **Normalized Scaling**:
    - **Range**: Normalized scaling (or min-max scaling) scales the features to a range between 0 and 1.
    - **Formula**: The formula for normalized scaling is:

    $$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

    **Advantage**: Preserves the shape of the original distribution while scaling the values.
3. **Standardized Scaling**:
    1. **Mean and Standard Deviation**: Standardized scaling (or z-score scaling) scales the features to have a mean of 0 and a standard deviation of 1.
    2. **Formula**: The formula for standardized scaling is:

    $$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

**Advantage**: Makes the distribution of each feature have a mean of 0 and a standard deviation of 1, which can be helpful when features have different units or distributions.

---

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

The VIF (Variance Inflation Factor) measures how much the variance of the estimated regression coefficients increases due to multicollinearity. A VIF value of infinity occurs when there is perfect multicollinearity between one or more predictor variables.

Perfect multicollinearity means that one or more independent variables can be exactly predicted from other independent variables in the model. In other words, there is a linear relationship between the predictor variables. When this happens, the regression coefficients cannot be estimated because there are infinitely many solutions that perfectly fit the data.

Perfect multicollinearity can arise due to various reasons, including:

1. Linear Dependency : One variable is a linear combination of other variables in the model.
2. Dummy Variable TrapWhen creating dummy variables from categorical variables, perfect multicollinearity can occur if one dummy variable can be exactly predicted from other dummy variables.
3. Data ErrorsData errors, duplicates, or highly correlated variables can lead to perfect multicollinearity.

To address this issue, it's essential to identify the source of multicollinearity and take appropriate actions, such as removing redundant variables, combining variables, or using dimensionality reduction techniques like PCA (Principal Component Analysis). Handling multicollinearity helps ensure the stability and reliability of the regression model.

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

Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- Interpretations
  - o Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
  - o Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - o X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - o Different distributions – If all the data points are lying away from the straight line.

- Advantages
  - o Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
  - o The plot has a provision to mention the sample size as well