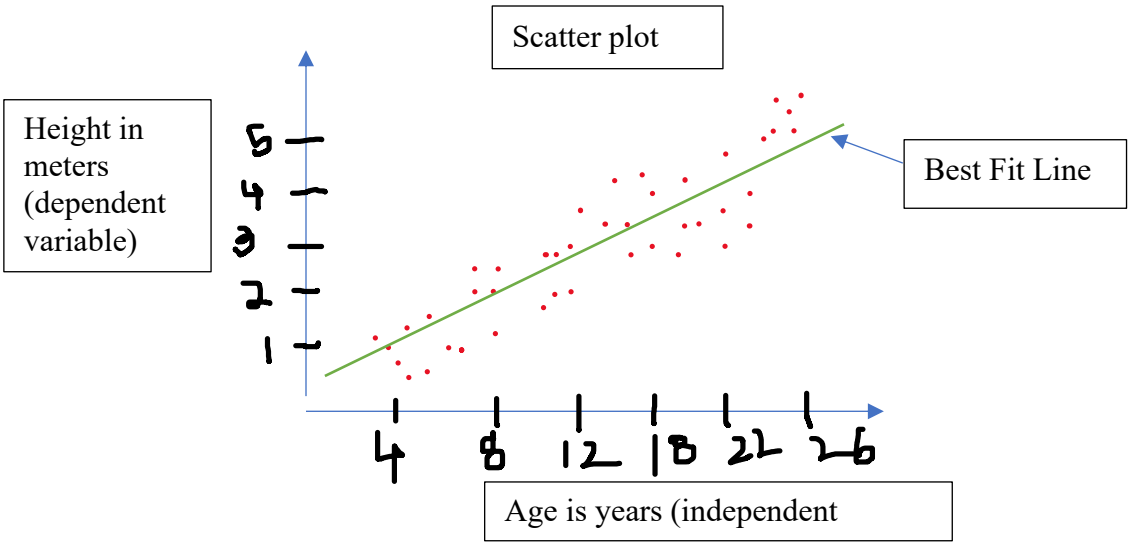
# 1. Explain the linear regression algorithm in detail.

Linear Regression is an algorithm that predicts the output of dependent variable using independent variable(s) by minimizing the error, thereby making the most accurate predictions
There are two types of Linear Regression Models:

1. Simple Linear Regression Model
2. Multiple Linear Regression Model

**Simple Linear Regression Model**

Simple Linear Regression Model explains the relationship between dependent variable and only one independent variable using a straight line. The straight line is a best fit line and it can be plotted on scatter plot
Lets consider an example of linear relationship between age and height




From above plot we can see that there is a positive relationship between independent variable age and dependent variable height. As the age increases, height of an individual increases.

The standard equation of Simple Linear Regression is:

$$Y = \beta 0 + \beta 1(X)$$

Where $\beta 0$ = intercept
And $\beta 1$ = slope
Y = dependent or target variable
X = independent variable

**Multiple Liner Regression**

Multiple Linear Regression Model explains the relationship between dependent variable and more than one independent variable. The main objective is to find a linear equation for different values of independent variable that can best determine the dependent or target variable

The standard equation of Multiple Linear Regression is:

$$Y = \beta0 + \beta1(X1) + \beta2(X2) + \beta3(X3) + \ldots\ldots\ldots\ldots.+\beta p(Xp)$$

Where $\beta0$ = intercept

$\beta1,\beta2\ldots\ldots,\beta p$ = coefficients

X1,X2…….,Xp = independent variables/ predictors

n = number of data points

Y = dependent or target variable

The whole method of Linear Regression model is to find a best fit line that passes through maximum data points and at the same time, residuals have to be monitored such that the model is not overfitted

**Residual**

The difference between the true value of $Y$ and the predicted value of $Y$ (denoted as $Y\hat{}$i) is known as a "residual." We calculate a residual as $Y$i $-Y\hat{}$i and denote each residual as e

Residual can be explained by using expression as:

$$e = Y\text{i} - Y\hat{}\text{i}$$

Say for Simple Linear Regression,

$$Y = \beta0 + \beta1(X1)$$

Where X1 = predictor

$$e = Y - (\beta0 + \beta1(X1))$$

There are some positive residuals and some negative residuals i.e. the predicted values are sometimes less or sometimes more than the actual values

Now if we add up all these residuals then the positive and negative amount might cancelled out and we wont get the proper sum of residuals. So to avoid that there is a term called residual sum of squares, where we square the residuals and sum all these square of residuals which is termed as residual sum of squares

For residual sum of squares,

$$RSS = e_1^2 + e_2^2 + e_3^2 + \ldots\ldots\ldots + e_n^2$$

Where n = number of observations/ data points

e = residual between actual and predicted values

Expanding error terms,

$$RSS = [Y1 - (\beta 0 + \beta 1(X1)]^2 + [Y2 - (\beta 0 + \beta 1(X2)]^2 + \ldots\ldots\ldots + [Yn - (\beta 0 + \beta n(Xn)]^2$$

