*** Understanding Linear Regression:**

Points to be covered:

- **What is Linear Regression?**
- **What are the types of Linear Regression.**
- **What are th eassumptions made in Linear Regression?**
- **How does the Slop, Intercept and Error is calculated in Linear Regression?**
- **How does Linear Regression works?**

- **What is Linear Regression?**

Linear Regression is a statistical supervised learning technique used to predict the continuous (dependent) variable given a set of independent variable. For predicting continuous variable it forms a linear relationship with one or more independent variables.

Mathematically, Linear Regression can be presented as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where, y = Dependent Variable,

x = Independent Variable,

β_0 = Intercept

β_1 = Slope

ε = Error

- **Types of Linear Regression:**

1) Simple Linear Regression:

Simple Linear Regression helps to find the relationship between one continuous/dependent variable and one independent variable. It uses linear equetion to form the relationship between two variables.

It is represented as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where, y = Dependent Variable,

x = Independent Variable,

β_0 = Intercept

β_1 = Slope

ε = Error

2) Multiple Linear Regression:

Multiple Linear Regression is used to form a relationship between two or mode independent variable and one continuous (dependent variable). Independent variables can be continuous or categorical.

Multiple Linear Regression can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2_2 + \beta_3 x_3 + ... \beta_n x_n + \varepsilon$$

where, y = Dependent Variable,

x_1 to x_n = Independent Variables,

β_0 = Intercept

β_1 to β_n = Slope coefficient for each independent variable

ε = Error

- **What are th eassumptions made in Linear Regression?**

Before moving further, we need to know some assumptions of linear regression. Linear regression gives better results if our dataset follows these assumtions.

1. The relationship between independent variable and dependent variable should be linear and additive.
2. The independent variables should not be correlated to each other. The correlation in independent variables leads to multicollinearity.
3. Variance between error terms should be constant. If the variance between error terms is not constant then it leads to heteroskedesticity.
4. The error terms should not be correlated i.e. error at *epsilon t must not* be correlated to error at (et+1).
5. The error terms and dependent variables must be normally distributed.

- The performance of linear regression model is dependent upon fulfilment of these assumptions.

- **How does the Slop, Intercept and Error is calculated in Linear Regression?**
  We will understand the concepts of Slope, Intercept and Error with the help of Simple Linear Regression.

1) What is slope?

In regression, slope is a coefficient which tells us about amount change in y with respect to change in x.

It is denoted as β_0 in linear regression formula

$y = \beta_0 + \beta_1 x + \varepsilon$

and this can be calculated as:

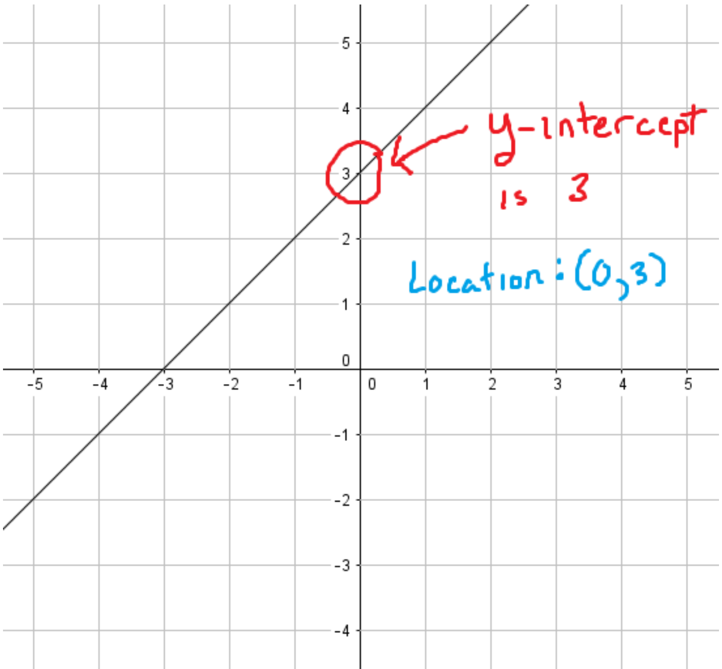
$$m = \frac{\sum_{i=1}^{n}(x_i - x_{mean})(y_i - y_{mean})}{\sum_{i=1}^{n}(x_i - x_{mean})^2}$$

2) What is intercept?

In linear regression formula y = ax+b, the intercept is denoted as b. It is the point of intercection where the regression line (y=ax+b) crosses the y axis.
This can be calculated as:

$\beta_0 = y_{mean} - \beta_1(x_{mean})$

After calculating the slope and intercept we can put them into equation y = beta0+beta1x+error to get a regression line as:



3) What is an error?

The error in linear regression is defined as:
‘The error is the difference between actual y value and predicted y value.’
In prediction process of regression error is inevitable part and we can not completly eliminate it. So, we try to reduce it to the lowest value. To reduce the error regression uses Ordinary Least Square (OLS) method. In this method OLS tries to reduce sum of squared errors - ∑[Actual y – Predicted y]**2 by determining possible values for regression coefficients (slope and intercept)

The estimation of errors in OLS method can be divided into three parts:

1) Resudual Sum of Square (RSS):

RSS = ∑[Actual(y) – Predicted(y)]^2

2) Explained Sum of Square(ESS):

ESS = ∑[Predicted(y) – Mean(ymean)]^2

3) Total Sum of Square(TSS):

TSS = ∑[Actual(y) - Mean(ymean)]^2

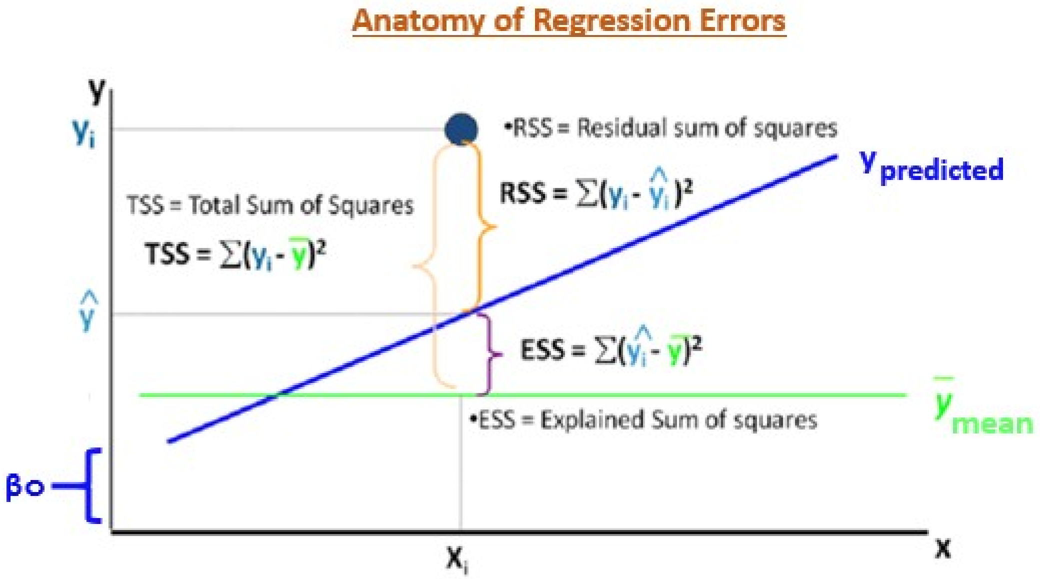

Anatomy of Regression Errors
y
y_i
•RSS = Residual sum of squares
$y_{predicted}$
RSS = $\Sigma(y_i - \hat{y}_i)^2$
TSS = Total Sum of Squares
TSS = $\Sigma(y_i - \bar{y})^2$
$\hat{y}$
ESS = $\Sigma(\hat{y}_i - \bar{y})^2$
$\bar{y}_{mean}$
•ESS = Explained Sum of squares
βo
X_i
x

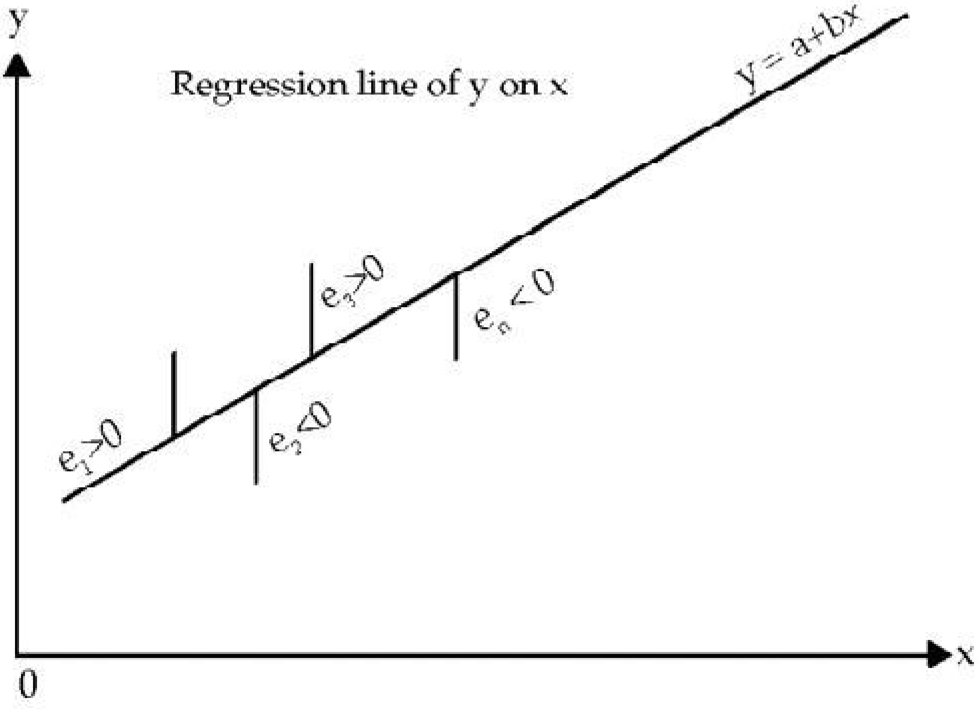
Figure from Hackerearth.com

These parts of error estimation are have important role in calculation of Coefficient of Determination (r squared).

$$R2 = 1 - \left(\frac{SSE}{TSS}\right)$$

(Coefficient of Determination: The coefficient of Determination is used to explain how much variability of one factor can caused by its relationship to another factor. Ref: Investopedia.com)

- **How does Linear Regression works?**
  The processof linear regression is to find out the Least Sqaure Regression Line (LSRL).
  The Least Sqaure Regression Line is the best fit line with having very low error value.
  LSRL described with the equation y = mx+b where b is intercept, m is slope.
  Properties of Regression Line:
  - The line minimizes the error i.e. (Actual y – Predicted y)
  - The best fit line passes through the mean of independent and dependent variables.


y
Regression line of y on x
y = a+bx
$e_1>0$
$e_2<0$
$e_3>0$
$e_n<0$
0
x