Linear Regression Assignment

By: Jayshree Sahoo

Assignment based Subjective Question:

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

Answer:

I have used both bar and boxplot for the analysis of the effect of categorical variables on the dependent variable (target variable) are as follows:

- During fall season the demand increases and there is significant decrease in winter season.
- There was a significant increase in demand of rental bike in year 2019 (i.e. in year column 0 represents 2018 and 1 represents 2019).
- The demand increases continuously with each month until June.
- June and September month has highest demand. After September, there is a continuous decrease in demands. In addition, we can see that the demand decreases in end and begging of the year.
- If there is a holiday, demand decreases.
- Working day does not show any significant change. Demand for rental bike is same in weekdays and weekends.
- Demand increases if the Weather is pretty clear (i.e. A).

Ques 2: Why is it important to use drop_first = True during dummy variable creation?

Answer:

During the dummy variable creation it is important to use drop_first = True because it helps us to reduce one extra column. It will also help us to reduce correlation between the dummies. It also reduces the multi collinearity issue, without any fear of information loss.

For example: A categorical column has 3 unique categorical value say A, B, C and we need to create a dummy variable of this column. Even if we drop, one of the value say A, we will get the same information. If it is not B and C then definitely it is A.

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

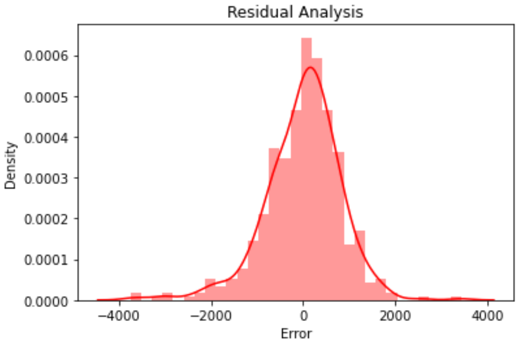
Answer: By looking at the pair-plot, "temp" has the highest correlation with the target variable.

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

Answer:

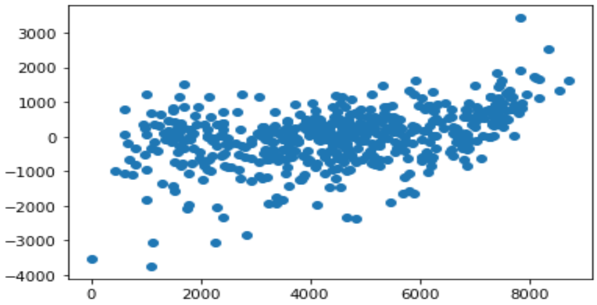
To validate the assumptions of Linear Regression after building the model on the training set:

I. **Residual analysis**: Error terms are normally distributed around the mean (i.e. 0).



II. **Homoscedasticity :**

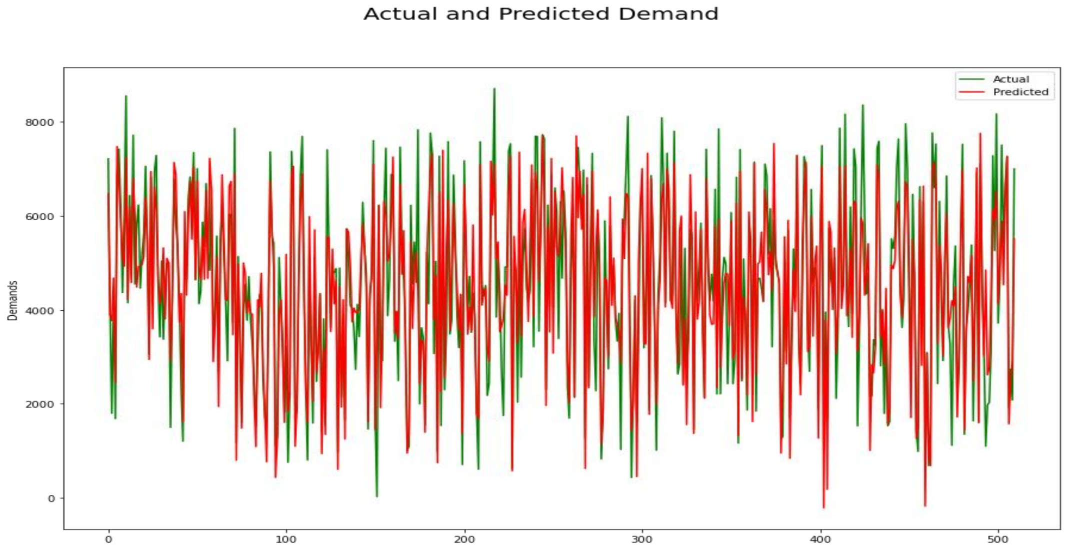
Error terms does not shown any pattern



III. **Multi-collinearity :**

- The VIF's of the features are in desired range (VIF < 5).
- Also by looking at heat-map there no multi collinearity between the features.

