

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?

Ans. The demand of bike is less in the month of spring when compared with other seasons

- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non-working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

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

Ans. Dropping the first columns as (p-1) dummies can explain p categories.

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

Ans. By looking at the pair plot temp variable *has the highest (0.63) correlation with target variable 'cnt'*.

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

Ans. The assumptions are as follows:

1. Residual errors follow normal distribution.
2. Maintains linear relation between dependant variable (Test and Predicted).

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

Ans. Based on final model top three features contributing significantly towards explaining the demand are:

0. Temperature (0.552)
1. weathersit: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)

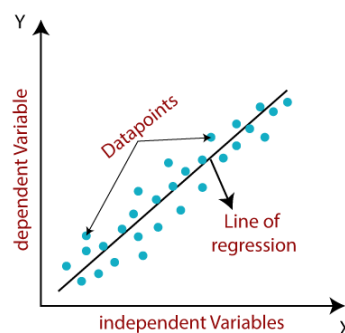
2. year (0.256)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1x + \epsilon$$

Types of Linear Regression:

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression**
- **Multiple Linear regression**

Linear Regression Line:

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship**
- **Negative Linear Relationship**

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error. The different values for weights or the coefficient of lines (a_0 , a_1) gives a different line of regression, so to calculate this we use cost function.

Cost function-

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing. We can use the cost function to find the accuracy of the **mapping function**. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$MSE = \frac{1}{N} \sum_{i=1}^n (y_i - (a_1 x_i + a_0))^2$$

Residuals: The distance between the actual value and predicted values is called residual.

Gradient Descent: It is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

R-squared method:

R-squared is a statistical method that determines the goodness of fit. It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%. The high value of R-square represents a good model. It is also called a **coefficient of determination**, or **coefficient of multiple determination** for multiple regression.

$$R\text{-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression:

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target**
- **Small or no multicollinearity between the features**
- **Homoscedasticity Assumption**
- **Normal distribution of error terms**
- **No autocorrelations**

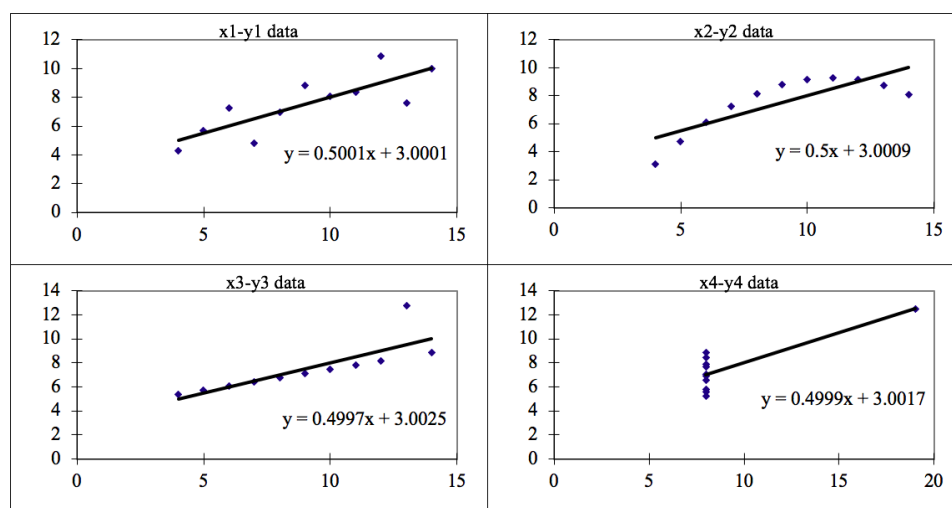
2. Explain the Anscombe's quartet in details.

Ans. **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets. The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4	
1	10	8.04	10	9.14	10	7.46	8	6.58	
2	8	6.95	8	8.14	8	6.77	8	5.76	
3	13	7.58	13	8.74	13	12.74	8	7.71	
4	9	8.81	9	8.77	9	7.11	8	8.84	
5	11	8.33	11	9.26	11	7.81	8	8.47	
6	14	9.96	14	8.1	14	8.84	8	7.04	
7	6	7.24	6	6.13	6	6.08	8	5.25	
8	4	4.26	4	3.1	4	5.39	19	12.5	
9	12	10.84	12	9.13	12	8.15	8	5.56	
10	7	4.82	7	7.26	7	6.42	8	7.91	
11	5	5.68	5	4.74	5	5.73	8	6.89	
Summary Statistics									
N	11	11	11	11	11	11	11	11	
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50	
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94	
r	0.82		0.82		0.82		0.82		