In General terms for multiple linear regression with n datapoints

$$RSS = \sum_{i=1}^{n} (Yi - \hat{Y}i)^2$$

This is also known as loss function, where the "loss" here is the sum of squared errors
The most common way to define the

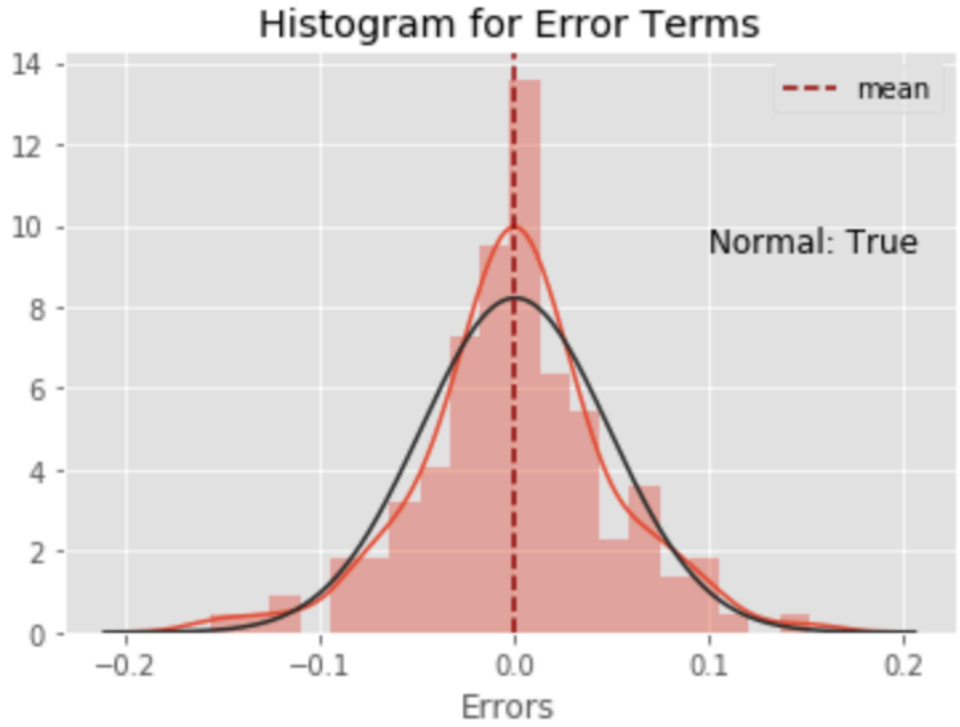
**Best Fit Line**

Best Fit Line is formed by minimizing the expression of RSS (Residual Sum of Squares) which is sum of squares of residuals of each data point in the plot
The best way to define the best fit line is such that residuals is close to zero as much as possible

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

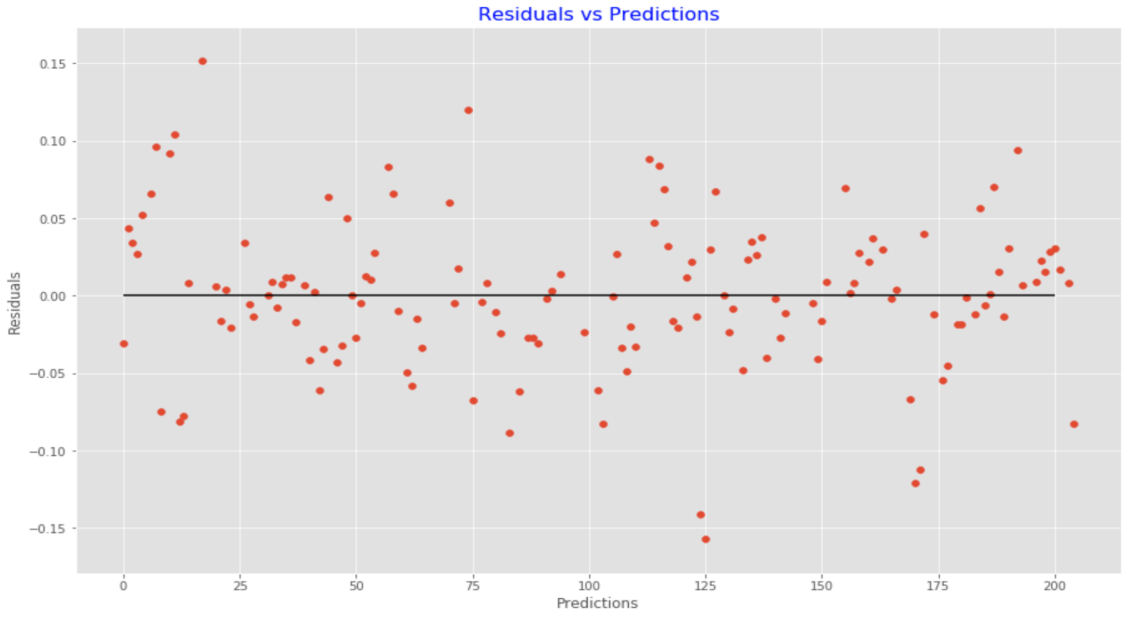
Assumptions of linear regression model regarding errors or residuals are:

