Name : Dhananjaya BL

Batch ID: 1993

Topic : Multiple Linear Regression

Assignment based subjective questions

Q1 : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANS :

- There were 6 categorical variables considered for analysis – ‘season’, ‘month’, ‘holiday’, ‘weekday’, ‘workingday’ and ‘weathersit’
- Variable ‘holiday’ was found to be not a significant variable
- Factors of ‘season’ – ‘summer’ and ‘winter’ were found significant and were found to have positive impact on the demand
- Only ‘month_9’ was found to be significant among the factors (12) of ‘month’ variable and has the positive impact on the demand
- Only ‘weekday_6’ was found to be significant among the factors (7) of ‘weekdays’ variable and has the positive impact on the demand
- If it is a ‘Workingday’, then the demand for the bikes expected to go up
- Type of weather – type 2 and type 3 are fund to have negative impact on the demand for the bikes

MLR model equation:

$count$=1295.4+(5307.68$*temp$)−(1153.59$*windspeed$)+(745.78$*season$2)+(1228.76$*season$4)+(792.85$*mnth$9)+(592.63$*weekday$6)+(565.29$*workingday$1)−(685.62$*weathersit$2)−(2876.88$*weathersit$3)

Q2: Why is it important to use drop_first=True during dummy variable creation?

ANS:

If a categorical variable has k levels, and when we create dummy variables there will be k binary variables. However k-1 dummy variables can represent the categorical variable with k levels without losing any information.

If we include all k binary dummy variables, it will introduce redundancy, that is increases in the degree of freedom by 1, without increasing the information gain. This will have impact on the model performance and also it will introduces multi-collinearity in the data and solution will not be reliable.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

ANS:

Pair-plot only give the direction of association or correlation but does not quantify the strength of association.

From Pair-plot we see that both `temp` and `atemp` has strong correlation with target variable. However when we compute the Pearson Correlation Coefficients between 'atemp' and 'cnt' = 0.6307 and 'tem' and 'cnt' = 0.627.

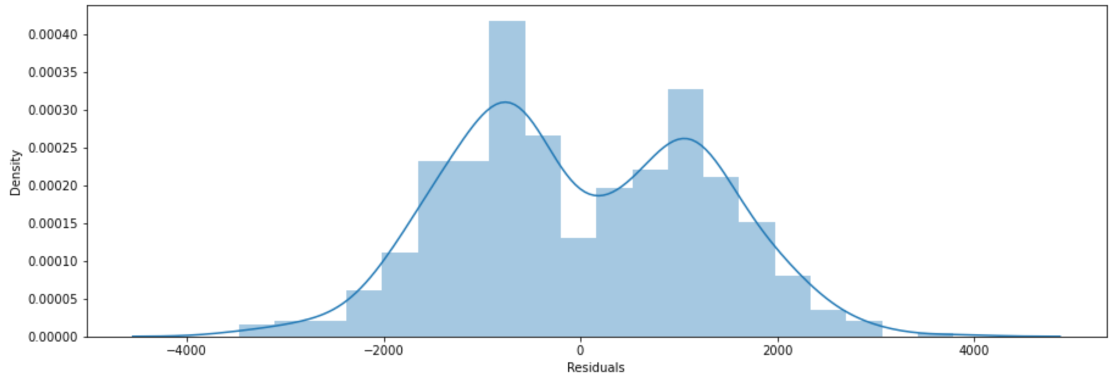
'atemp` has slightly higher correlation with 'cnt' compared to 'temp`

