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

Linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.

This model is to find a linear relationship between the input variable(s) ***X*** and the single output variable ***y***.
***Simple linear regression:*** When there is only single independent/feature variable X then it is called as simple linear regression.
***Multiple linear regression:*** When there are multiple independent/feature variables $X_i$ then it is called as Multiple linear regression.

- The independent variable is also known as the predictor variable.
- The dependent variables are also known as the output variables.

$Y = \beta_0 + \beta_1 X$ (SLR)

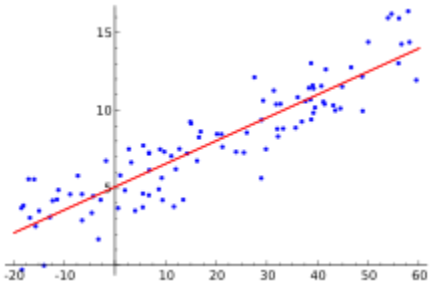
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$ (MLR)

*Where:*

*Y = how far up ↑ and X = how far along →*

$\beta_1, \beta_2 .. B_p$ = Slope or Gradient (how steep the line is)

$\beta_0$ = value of Y when X=0 (Y-intercept)

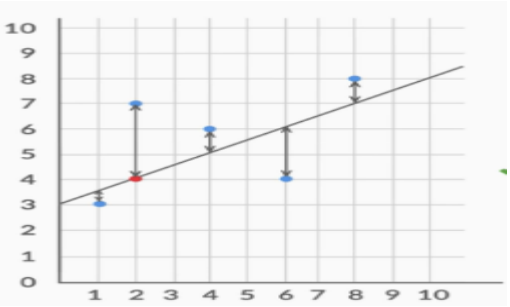



As part of linear regression, there can be multiple lines which can be drawn from the data points as part of scatter plot but regression model can help to identify model that is best fit line from the data points.

***Cost Function:***
The cost function helps to figure out the best possible values for $\beta_0$, $\beta_1$, $\beta_2$ etc.. which would provide the ***best fit line*** for the data points. We need to convert this problem into a minimization problem where we would like to *minimize the error between the predicted value and the actual value.*

It means that given a regression line through the data, we calculate the distance from each actual data point to the regression line (predicated values), square it, and sum all of the squared errors together. This is called ***Residual Sum of Squares (RSS)***

Then we divide this RSS values by total number of data points which provides average squared error of all the data points and it is called ***Mean Square Error (MSE)*** . MSE is also known as cost function using which we need to identify optimal values of co-efficients and interceptor such that MSE values settles at minima.