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

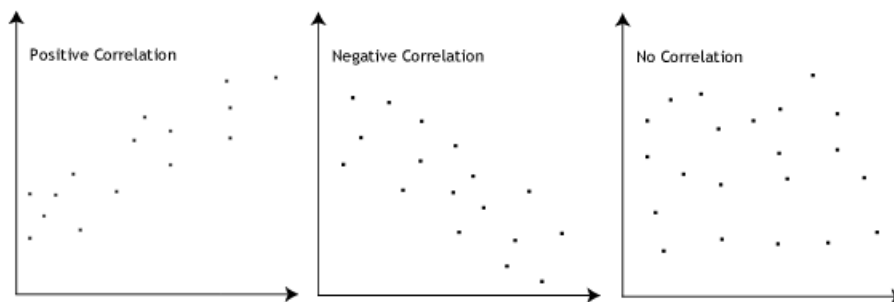
3. What is Pearson's R?

Ans. In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their

standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1 .

The Pearson's correlation coefficient varies between -1 and $+1$ where:

- $r = 1$ 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
- $r > 0 < 5$ means there is a weak association
- $r > 5 < 8$ means there is a moderate association
- $r > 8$ means there is a strong association



Pearson r Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Ans. **Scaling:**

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

The purpose of doing Scaling:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

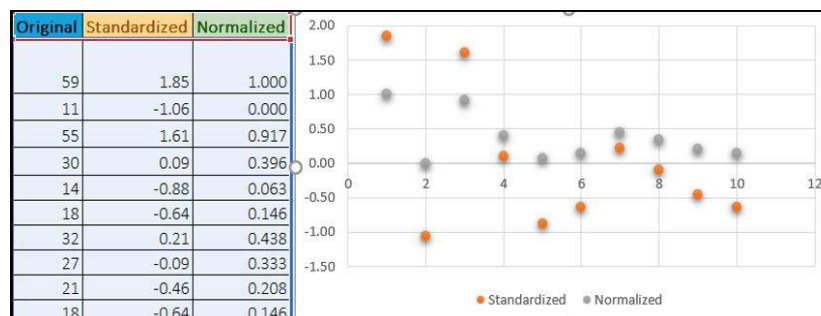
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Example:

Below shows example of Standardized and Normalized scaling on original values.



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

Ans. If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1 - R^2)$ infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

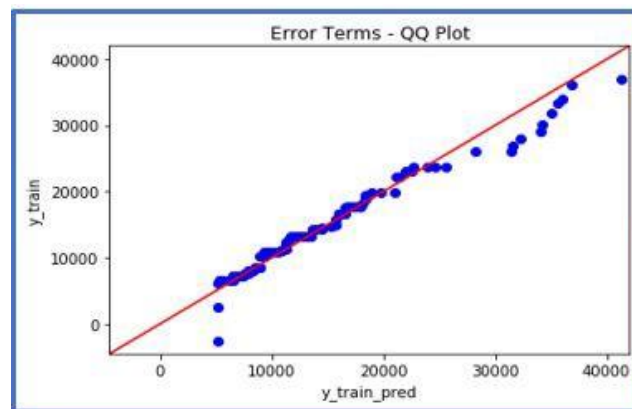
Few advantages:

- a) It can be used with sample sizes also
- b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

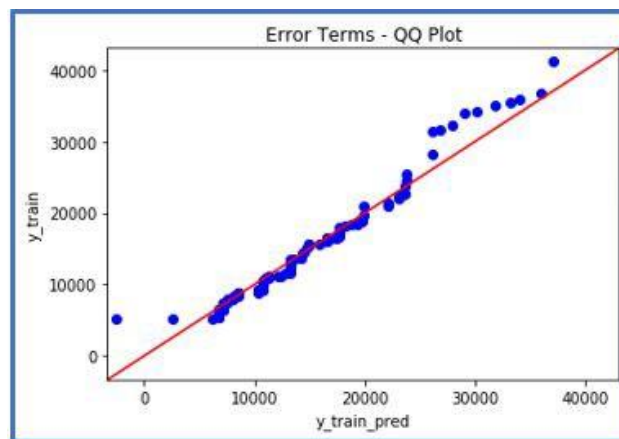
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- a) **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x - axis
- b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



- c) **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



- d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis