Till now we have learned various concepts of simple linear regression. In simple linear regression, we talked about the linear relationship between an independent variable with a dependent variable.

But, here arises one question

What if we have more than one independent (explanatory) variable?

This question is stuck in your mind too.

No need to worry, here is the answer.

*When we have more than one independent variable to predict any dependent variable, having linear relationships we use* ***multiple regression*** *techniques.*

***Multiple linear regression explains the relationship between multiple independent variables with one dependent or predicting variable.***

Thus the model would be

**Y = b0+b1 x1+b2 x2+b3x3 +…..+bn xn +e**

Where xI ‘s are explanatory variables and bi’s are coefficients of regression.

Note that the j regression coefficient b j represents the expected change in y per unit change in the independent variable X . Assuming E( e)= 0,

**bj =d (E(y))/d xj**

For example, a biostatistician wants to predict the height of a child, various factors affect the height of a child including various environmental factors and nutritional factors.

But when we use the multiple regression model we need to consider some important assumptions.

1. *The relationship between the dependent and independent variables is linear.*
2. *Independents variables are not highly correlated with each other.*
3. *The residual term is distributed normally.*
4. *The size of the error remains the same across the different values of an independent variable. That is data is homoscedastic in nature.*

*(homoscedastic – when the variance of error term remain constant)*

So, now we know what to do when we have multiple explanatory variables.

Ohh, wait before ending ….

I want to show some light on a very important topic …..

We have a very important assumption that independent variables are not highly correlated but in many cases, there is the possibility that the two independent variables are highly correlated with each other. This is a violation of our assumption, called Multicollinearity.

Do we need to be afraid of multicollinearity?

Due to multicollinearity, the estimators of regression coefficients remain unbiased but the variance of the estimators becomes large. Thus, it will not give precise results hence we would not be able to determine how different explanatory variables influence dependent variables individually.

How to deal with this??

Increase the size of the data.(If possible)

Drop some variables that are collinear.

Use prior information regarding the coffecients.

Happy learning 😊