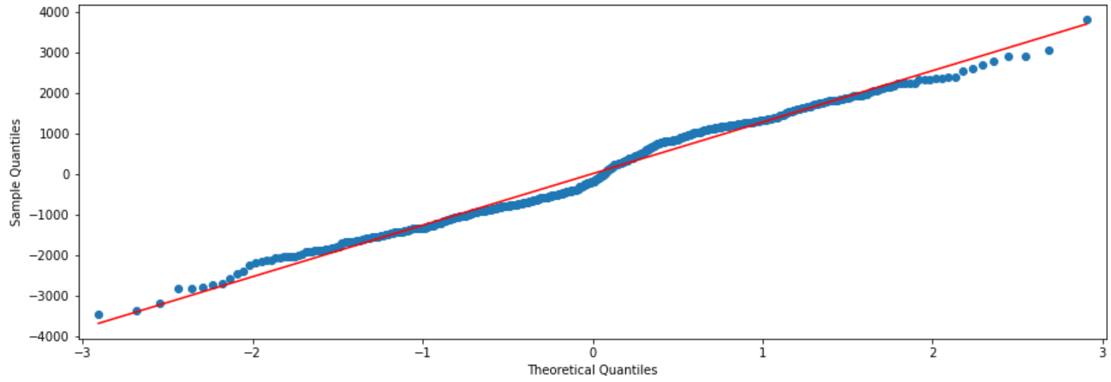
Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

ANS:

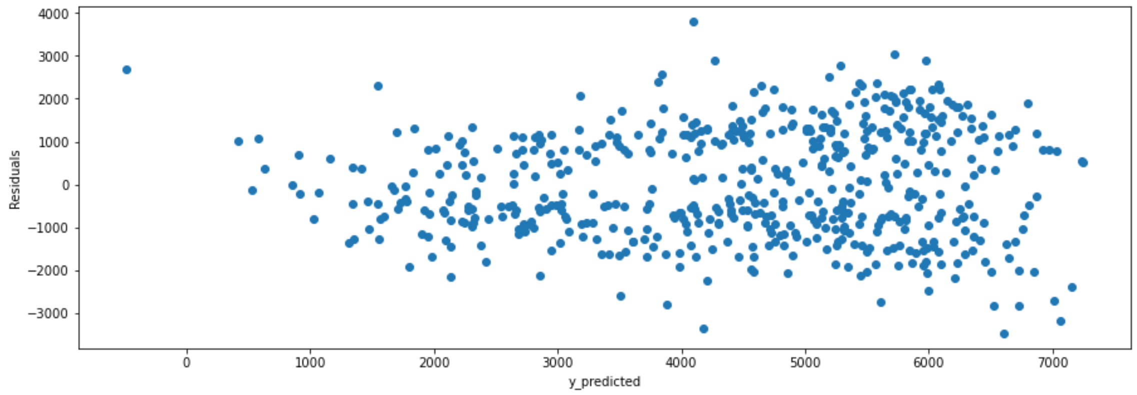
Following assumption are tested after modelling on the training set:

- Error follows normal distribution
  - Computed the predicted values for training data
  - Calculated the residuals = (y_actual – y_predicted)
  - Visual inspection of histogram / distribution plot of error terms
  - Used QQplot to confirm the normality of residuals
  - Checked the mean of residuals, which was close to zero

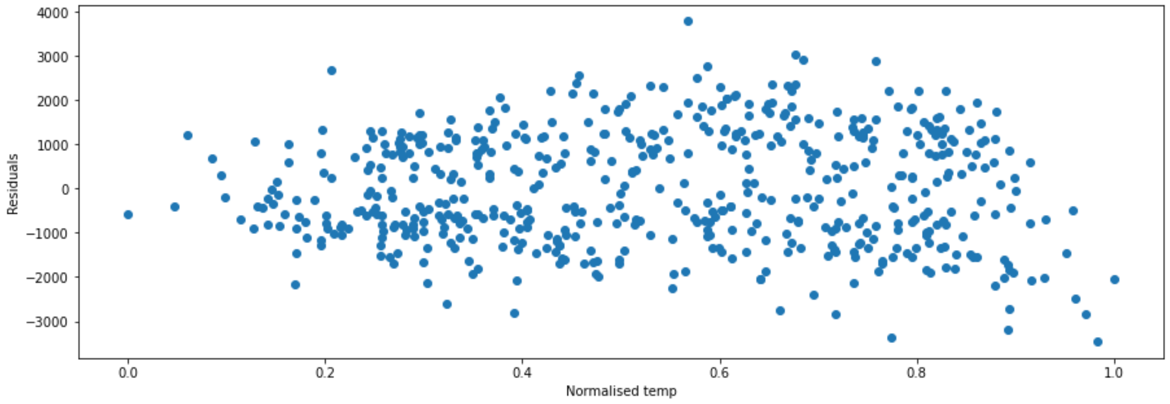




- Error terms are independent
    - Plotted the Residuals vs Precited values for train dataset
    - Observed no relationship between them, can infer errors are independent



- Error terms have constant variance (homoscedasticity)
    - Plotted the Residuals vs Normalized X variable (temp, windspeed)
    - Found no observable patterns, but a random scatter



Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

ANS:

MLR model equation:

$$count=1295.4+(5307.68*temp)-(1153.59*windspeed)+(745.78*season2)+(1228.76*season4)+(792.85*mnth9)+(592.63*weekday6)+(565.29*workingday1)-(685.62*weathersit2)-(2876.88*weathersit3)$$

For the above model, based on the regression coefficients, the top three features are

- ‘temp’ – temperature in degree Celsius
- ‘weathersit_3’ - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- ‘windspeed’ - wind speed

‘temp’ has a positive impact on the demand where are weathersit_3 and windspeed has negative impact on the demand

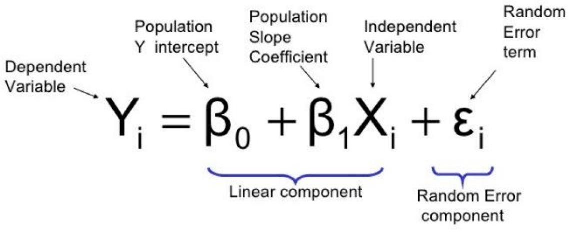
---

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

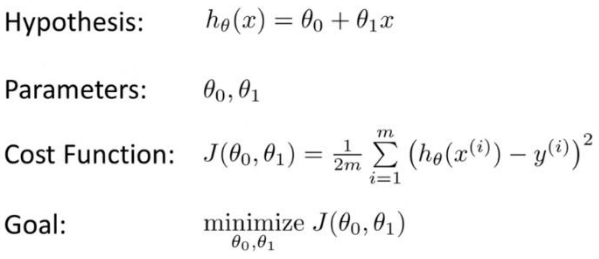
ANS:

- Linear Regression is one of the supervised machine learning algorithms
- The fundamental assumption of linear regression is that there is a linear relationship between independent variables and dependent variable
- If there is only one independent variable, then it is known as simple linear regression
- If there are more than one independent variable, then it is known as multiple linear regression
- In linear regression the dependent variable is of continuous data type, independent variables can be of continuous or categorical
- Simple Linear Regression is a technique to find the best fit line, that explains the maximum variance in the dependent variable, by minimizing the Residual Sum of Squares
- In Multiple Linear Regression the process is to fit the best hyper-plane in n-dimensional data (n-dimensions = n-number of features)
- One of the method to obtain the best fit solution is through Ordinary Least Square method
- In this method the objective is to minimise the Residual Sum of Square, which is also the cost function.
- Residual Sum of Square is sum of squared difference between the actual values to the predicted values
- Gradient Descent optimization technique is used to reduce the RSS, iteratively
- Simple Linear Regression can be expressed as :



$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

- B0 and B1 are obtained using Gradient Descent approach, by minimizing the following cost function



Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

- Significance of the features used in the model is checked though t-test, if p-value is less than set significance level (0.01, 0.05, 0.1) then the feature is considered significant
- Model significance tested via F-test
- The strength of the Linear Regression model can be assessed by
  - R-squared (Adjusted R-squared in case of MLR) / co-efficient of determination
  - Residual Sum of Squares Error (RSE)