IV. The graph for actual and predicted value almost overlap each other.



V. The R2-Score (0.822) and Adj. R2-Score (0.819) are almost similar. Hence, this ensure the good fitness of the model.

<u>Ques 5</u>: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: On the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- Temp
- Weather_B
- Year

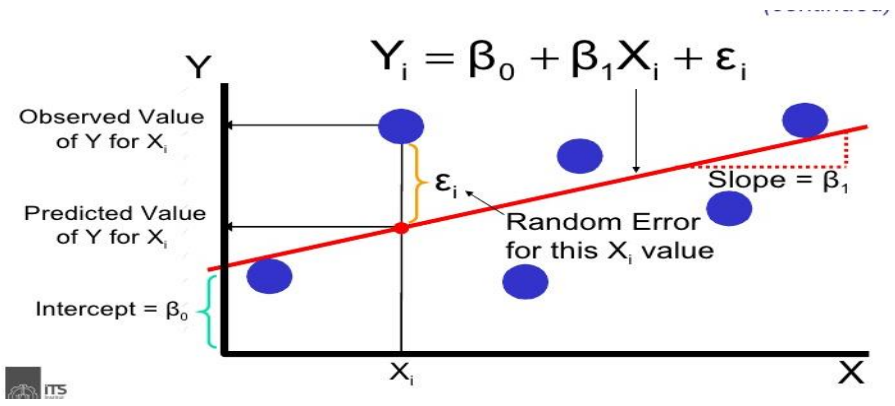
<u>General Subjective Questions</u>:

<u>Ques 1</u>: Explain the linear regression algorithm in detail.

Answer:

- Linear regression algorithm is one of the very basic forms of machine learning algorithm where we train a model to predict the behaviour of your data based on some variables.
- It shows the linear relationship between the dependent variable (target variable) and multiple independent variables. Linear regression shows that, how the dependent variable change w.r.t. the change in independent variable.

- The linear regression model provides a sloped straight line representing the relationship between the variables, this id known as regression line.
  As shown in figure:


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Slope = β_1
Random Error for this X_i value
X_i
X
ITS

Mathematical representation of the linear regression:

Yi= βo+ β1*Xi + ϵ

Where,

Yi = Dependent Variable (Target Variable)
Xi = Independent Variable (Predictor Variable)
βo = Intercept
β1= Slope of the Line
ϵ = Error

- There are mainly two types of linear regression algorithm:

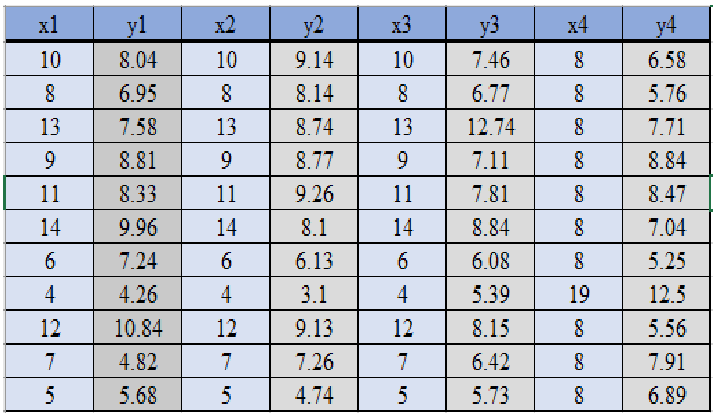
1. **Simple Linear Regression**: If a single independent variable are used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression
2. **Multiple Linear Regression:** If more than one variable are used to predict the value of the numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression

Ques 2: Explain the Anscombe's quartet in detail.

Answer:

- Anscombe's quartet comprises of the four data set and each data set consists of eleven (x, y) points. The basic thing to analyse 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 behaviour irrespective of statistical analysis.

- The data sets:

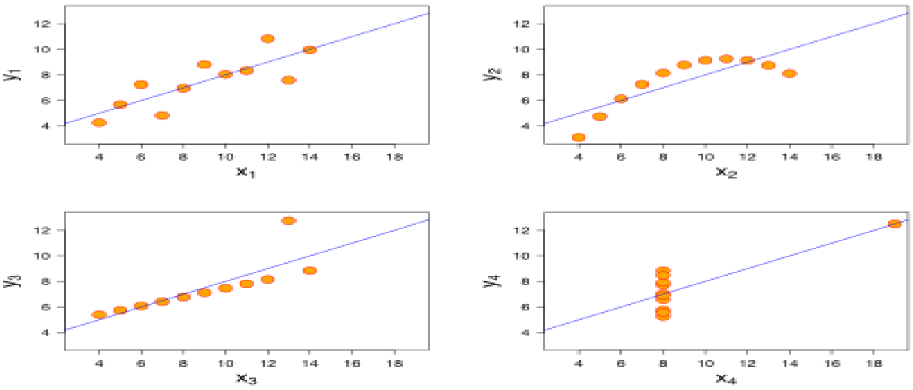


x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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

- Statistical Analysis for the dataset:

  - The average x value is 9 for each dataset.
  - The average y value is 7.50 for each dataset.
  - The variance for x is 11 and the variance for y is 4.12
  - The correlation between x and y is 0.816 for each dataset
  - A linear regression for each dataset follows the equation y = 0.5x + 3

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



  - Graph-1: The first graph represents the simple linear regression graph with a good fit and shows a good linear relationship between data points.