1. **Independence**: Error terms must be independent of each other. In other words, if ei and ej are two error terms then ei and ej must be independent of one another given i is not equal to j
2. **Normality**: The residuals follow a Normal distribution with mean of 0
   It should be something like below:




3. **<u>Equality of Variance</u>**: Also known as homoscedasticity of errors. The residuals or errors should have a roughly consistent pattern, regardless of the value of X. There should be no relationship between X and the residuals. The residuals should have constant variance

   The pattern should be something like below:



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

**Coefficient correlation**

1. Coefficient of correlation is the linear correlation coefficient denoted as 'r'. It measures the strength and direction of relationship between two variables.
2. The value of r lies in the range of -1 to 1. Positive value of r means there is a positive relationship between variables (say x and y) and negative value of r means there is a negative relationship between variables (say x and y). Value of 0 means there is no correlation between the two variables

3. The mathematical formula for computing r is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where n is the number of datapoints and x and y are two variables

To derive this formula, there are various steps:

1. We have pair of variable x and y and there are n number of data points so at any ith point, the pair of variable is denoted as ($x_i$, $y_i$)
2. For calculations,
    a. calculate $\bar{x}$, the mean of all data points of $x_i$
    b. calculate $\bar{y}$, the mean of all data points of $y_i$
    c. calculate $s_x$, sample standard deviation of all data points of $x_i$
    d. calculate $s_y$, sample standard deviation of all data points of $y_i$
    e. Use the formula, $(z_x)_i = (x_i - \bar{x})/ s_x$ and calculate standardized value for each $x_i$
    f. Use the formula, $(z_y)_i = (y_i - \bar{y})/ s_y$ and calculate standardized value for each $y_i$
    g. Multiply corresponding standardized values: $(z_x)_i(z_y)_i$
    h. Do this for all data points and sum up the products
    i. Divide this sum by (n-1), where n is the number of data points

Below matrix can show the coefficient correlation between any two variables at a time

| | price | wheelbase | carlength | carwidth | carheight | curbweight | enginesize | boreratio | stroke |
|---|---|---|---|---|---|---|---|---|---|
| **price** | 1 | 0.577816 | 0.68292 | 0.759325 | 0.119336 | 0.835305 | 0.874145 | 0.553173 | 0.0794431 |
| **wheelbase** | 0.577816 | 1 | 0.874587 | 0.795144 | 0.589435 | 0.776386 | 0.569329 | 0.48875 | 0.160959 |
| **carlength** | 0.68292 | 0.874587 | 1 | 0.841118 | 0.491029 | 0.877728 | 0.68336 | 0.606454 | 0.129533 |
| **carwidth** | 0.759325 | 0.795144 | 0.841118 | 1 | 0.27921 | 0.867032 | 0.735433 | 0.55915 | 0.182942 |
| **carheight** | 0.119336 | 0.589435 | 0.491029 | 0.27921 | 1 | 0.295572 | 0.0671487 | 0.171071 | -0.0553067 |
| **curbweight** | 0.835305 | 0.776386 | 0.877728 | 0.867032 | 0.295572 | 1 | 0.850594 | 0.64848 | 0.16879 |
| **enginesize** | 0.874145 | 0.569329 | 0.68336 | 0.735433 | 0.0671487 | 0.850594 | 1 | 0.583774 | 0.203129 |
| **boreratio** | 0.553173 | 0.48875 | 0.606454 | 0.55915 | 0.171071 | 0.64848 | 0.583774 | 1 | -0.055909 |
| **stroke** | 0.0794431 | 0.160959 | 0.129533 | 0.182942 | -0.0553067 | 0.16879 | 0.203129 | -0.055909 | 1 |

**Coefficient of determination**

1. Coefficient of determination is also denoted as r-squared. We can multiply r times r to get r-squared.
2. It shows the percentage variation in y (dependent variable) which is explained by all other independent variables. It's the ratio of explained variation to the total variation
3. Its always between the range of 0 and 1. Closer to 1 is a good value of r-squared.
4. The coefficient of determination is a measure how well the regression line represents the data. If the regression line passes through most of the points, it would be able to explain most of the variation. The further the line away from points, the less it is able to explain

   Like in below example, 93.7% of the variation in y can be explained by the model

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | price | **R-squared:** | 0.937 |
| **Model:** | OLS | **Adj. R-squared:** | 0.934 |
| **Method:** | Least Squares | **F-statistic:** | 288.3 |
| **Date:** | Fri, 17 Jan 2020 | **Prob (F-statistic):** | 6.65e-78 |
| **Time:** | 14:15:52 | **Log-Likelihood:** | 229.76 |
| **No. Observations:** | 143 | **AIC:** | -443.5 |
| **Df Residuals:** | 135 | **BIC:** | -419.8 |
| **Df Model:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

For simple linear regression, correlation can very well explain the relationship between x and y. For multiple linear regression, because there are multiple variables involved so r-squared is a better term

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

Anscombe's quartet comprises of 4 dataset that have nearly identical descriptive summary, eg.mean, standard deviation etc.

The data set is as below:

**Anscombe's quartet**

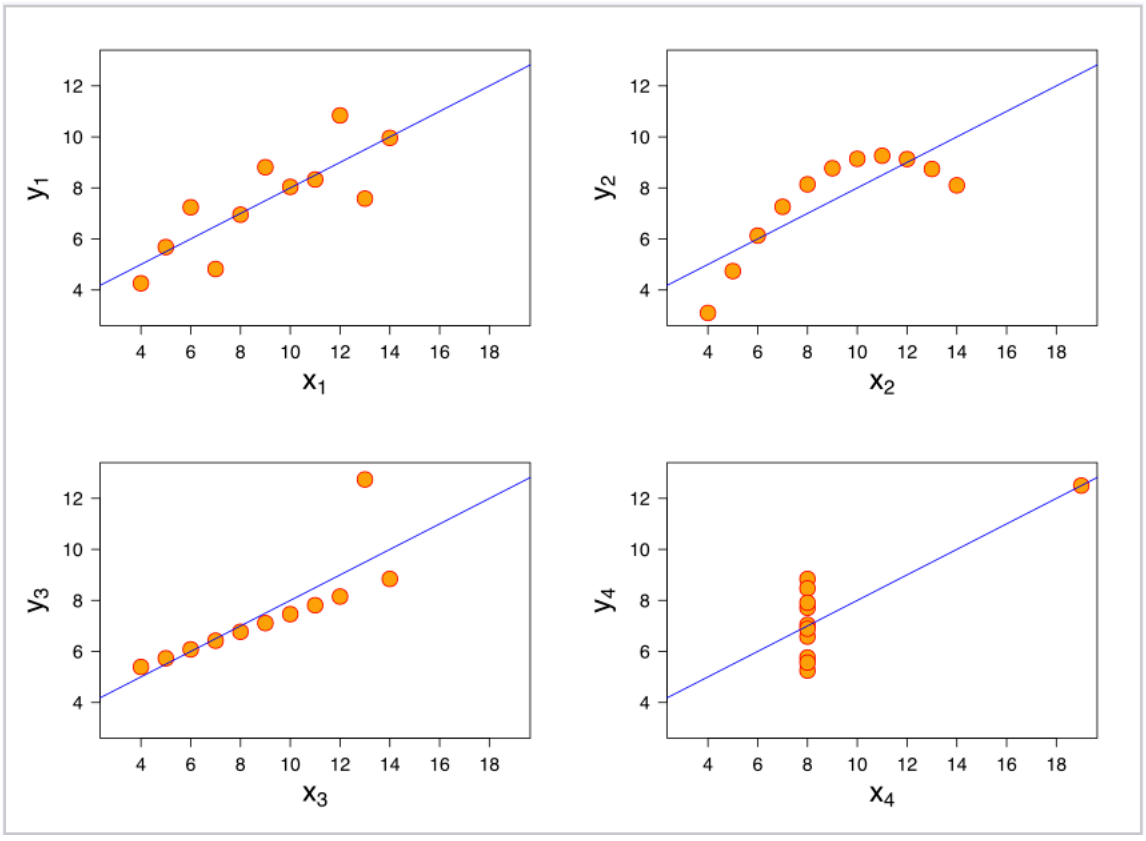
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

Statistical summary:

For all four datasets:

| Property | Value | Accuracy |
|---|---|---|
| Mean of $x$ | 9 | exact |
| Sample variance of $x$ | 11 | exact |
| Mean of $y$ | 7.50 | to 2 decimal places |
| Sample variance of $y$ | 4.125 | ±0.003 |
| Correlation between $x$ and $y$ | 0.816 | to 3 decimal places |
| Linear regression line | $y = 3.00 + 0.500x$ | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression | 0.67 | to 2 decimal places |

But if we plot it using graphs then we get as below:



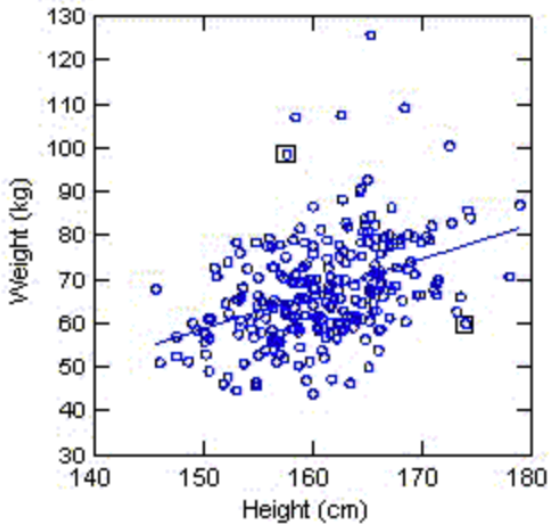
From above graph we can see that even though the statistical summary of four datasets is same but the relationship of variables is completely different

1. 1st scatter plot tells that there is a simple linear relationship between x and y
2. 2nd plot tells that the relationship is not linear and instead of using Pearson correlation between two variable to find correlation we should use coefficient of determination. The relationship is polynomial
3. The 3rd plot tells us that there is linear relationship between x and y however there is an existence of outlier in the dataset. The calculated regression has been impacted due to this outlier. We should treat this outler and should model again and then replot
4. The 4th plot tells us that one high-leverage point (at the top right extreme) is enough to determine the relationship between two variables whereas the other data points do not indicate any relationship between the variables

Anscombe's quartet tells us that graphing data prior to analysis is a good practice. Outliers should be treated well before analysing data and statistics of the dataset do not fully depict the dataset in its entirety

## 5. What is Pearson's R?

1. Pearson's R is the numerical summary of the strength of linear relationship between the variables.
2. Its value is within the range of -1 and +1
3. Negative value means there is a negative relationship between the variables , positive value means there is positive relationship between variables, zero value means there is no linear relationship between the variables
4. Please note that the relationship is "on average", not for any arbitrary pair of observation
   Consider the plot below – relationship between height and weight of people:



This scatter plot depicts that there is a positive relationship between the variables height and weight. Yet there are points where for taller height the weight is less

5. As the correlation coefficient increases in value, the data points are more aligned towards the best fit line or in other words for higher correlation coefficient we can say that straight best fit line passes through most of the data points
6. The formula for Pearson correlation coefficient when applied to population is represented as :

   $\rho X,Y = cov(X,Y) / (\sigma X \sigma Y)$

   Where:
   cov is the covariance
   $\sigma X$ is the Standard Deviation of X
   $\sigma Y$ is the Standard Deviation of Y

   Pearson correlation coefficient when applied to sample is represented as:

   $$r_{xy} = \sum x_i y_i - n\bar{x}\bar{y} \,/\, \left(\sqrt{(\sum x_i^2 - n\bar{x}^2)} \quad \sqrt{(\sum y_i^2 - n\bar{y}^2)}\right)$$

7. The t-test is used to find if the correlation coefficient is significantly different from zero and that there is an association between two variables. The underlying assumption is that the data is from a normal distribution sampled randomly. If this is not true then its better to use Spearman's coefficient rank correlation

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

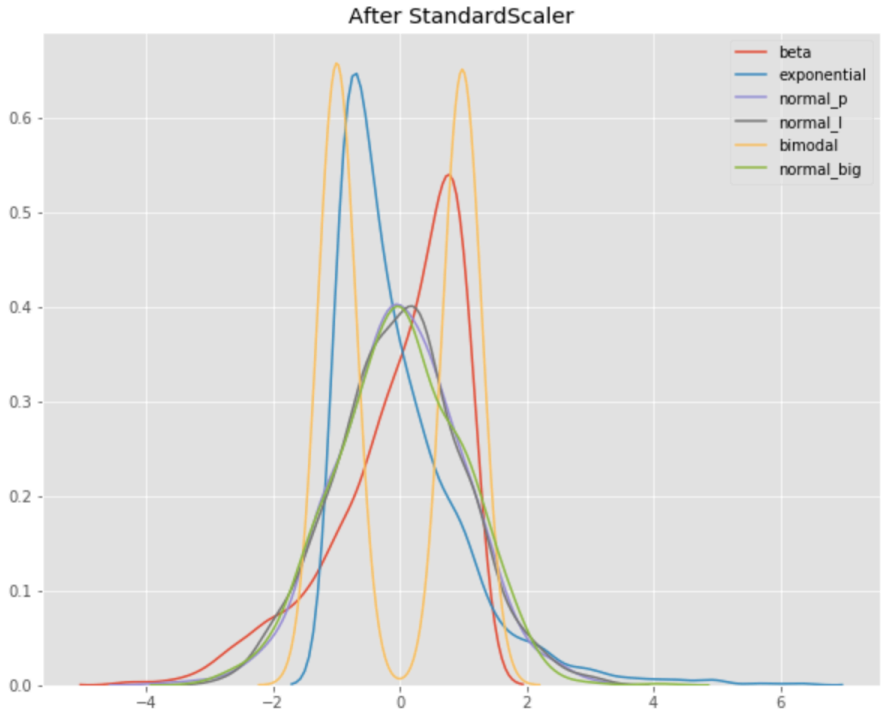
Feature scaling is a method to bring all feature values at one scale. In any dataset, it can be that many independent variables have very different scale. The machine learning algorithm will then weight greater values with higher coefficients and smaller values with smaller coefficients which is not fair and correct for the model