    R-squared can be given as 1- Unexplained Variance / Total Variance = 1 – (RSS)/(TSS)
    R-squared value varies between 0 and 1 (some time negative)
    Higher the R-squared better the model

    Generally, RSE must be as low as possible, when comparing RSE from different models

Key Assumptions of Linear Regression:

- The Regression model is linear in parameters
- The residual of the regression follows normal distribution
- The residual of the regression has constant variance (homoscedasticity)
- There is no auto correlation between errors – errors are independent
- The Independent variables are assumed to be non-stochastic (not random)
- There is no perfect multi-collinearity among independent variables

Linear Regression Models can be used for;

- Predicting the target variable value, given independent variable values
- Identify the variables the affect the target variable
- To explain the variation in the target variable using independent variables

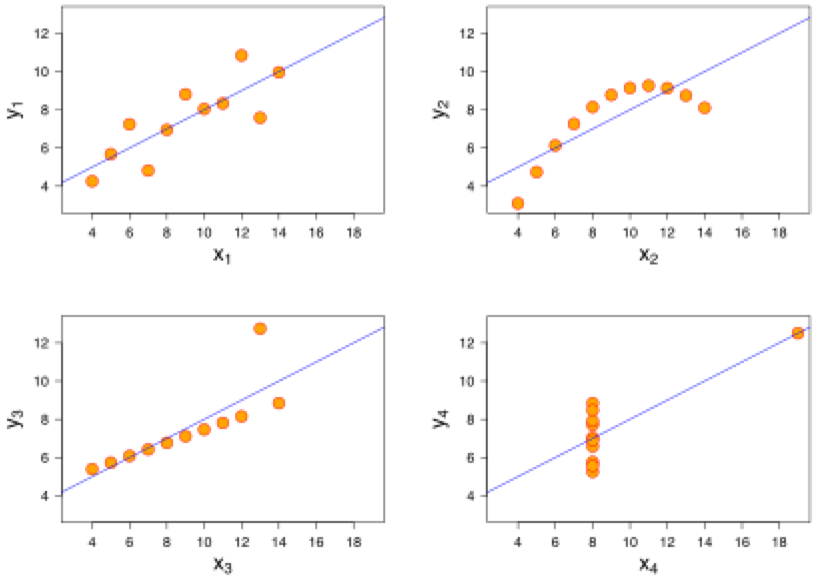
Eg for Linear Regression Application

1. Predicting the amount of rain fall, given other environmental parameters
2. Predicting the growth of GDP of the nation, given macro economical parameters
3. Predicting the Blood Pressure, given the information like activity levels, age, gender, health condition and other data

Q2: Explain the Anscombe’s quartet in detail

ANS:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. -- Wikipedia



Scatter plot 1: shows fitment of the linear regression model pretty well.

Scatter plot 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Scatter plot 3: the distribution is linear but shows the effect of outliers on regression line, pulling the line towards the outlier

Scatter plot 4: there is no linear relationship between two variables but an outlier creates the leverage points and produces high correlation co-efficient

What we understand is though the data may exhibit identical descriptive statistics but vary in their distribution. These types of data can fit a good regression line but may fail to generalize. It shows the importance of data visualization, before implementing any algorithm.

Q3: What is Pearson's R?

ANS:

- Pearson' R is also known as Pearson Correlation Coefficient or Pearson product-moment correlation coefficient.
- is a measure of linear correlation between two sets of data. It indicates the strength of the linear relationship between two variables
- It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.
- Correlation coefficient equal to 1 or -1 indicates the perfect correlation, 0 indicates the no correlation.
- Correlation coefficient equal to > 0 indicates positive correlation and < 0 indicates negative correlation
- It does not capture the non-linear or any other form of relationship between variables.