  - Graph-2: The second graph represents the non-linear relationship between the data point. Also shows this could not fit the linear regression line on the data points.

- Graph-3: The third graph also shows the linear relationship between the data points, but there is single outliers.

- Graph-4: The fourth graph shows that the value of X is constant except one outlier.

- **Anscombe's Quartet** is the modal example to demonstrate the importance of data visualization. Even if the four data sets have same statistical representation but when we plot, we get different graph. It shows us the importance of looking at a data set graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Ques 3: What is Pearson's R?

Answer:

- The Pearson's R also known as Pearson correlation coefficient is the measure of the linear correlation between two variables.

- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**.**

- Pearson's R Formula:

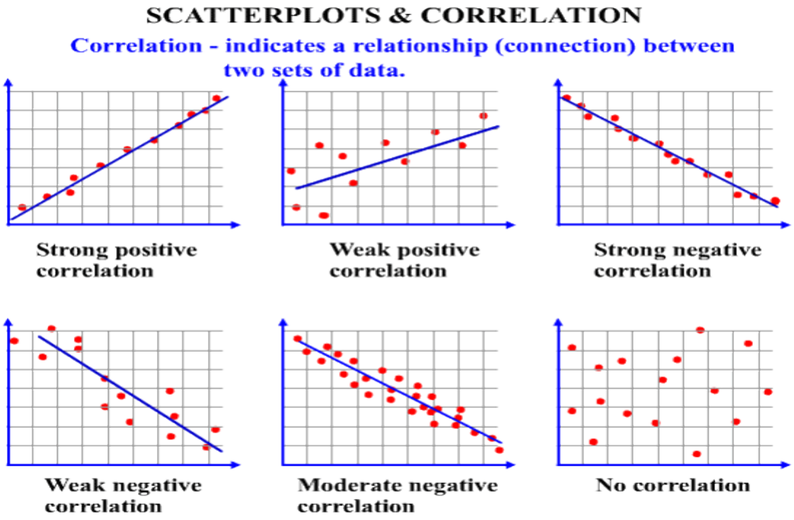
$$r = \frac{N \sum xy - \sum x \sum y}{\sqrt{(N \sum x^2 - \sum x^2)\,(N \sum y^2 - \sum y^2)}}$$

- r = correlation coefficient
- N = number of pair
- $\sum xy$ = sum of the product of the pairs of variables
- $\sum x$ = sum of the x variables
- $\sum y$ = sum of y variables
- $\sum x^2$ = sum of squares of x
- $\sum y^2$ = sum of squares of y

- The value of r varies between -1 to 1

  - r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
  - r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
  - r = 0 means there is no linear association
  - r > 0 < 0.5 means there is a weak association

- r > 0.5 < 0.8 means there is a moderate association



<u>Ques 4</u>: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

- Scaling is the method used to normalize the independent variables. It is performed to bring all the independent variable in same scale.
- The scaling is important because, Most of the time the data features varies in range, if scaling is not done then the model will consider higher value as high and lower value as low. For example, If the temperature of city = 34◦ C and temperature of another city = 93.2◦F, even the temperature of both the cities are same but the machine learning model will consider 93.4 as the higher value and this will lead to wrong prediction.
- If scaling is not done, it just affects the coefficient of the variables but not the other parameters such as t-statistic, F-statistic, p-values, R-squared, etc.

- The Methods of Scaling :

  - Min-Max Scaling:
    - It is the simplest method and consists of rescaling the range of features to scale the range in [0, 1].
    - Min-Max Scaler handles the outliers present in data.
    - Min-Max Scaler Formula: $x' = \frac{x - min(x)}{max(x) - min(x)}$
  - Standardized Scaling:
    - It scales the data between zero mean and unit standard deviation.
    - It is mainly used when the data is normally distributed across the mean.
    - Standardized Scaler Formula: $x' = \frac{x - mean(x)}{\sigma}$

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Ques 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

- The VIF = infinite means that there is a perfect linear relationship between variables.
- In the case of perfect correlation, we get R2 =1, which lead to 1/ (1-R2) infinity.
- An infinite VIF value also indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

Ques 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

- Q-Q plot also known as Quantile-Quantile plot is a graphical tool to validate if two dataset are coming from the populations with common distribution.
- Most of the time we assume that the given data is normally distributed around the mean for ease of inferring something useful information. Q-Q plots are one of the way to prove our assumptions correctness. Using Q-Q plot not only normal distribution, we can test other distributions such as uniform distribution etc.
- The Quantile are the breakpoints that divides the numerical data into equal sized bins. Percentiles are a type of quantiles that divide the data into 100 equal bins; quartiles divide the data into 4 equal parts and so on.
- It also compares the quantile of two datasets. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a straight line.

- A Q–Q plot compares the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- Importance of Q-Q plot:
    - Q-Q plots detects whether the two data sets come from populations with a common distribution.
    - A Q–Q plot compares the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
    - Q-Q plot can detect outliers, shifts in scale, symmetry etc. simultaneously.