Scaling is needed for:
1. Ease of interpretation
2. Faster convergence of gradient descent methods

1. **<u>Standardized Scaling (Standard Scalar)</u>**

a. Standardization is a technique to change the scale of feature values between -1 and 1
b. Standard Scalar is one such Scaling method that standardizes the feature values by subtracting the mean and then dividing all the values by standard deviation.
c. With Standard Scalar, standard deviation is 1, mean is 0 and all the values lie within the range of -1 and 1

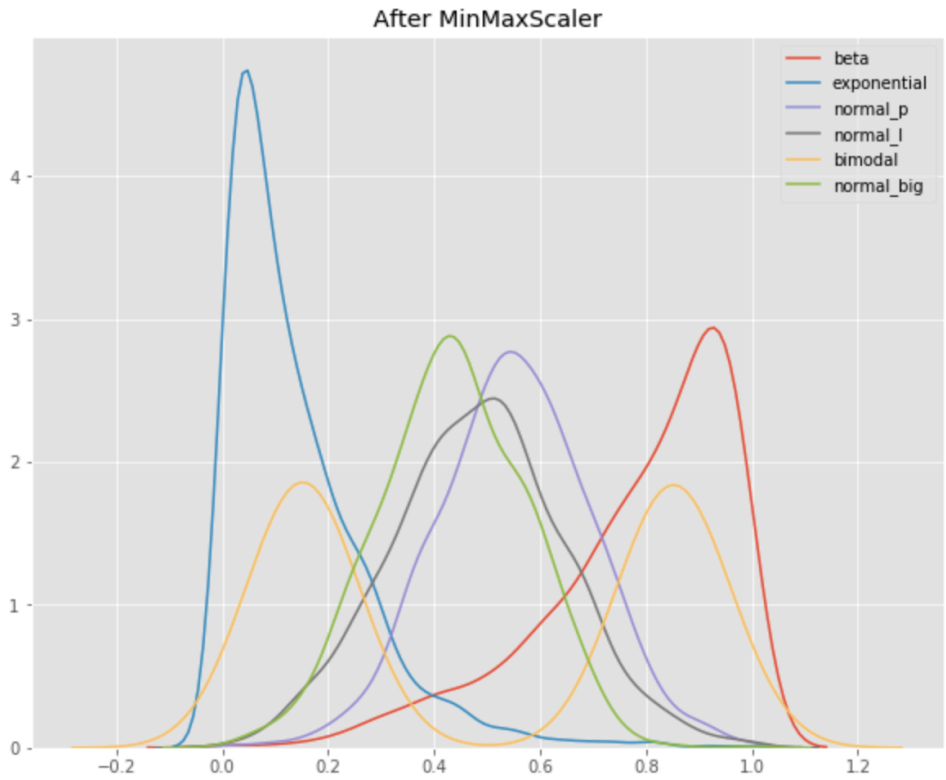
$$X_{changed} = (X - \mu) / \sigma$$




**2. <u>Normalized Scaling (MinMax Scalar)</u>**

a. Normalization is technique to change the scale of feature values between 0 and 1
b. MinMax Scalar is one such scaling method of Normalized Scaling
c. For each feature values, MinMax Scalar subtracts the minimum value in the feature and then divides the value by the difference between original maximum and original minimum
d. The values are within the range of 0 and 1
e. It preserves the shape of original distribution, however the outliers from the data set are lost

$$X_{changed} = (X - X_{min}) / (X_{max} - X_{min})$$




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

1. VIF i.e. Variance Inflation Factor is an index that provides the measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, regression model is fitted between the independent variables.