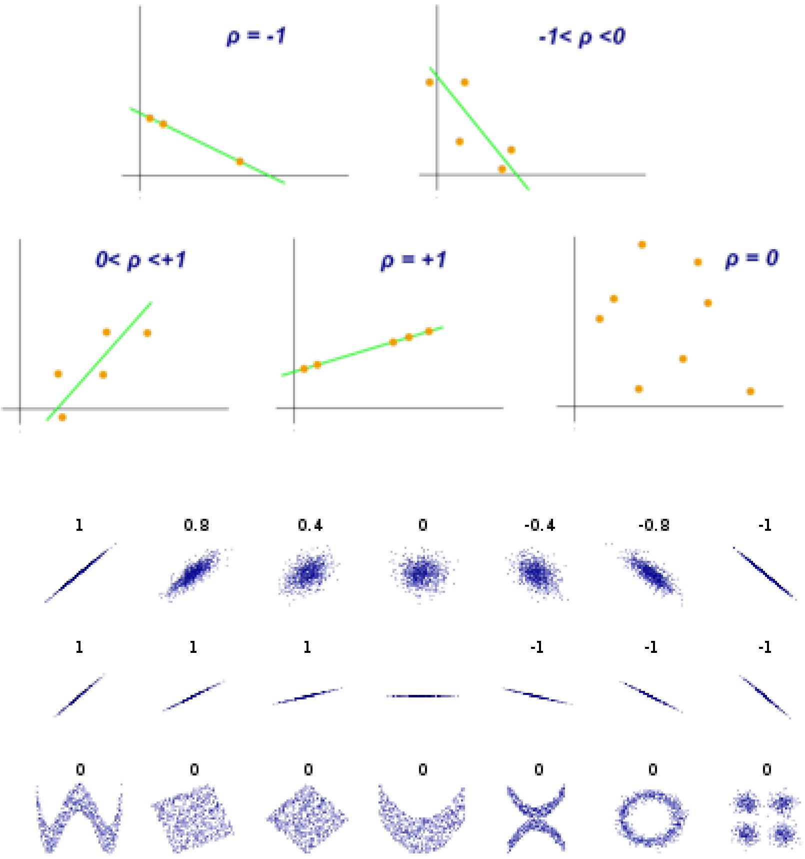
- Given a pair of random variables (X,Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

- Under heavy noise conditions, extracting the correlation coefficient between two sets of stochastic variables is nontrivial, correlation coefficient is affected by outliers.
- Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

ANS:

Scaling : scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is generally performed during the data preprocessing step.

Eg: if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (Rs 25,000– Rs 75,000), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range

Methods of Scaling

- Min-Max Scaling (Normalization) - rescaling the range of features to scale the range in [0, 1]. Given as

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

- Standardization - Feature standardization makes the values of each feature in the data have zero mean and unit variance.

$$x' = \frac{x - \bar{x}}{\sigma}$$

- MaxAbs Scaler - This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/center the data

$$x_{scaled} = \frac{x}{max(|x|)}$$

- Robust Scaler - This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). This is best employed when outliers influence the sample mean / variance.