Residual/Error = $e_i = y_i - y_{pred}$

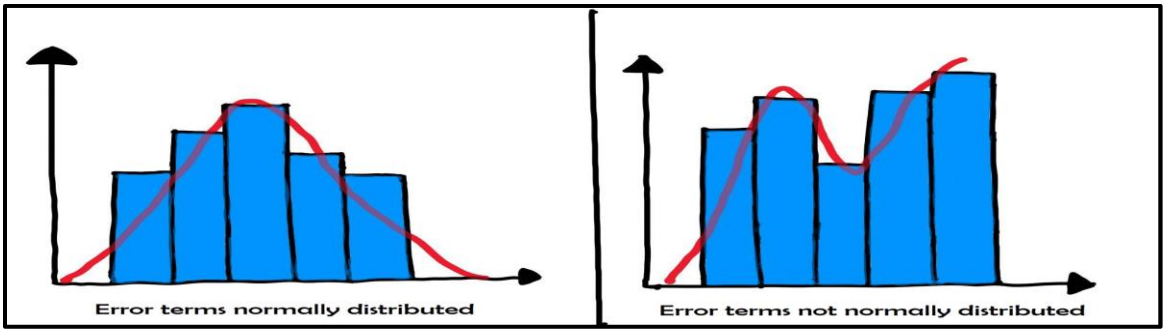
RSS = $e_1^2 + e_2^2 + .....+e_n^2$

RSS = $\sum_{i=0}^{n} (Yi - \beta 0 - \beta 1 X) 2$

MSE = RSS/n

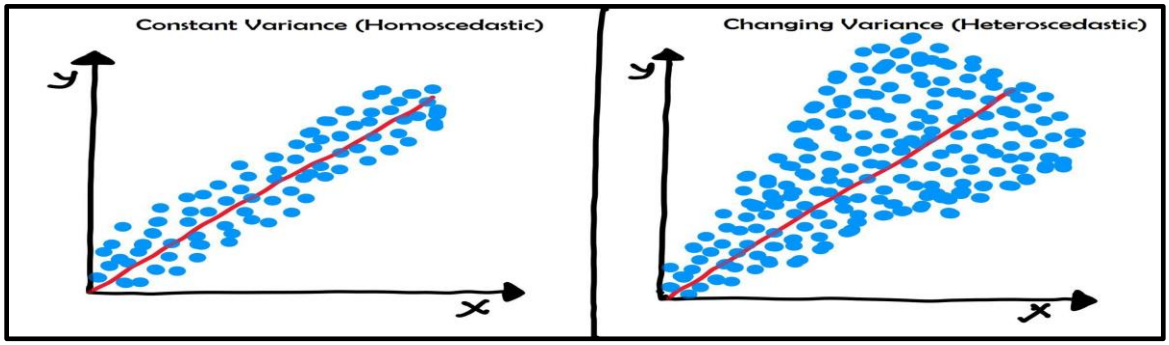
## 2. *What are the assumptions of linear regression regarding residuals?*

a) ***Normality assumption***: It is assumed that the error terms, ε(i), are normally distributed. If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.




b) ***Zero mean assumption***: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

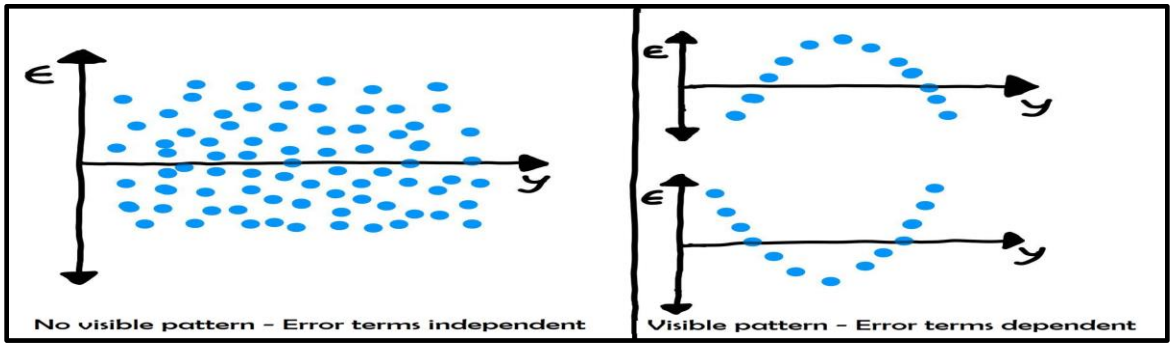
*c)* ***Constant variance assumption***: It is assumed that the residual terms have the same (but unknown) variance, $\sigma^2$ . This assumption is also known as the assumption of ***homogeneity*** *or* ***homoscedasticity.***




d) ***Independent error assumption***: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves.

If some correlation is present, it implies that there is some relation that the regression model is not able to identify