2. VIF is calculated as:

   VIF = 1/ (1-$R^2$)
   Where $R^2$ is the coefficient of determination

3. If all the independent variables are orthogonal to each other i.e. there is no relationship then $R^2$ = 0 and therefore VIF = 1
4. If there is a perfect correlation then VIF = infinity

   Example

### Checking VIF

```
In [6309]:  1  check_vif(X_train_sm)
```

Out[6309]:

| | Features | VIF |
|---|---|---|
| 10 | fueltype_gas | inf |
| 18 | fuelsystem_idi | inf |
| 6 | compressionratio | 62.06 |
| 3 | enginesize | 14.41 |
| 2 | curbweight | 12.90 |
| 7 | horsepower | 12.05 |

Above example, it means that fueltype_gas feature can be very well explained by all other features in the model

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

Gauss-Markov theorem states that given assumptions of linear regression is true, OLS (Ordinary Least Square) estimator will have following properties:

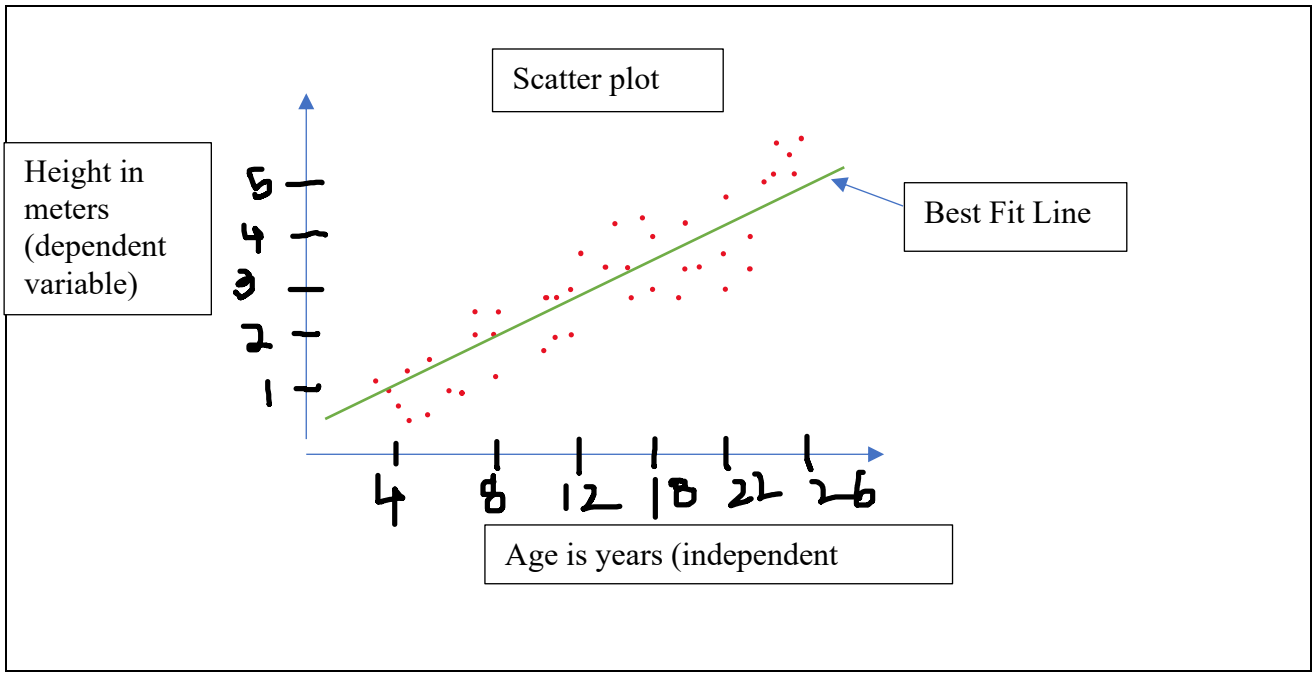
1. Unbiased
2. Minimum Variance
3. Consistent
4. Normally distributed

Such estimator is called as BLUE – Best Linear Unbiased estimate

Assumptions of linear regression:

1. There is linear relationship between X and Y
2. Error terms are normally distributed (not X and Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Consider it with an example for a relationship between age and height.

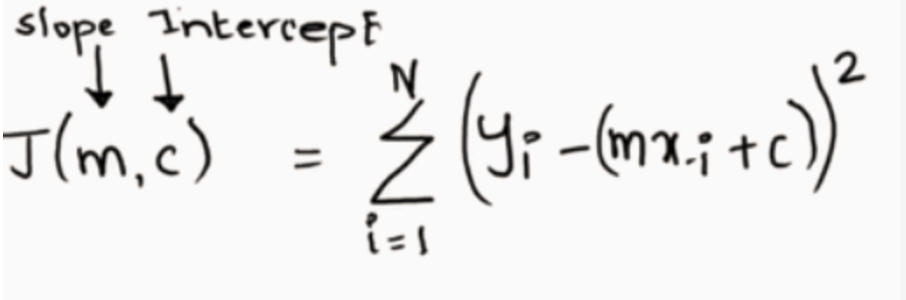



Some people are tall, some are small but somewhere there is an average across all people of given age. The average of data as we move across the different ages forms a line in scatter plot. The Gauss-Markov theorem states that if people generally have about the same variance in height at each age, then the 'ordinary least squares' estimate of the mean would give you the best unbiased estimate of how people's height varies with age on average.

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

Gradient descent is:

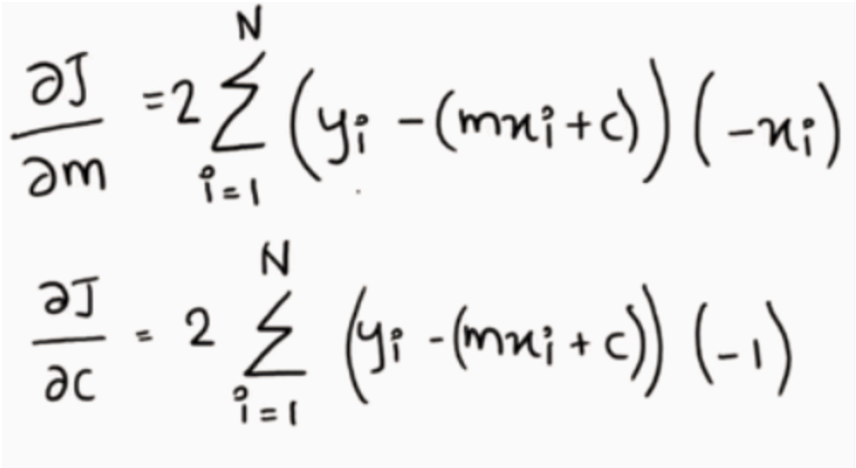
1. an iterative method which is used to identify the optimal value of parameters by optimizing an objective function.
2. It works by:
   a. Start by making a guess for the optimal parameter value i.e. instantiate the model.
   b. Calculate the loss given that learning rate and instantiated value.
   c. Update guess to decrease loss.
   d. Keep going until loss is "sufficiently minimized."
3. Cost function known as Mean Squared Error (MSE) is defined as:




**Cost Function**

where m = slope
c = intercept
N = number of observations

Now gradient descent is an optimization algorithm which is used to coefficients of function that minimizes the cost function.
The derivative of above defined cost function can be defined as below, derivative of m and c is:

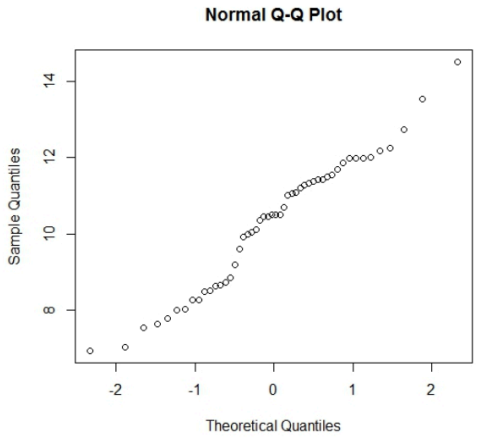



**Derivatives**

This will be done in iterations, and m and c are updated at each iteration till the cost function is minimized
Learning rate controls the step taken in downward direction in each iteration

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

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, then we there will be a roughly straight line between the points.
Below is an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions



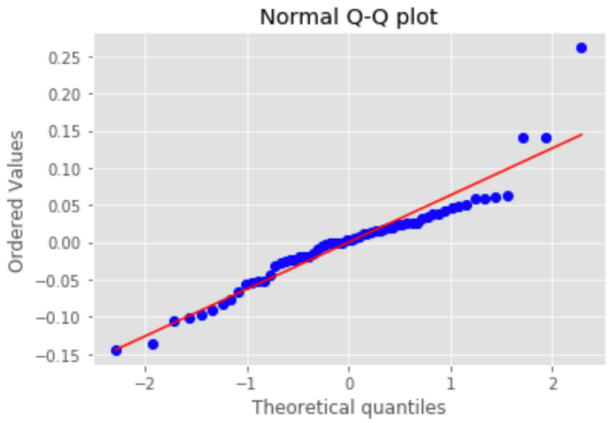

Q-Q plot help us assess if a set of data came from theoretical distribution such as Normal or exponential

Q-Q plots take sample data, then sort the data in ascending order and then plot them versus quantiles calculated from theoretical distribution. The number of quartiles depend upon the sample data size.

**Example**

For instance, below is the normal distribution of error terms and can be seen from Q-Q plot that the error terms distribution is normal. This is because the data points closely follow the straight line

```
stats.probplot(x = (y_test_scaled - y_test_pred)['price'], dist="norm", plot=plt)
plt.title("Normal Q-Q plot")
plt.show()
```




For instance, below is not the normal distribution of predicted test values of feature price and same can be seen from Q-Q plot as well. This is because the data points are not closely following the straight line

```
stats.probplot(x = y_test_pred['price'], dist="norm", plot=plt)
plt.title("Normal Q-Q plot")
plt.show()
```