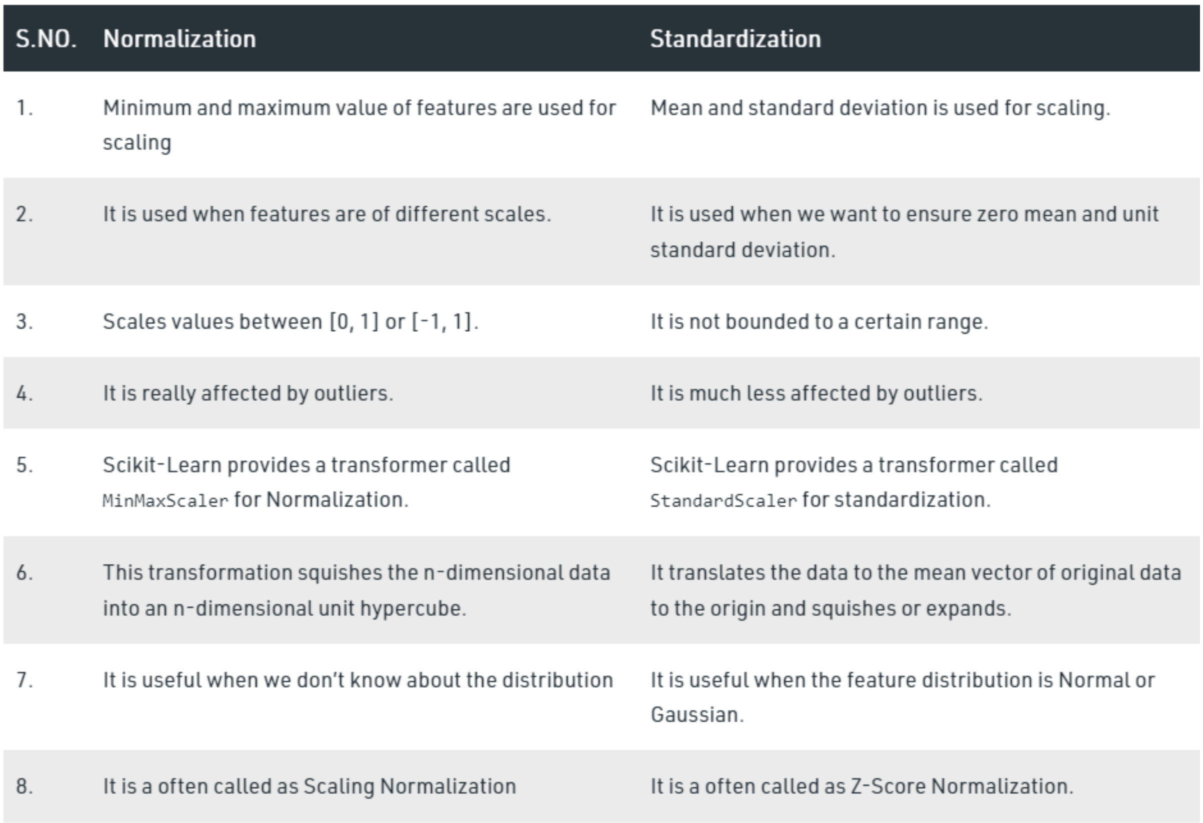
$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Why is scaling performed?

Features with different scale / range lead to difficulties to visualize the data and, they can degrade the predictive performance of many machine learning algorithms. Unscaled data can also slow down or even prevent the convergence of many gradient-based estimators.

Machine learning algorithms gives more weightage feature to large values compared to feature with smaller values, but in reality, both features may be of equal importance.

What is the difference between normalized scaling and standardized scaling?



S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

ANS:

VIF : Variance Inflation Factor is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to multi-collinearity. To determine VIF, we fit a regression model between the independent variables.

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

VIF for ith features is given as above. R_i^2 represents the R-squared obtained taking ith features as dependent variable and all others as independent variables.

- If all the independent variables are orthogonal to each other, then VIF = 1.0.
- A large value of VIF indicates that there is a correlation between the variables.
- If there is perfect correlation, then VIF = infinity
- Where there is a perfect correlation among independent variables then the R-squared approaches close to 1, then the denominator in VIF formula is zero or close to zero, taking the VIF number to infinity or close to infinity.
- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).
- The standard error of the coefficient determines the confidence interval of the model coefficients.
- If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.
- A general rule of thumb is that if VIF > 10 then there is multicollinearity, for model development purpose VIF > 5 is used for feature selection along with p-value.

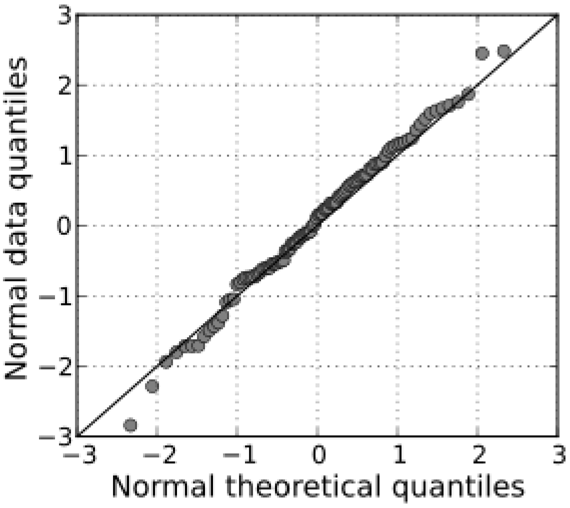
Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