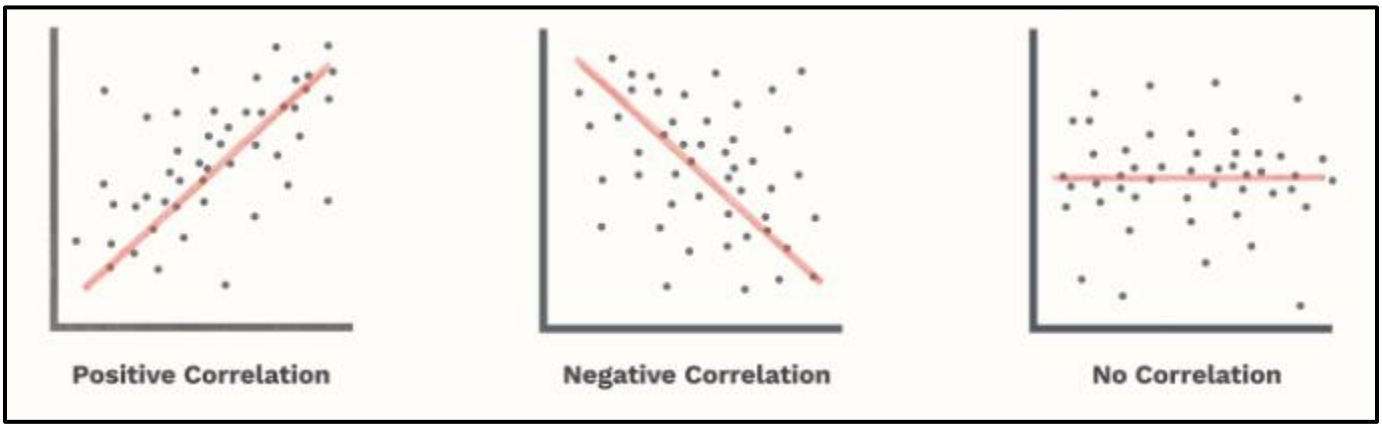
If the independent variables are not linearly independent of each other, the uniqueness of the least square's solution (or normal equation solution) is lost.



## 3. *What is the coefficient of correlation and the coefficient of determination?*

a) ***Coefficient of Correlation (r):***

- ✓ It measures the strength and the direction of a linear relationship between two variables (x and y) with possible values between -1 and 1.
- ✓ ***Positive Correlation:*** It indicates that two variables are in perfect harmony. They rise and fall together. +1 is perfect +ve correlation
- ✓ ***Negative Correlation:*** It indicates that two variables are perfect opposites. One goes up and other goes down. -1 is perfect -ve correlation
- ✓ ***No correlation***: If there is no linear correlation or a weak linear correlation, r is close to 0.




b) ***Coefficient of determination ($r^2$):***

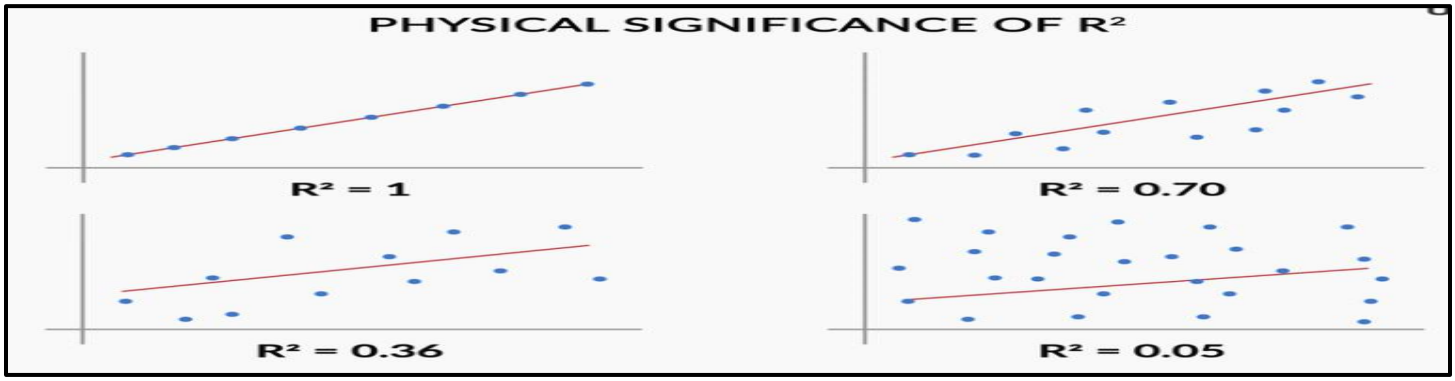
- ✓ Coefficient of determination ($r^2$) = Coefficient of Correlation (r) x Coefficient of Correlation (r)
- ✓ It provides percentage variation in y which is explained by all the x variables together
- ✓ Its value is (usually) between 0 and 1 and it indicates strength of Linear Regression model
- ✓ Higher the R2 value, data points are ***less scattered*** so it is a good model. Lesser the R2 value is more scattered the data points.
- ✓ It can also be calculated as below:

**$R^2 = 1 - (RSS/TSS)$**

Where

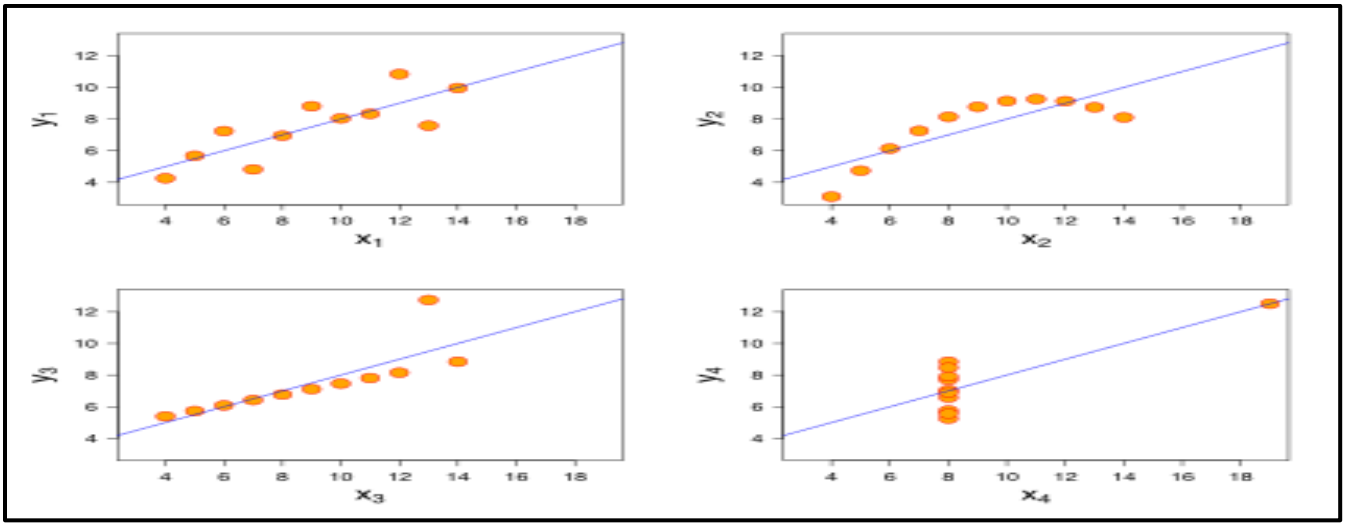
RSS = Residual Sum of Square

TSS = Total Sum of Square (It's square of (actual value - average value))



## 4. *Explain the Anscombe's quartet in detail.*

- Anscombe's Quartet was developed to demonstrate both the importance of graphing data before analyzing it and the ***effect of outliers*** and other ***influential observations*** on statistical properties.
- It was developed by statistician Francis Anscombe. It comprises four datasets that have nearly identical simple descriptive statistics and each dataset containing eleven (x,y) pairs.




a) The first scatter plot (x1 vs y1) appears to be a simple linear relationship
b) The second graph (x2 vs y2) is not distributed normally and the Pearson correlation coefficient is not relevant.
c) The third graph (x3 vs y3), the distribution is linear, but the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
d) The fourth graph (x4 vs y3) shows that one outlier point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

***In short. this quartet emphasizes the importance of visualization in Data Analysis***

## 5. What is Pearson's R?

Pearson's R has multiple names like ***Pearson correlation coefficient, Pearson product-moment correlation coefficient (PPMCC)*** *or* ***bivariate correlation.***

It is a measure of the strength of a linear association between two variables and is denoted by r. It tries to draw a line best fit through the data of two variables. For +ve, -ve and no correlation information, refer question 3.

***Formula:*** Below formula can be used to calculate Pearson's R.

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

where:

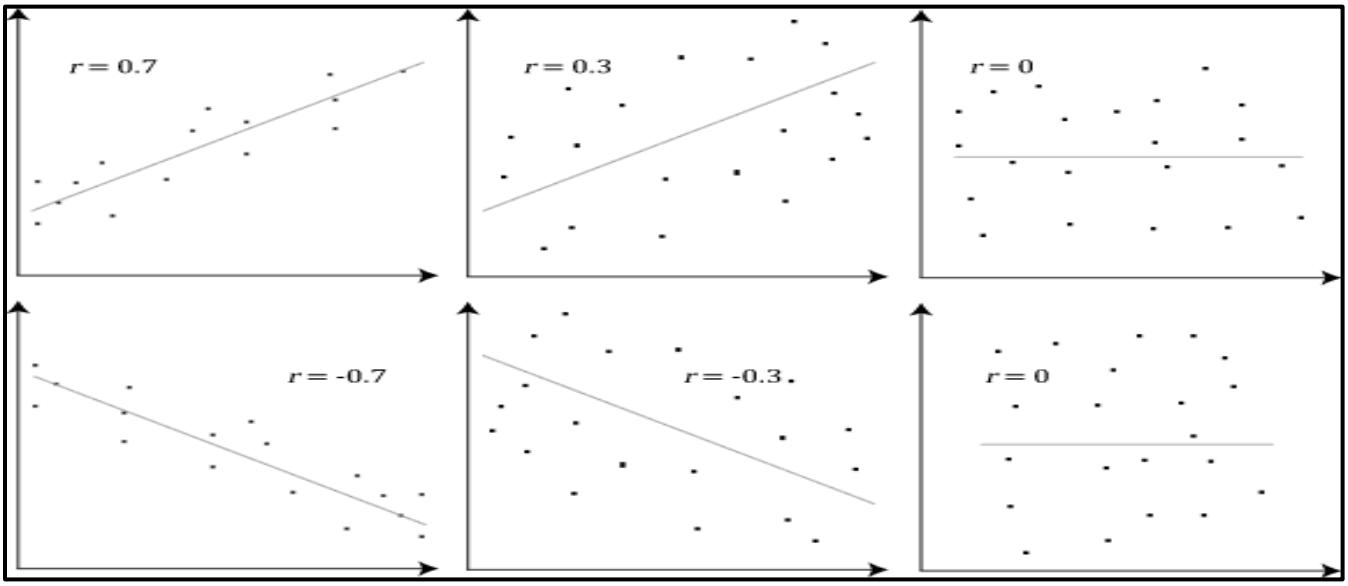
✓ cov is the covariance

- ✓ $\sigma_x$ is the standard deviation of X
- ✓ $\sigma_y$ is the standard deviation of Y

Below guidelines can be used to determine strength of association of two variables. This is only guideline so values can be interpreted differently case to case.

| Strength of Association | Positive | Negative |
|---|---|---|
| Small | .1 to .3 | -0.1 to -0.3 |
| Medium | .3 to .5 | -0.3 to -0.5 |
| Large | .5 to 1.0 | -0.5 to -1.0 |

Different relationships and their correlation coefficients are shown in the diagram below:




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

<u>What is scaling:</u>

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.

<u>Why is scaling:</u>

Most of the times, collected dataset 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.

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

*Standardization:*

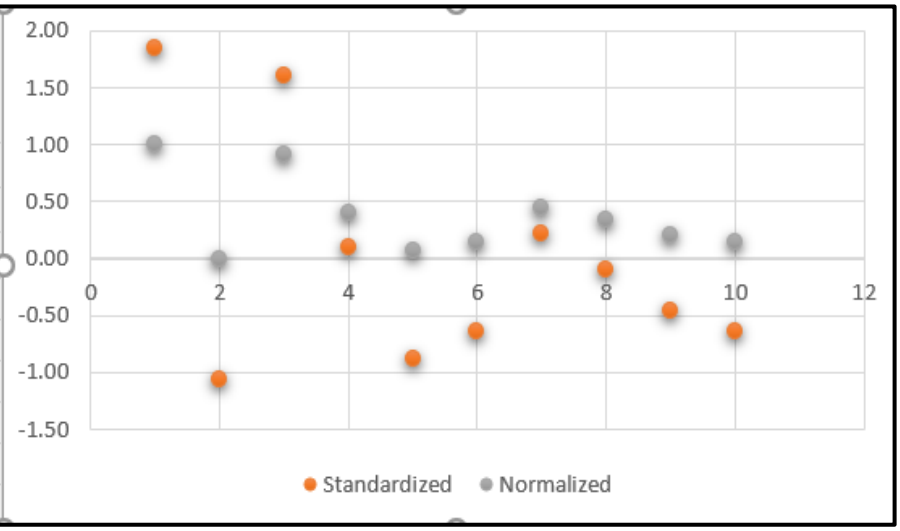
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- ***sklearn.preprocessing.scale*** helps to implement standardization in python.
- *One disadvantage of normalization over standardization is that it* loses *some information in the data, especially about* outliers.

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

*Example:*

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

| Original | Standardized | Normalized |
|---|---|---|
| 59 | 1.85 | 1.000 |
| 11 | -1.06 | 0.000 |
| 55 | 1.61 | 0.917 |
| 30 | 0.09 | 0.396 |
| 14 | -0.88 | 0.063 |
| 18 | -0.64 | 0.146 |
| 32 | 0.21 | 0.438 |
| 27 | -0.09 | 0.333 |
| 21 | -0.46 | 0.208 |
| 18 | -0.64 | 0.146 |




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

Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset. It is calculated as— $VIF_i = 1/(1-R_i^2)$

where *'i'* refers to the $I^{th}$ variable which is being represented as a linear combination of rest of the independent variables.

*When corresponding variable is being expressed exactly by a linear combination of other independent variables, $R^2$ becomes 1 hence denominator of VIF equation $(1-R^2) = 0$ so VIF is infinite.*

***For example:*** Below example shows that 'idi' fuel system and fueltype each has exact linear correlation with all other independent variables considered and hence VIF is inf so they can be dropped from linear regression model.

| Index | Features | VIF |
|---|---|---|
| 16 | idi | inf |
| 1 | fueltype | inf |
| 11 | compressionratio | 83.34 |
| 8 | enginesize | 40.24 |
| 6 | curbweight | 21.52 |
| 7 | cylindernumber | 21.07 |
| 4 | carlength | 12.7 |
| 18 | peugeot | 10.62 |
| 14 | l | 10.06 |
| 13 | citympg | 8.39 |
| 5 | carwidth | 8.28 |
| 9 | boreratio | 7.86 |
| 3 | wheelbase | 7.63 |
| 10 | stroke | 3.86 |
| 19 | porsche | 2.95 |
| 15 | rotor | 2.65 |
| 12 | peakrpm | 2.07 |
| 2 | enginelocation | 2.02 |

## 8. *What is the Gauss-Markov theorem?*

It tells us that if below of assumptions are met, the ordinary least squares (OLS) estimate for regression coefficients gives the ***B****est* ***L****inear* ***U****nbiased* ***E****stimate (BLUE)* possible.

There are five Gauss Markov assumptions/conditions:

1. ***Linearity***: the parameters we are estimating using the OLS method must be themselves linear.
2. ***Random***: our data must have been randomly sampled from the population.
3. ***Non-Collinearity***: the regressors being calculated aren't perfectly correlated with each other.
4. *Exogeneity*: the regressors aren't correlated with the error term.
5. *Homoscedasticity*: Residual terms have the same (but unknown) variance, $\sigma^2$ no matter what the values of regressors might be.

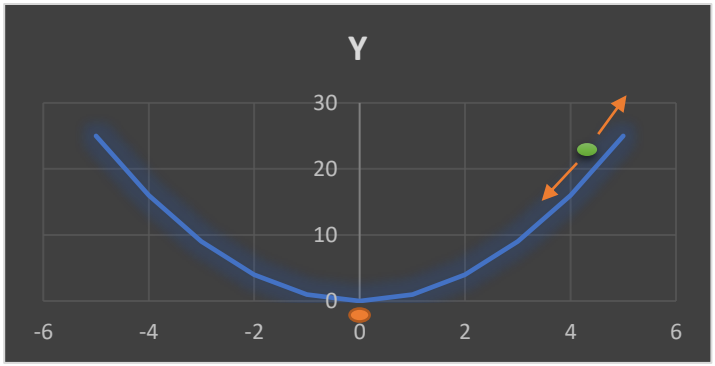
These condition check should be part of estimating regression coefficient. It these conditions are violated then we have to change experiment setup to fit the situation.

## 9. *Explain the gradient descent algorithm in detail.*

Gradient descent is an iterative optimization algorithm. In linear regression, it is used to optimize the cost function and find the values of the $\theta_1$ (estimators) corresponding to the optimized value of the cost function.

Consider we are walking down the graph below and we are at green point and our target is to reach to minima red point. We don't have any visibility so we can either go up or down.

Gradient descent helps here to decide how to reach to minima or have minimum cost function.




Mathematically, the aim of gradient descent for linear regression is to find the solution of ArgMin $J(\theta_0, \theta_1)$, where $J(\theta_0, \theta_1)$, is the cost function of the linear regression. It is given by —

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

Here, h is the linear hypothesis model, $h = \theta_0 + \theta_1 x$; y is the true output; and m is the number of data points in the training set.

We want to identify value of co-efficient $\theta_1$ of estimator, the equation will look like this:

$$\theta_1 = \theta_0 - \eta(\partial/\partial\theta)\, J(\theta)$$

Where η= the learning rate, which defines the speed at which we want to move towards negative of the gradient.

- ✓ It should be very small to reach solution at slow speed.
- ✓ A large value of learning rate may oscillate your solution, and you may skip the optimal solution (global minima)

For example:

- ✓ $J(\theta) = \theta^2$
- ✓ $(\partial/\partial\theta)\, J(\theta) = 2\theta$
- ✓ $\eta = 0.1$ (learning rate)
- ✓ $\theta_0 = 2$ (starting point on the curve)

*As per below table, you can verify that it took around ¬28 iterations to reach to minima = 0 with learning rate of 0.1.*

| $\theta_0$ | $(\partial/\partial\theta)\, J(\theta) = 2\theta$ | $\eta(\partial/\partial\theta)\, J(\theta)=0.1*2\theta$ | $\theta_1=\theta_0-\eta(\partial/\partial\theta)\, J(\theta)$ |
|---|---|---|---|
| 2.00 | 4.00 | 0.40 | 1.60 |
| 1.60 | 3.20 | 0.32 | 1.28 |
| 1.28 | 2.56 | 0.26 | 1.02 |
| 1.02 | 2.05 | 0.20 | 0.82 |
| 0.82 | 1.64 | 0.16 | 0.66 |
| 0.66 | 1.31 | 0.13 | 0.52 |
| 0.52 | 1.05 | 0.10 | 0.42 |
| 0.42 | 0.84 | 0.08 | 0.34 |
| 0.34 | 0.67 | 0.07 | 0.27 |
| 0.27 | 0.54 | 0.05 | 0.21 |

| | | | |
|---|---|---|---|
| 0.21 | 0.43 | 0.04 | 0.17 |
| 0.17 | 0.34 | 0.03 | 0.14 |
| 0.14 | 0.27 | 0.03 | 0.11 |
| 0.11 | 0.22 | 0.02 | 0.09 |
| 0.09 | 0.18 | 0.02 | 0.07 |
| 0.07 | 0.14 | 0.01 | 0.06 |
| 0.06 | 0.11 | 0.01 | 0.05 |
| 0.05 | 0.09 | 0.01 | 0.04 |
| 0.04 | 0.07 | 0.01 | 0.03 |
| 0.03 | 0.06 | 0.01 | 0.02 |
| 0.02 | 0.05 | 0.00 | 0.02 |
| 0.02 | 0.04 | 0.00 | 0.01 |
| 0.01 | 0.03 | 0.00 | 0.01 |
| 0.01 | 0.02 | 0.00 | 0.01 |
| 0.01 | 0.02 | 0.00 | 0.01 |
| 0.01 | 0.02 | 0.00 | 0.01 |
| 0.01 | 0.01 | 0.00 | 0.00 |

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

Quantile-quantile 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 datasets 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.

It is used to check following scenarios:
If two data sets --

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

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

Python:

statsmodels.api provide ***qqplot*** and ***qqplot_2samples*** to plot Q-Q graph for single and two different data sets respectively.