ANS:

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure

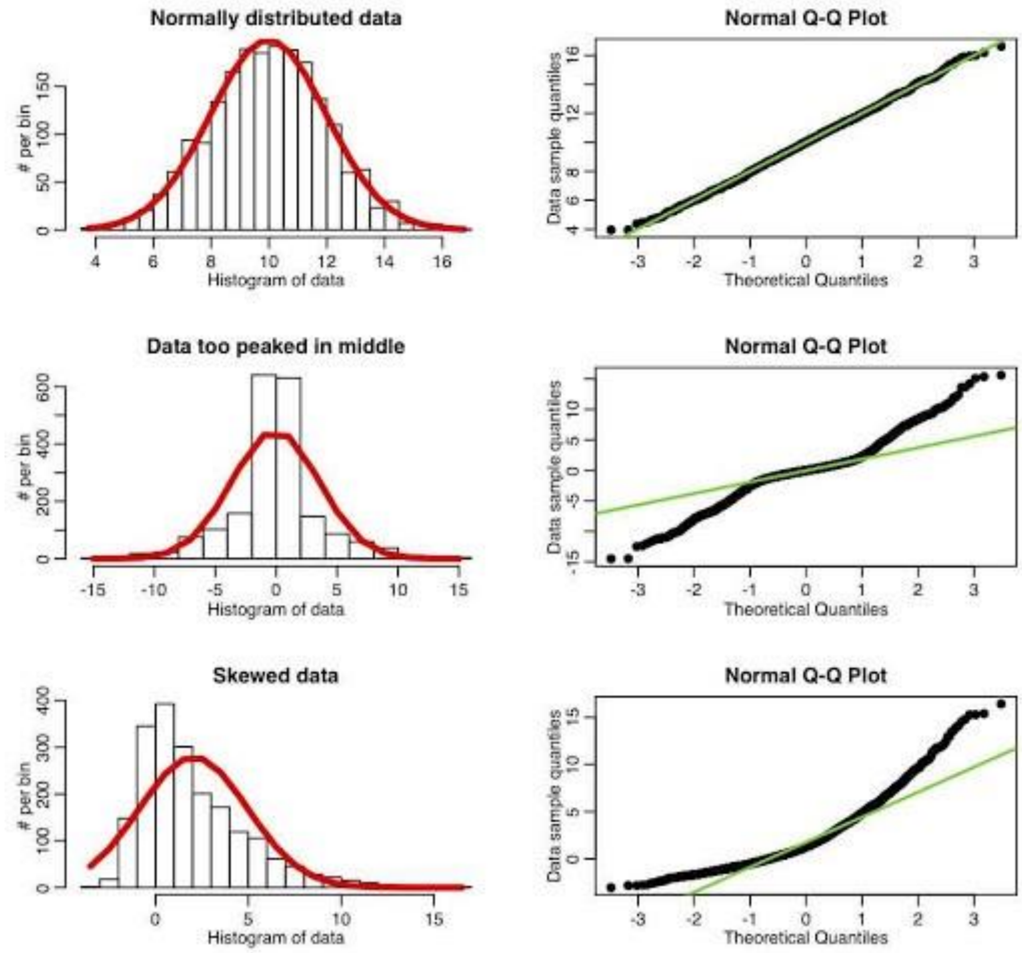


from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.

Importance of a Q-Q plot in linear regression:

- In linear regression, the residual obtained after making prediction, can be visualized using Q-Q plot, to determine if the residuals follows normal distribution.
- If the residuals fall on or close to 45-degree reference line, then we infer that residual are normally distributed.


Normally distributed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles