Introduction to Linear Regression

When you think of [**Regression**](javascript:glosItem%20=%20'regression';void%20openURL('../help/glossary.html','glossary',640,480)), think prediction. A regression uses the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable.

**Application of Linear Regression**

Businesses use regression to predict such things as future sales, stock prices, currency exchange rates, and productivity gains resulting from a training program

**Types of Regressions**

A regression models the past relationship between variables to predict their future behaviour. As an example. How can we formally test that there is a relationship between Wages and education spend in years. More importantly how can we expect our wage to increase in every year spent on our education i.e is it even worth of studying in high school.

The **dependent** variable in this instance is Wages and the **independent** variable is Education.

Usually, more than one independent variable influences the dependent variable. You can imagine in the above example that Wages are influenced by Education, also if we include other factors as well, such as age, gender, work experience, and sector. When one independent variable is used in a regression, it is called a [**simple regression**](javascript:glosItem%20=%20'simple_regression';void%20openURL('../help/glossary.html','glossary',640,480)); when two or more independent variables are used, it is called a [**multiple regression**](javascript:glosItem%20=%20'multiple_regression';void%20openURL('../help/glossary.html','glossary',640,480)).

The general formula for simple and multiple linear regression is given as:

Simple linear regression:

Wages(dependent variable) = (Y-Intercept) + Education(Independent Variable)

Y= β o + β 1X

Multiple regression equation:

Wages(dependent variable) = (Y-Intercept) + (Education) + (age) + (Gender) + (Work Experience) + (Sector)

Y = β 0 + β 1X1 + β 2X2 + β 3X3 + β 4X4 + β 5X5

So the best way to know the relationship between independent and dependent variable is by scatter plot.

**Scatter plot:**

Consider an above example of wages and education:



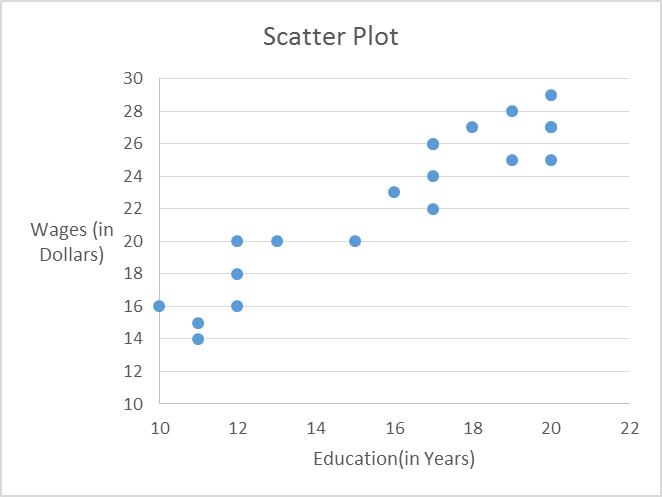
Let us consider data of 20 professionals of their years of education and Wages in dollars per hour.

**Note : Make sure the collected data is a representation of the population.**

In statistics we must ensure that our sample of individuals must represents our population. That means we must ensure the random sampling, this will allow us the make the inferences of our population at large.

So to represent the above individuals on their Wages and Education, the best way is the

**Scatter plot:**



This Scatter plot allowed us to accommodate all the individuals with their wages and years in Education.

Now to know the relationship between our variables or the pattern between them we use the **line of best fit.**

The line of Best fit is the line which represents the general pattern of the sample. **A regression line** is simply the line of best fit for a given sample.

Now we know that the equation of line is

**Y=mx + c**

Where m =slope

C= intercept of the line.

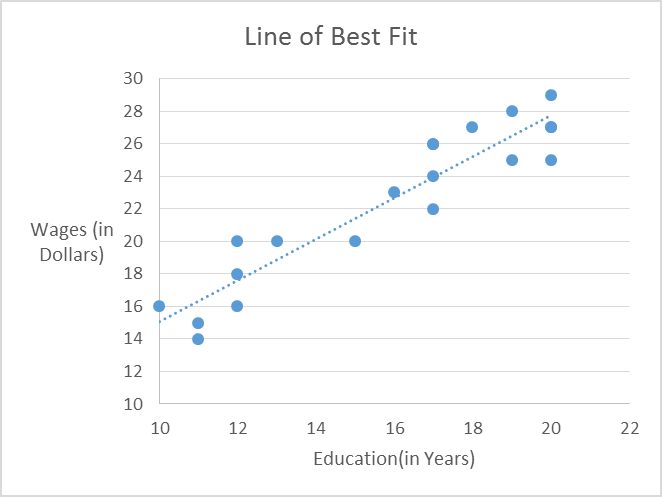
In regression analysis we represent the best fit line with

**Y=** β **0 +** β **1X**

(Pronounced as Beta not) β 0= Intercept

(Pronounced as Beta one) β 1=Slope of the line

Here Y= Wages and X = Education



β 0

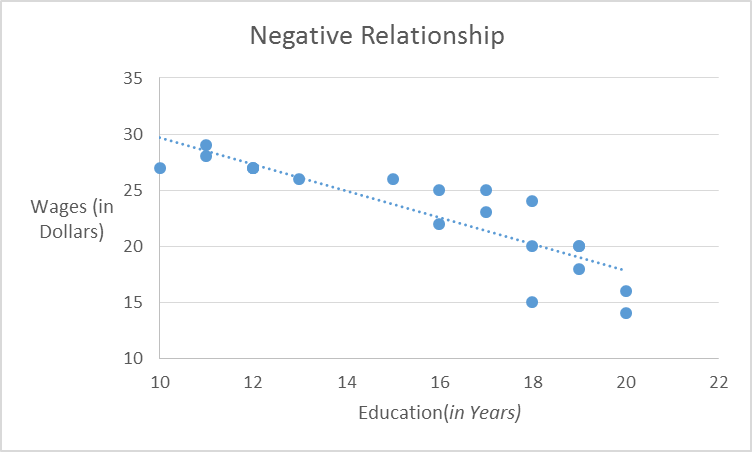
β1

So **Y=** β **0 +** β **1X**

Wages = β 0 + β1(Education)

so if β1 >0 it has a **positive relationship.**

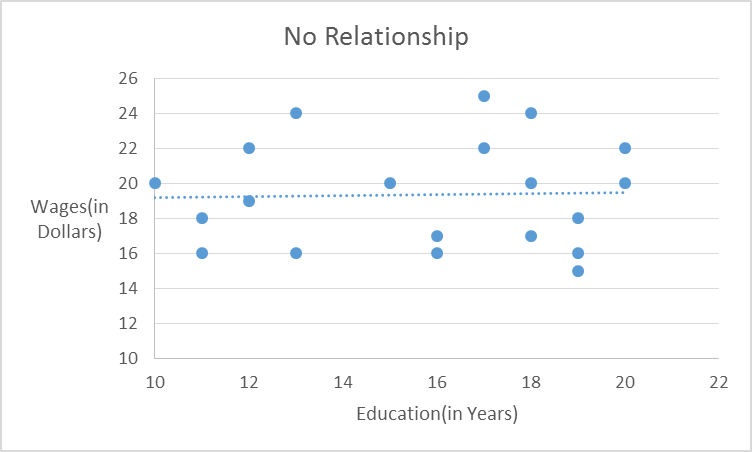
The above Shows the positive relationship between Wages and Education. The more Education a person attains the higher the wage it gets.



If β1 <0 it has a **negative relationship**. The regression line is in a downward direction.

There is an negative relationship between the Wages and Education. It has a general trend that the more educated is any individual the less pay they would get.

In this case the slope of regression line β1 is negative.



If β1 =0 it has a **No relationship**. The regression line is in a Straight direction.

There may be no relationship between Wages and Education. The Slope of the regression line β1 is zero.

**Estimation of regression line:**

Let suppose we get an estimated regression line as:

**Y=2.372 + 1.267x**

Means: Wages = 2.372 + 1.267(Education)

This means that the line cuts the Y-Axis at 2.372 (Dollars) and slope of the line is 1.267 (in Years)

**Now lets make a prediction:**

Suppose that for a Professional who is having an work experience of 12 years and we wanted to know about the wage of that person per hour in dollars then we simply replace x by 12 in the above equation as:

Wages = 2.372+1.267\*12

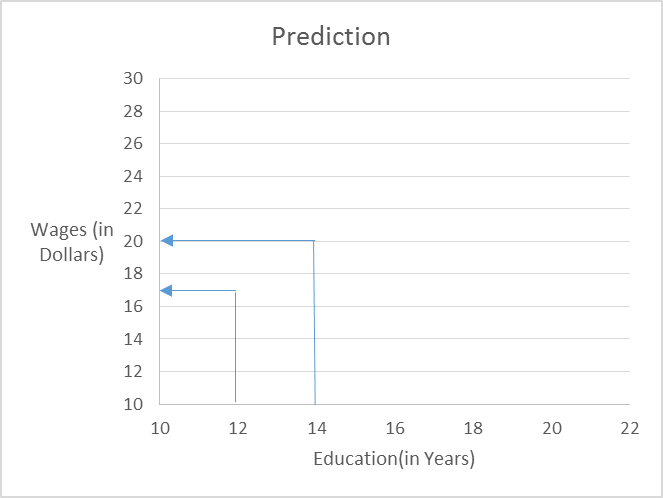
Wages = $17.57 per hour

**Lets take another example:**

To know about the Wage of a person who is having a 14 years of Education.

Wages = 2.372 + 1.267\*14

Wages = $20.11 Per hour



$17.57

$20.11

**Inference from Prediction:**

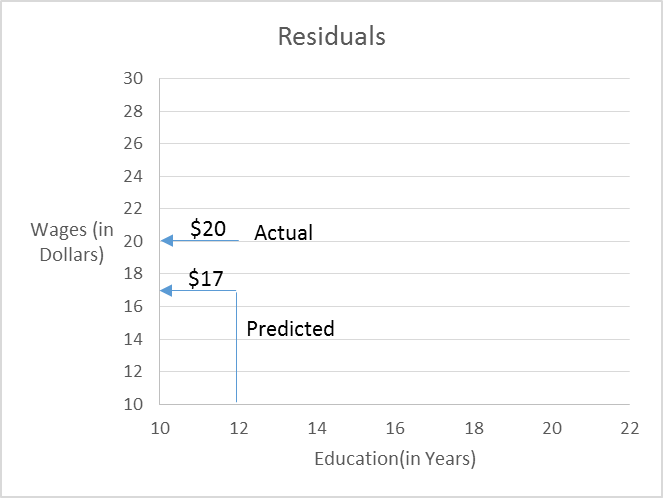
1) This means that for every 1 year addition of education the wages is expected to increase by $1.5 approx.

2) When education is Zero i.e (β 1=0) , the Wages is expected to be $2.372 per hour.

**Residuals:**

Residuals are the difference between the actual value and the predicted value.

Suppose as per our predictions, the wage for a professional who has a 12 years of education(Let say #11 from table) which is $17 per hour. Actually the wage of that professional is $20 per hour.



So difference between the actual and the predicted wages which is $3 are the residuals.

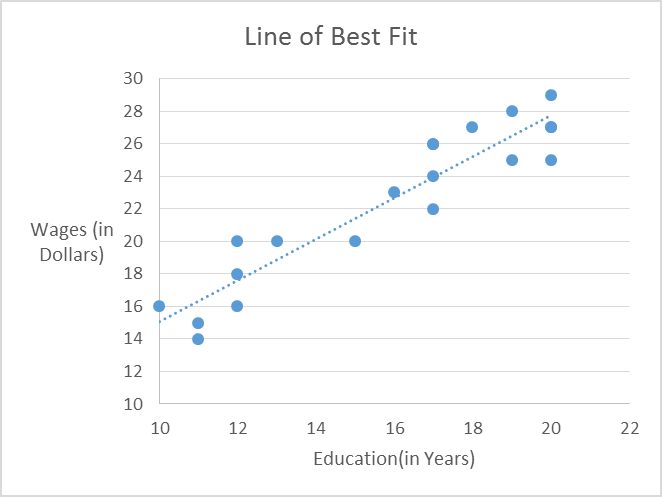
Thus Residuals =Actual Value- Predicted Value

Residuals =$20-$17 = $3

So Residuals are the other factors which doesn’t include into the regression equation. These are the factors that does have an effects on the wages but not contained into the model.

**Wages = β0 + β1(Education) + µ(Residuals)**

**Regression in R:**



**Linear Relationship-** The predictor (Xs) variables and dependent variables have a linear relationship. This can be easily verified by plotting the X vs Y in scatter plot. Here X vs Y plot is between Education and Wages. If the linear relationship doesn’t exist, either the variables need to be transformed or some other technique should be used.

In other words, the best fit line (as explained above) tends to return most accurate value of Y based on X  i.e. it evaluates a minimum difference between actual and predicted value of Y (lower prediction error).

**Methods to check for the errors:**

* Sum of all errors (∑error)
* Sum of absolute value of all errors (∑|error|)
* Sum of square of all errors (∑error^2)

**Let us now consider the above Regression equation as:**

Y=2.372+1.267x

Here Y = Wages and X= Education

We now have to calculate the Error terms which tells us about the difference of variability between the actual value of wages and predicted value of wages



**Sum of Errors:**

Sum of errors simply tells about the exact deviation between the predicted and the actual model values.

**Sum of Absolute Errors(SAE):**

The SAE is a more robust alternative to the popular least squares regression whenever there are outliers in the values of the response variable, or the errors follow a long tailed distribution, or the loss function is proportional to the absolute errors rather than their squared values.

**Sum of Square Error(SSE):**

It is also known as Residual sum of squares(RSS): It is the measure of the sum of the squares of all the predicted values with actual values. It is used as the optimal measure of the selection of the model

**Heteroscedasticity:**

The error between the predicted values and the actual values should be randomly distributed for all values of independent factors. This can be easily verified by plotting the error (residual) terms against each X. If there is no pattern there is Homoscedasticity, otherwise there is Heteroscedasticity (lack of constant variance). If Heteroscedasticity is present, this needs to be fixed prior to finalizing the model.

**No or little Multicollinearity**

The independent factors shouldn’t be correlated to each other. If they are collinear, some of these need to be excluded from the final model to provide stability to the model and estimated coefficients.

**Normality and Independence**

Residuals (errors), i.e. predicted minus actual data, should be normally distributed with mean of zero and constant standard deviation, and the residuals of independent factors should not be correlated to each other.

**Independence-**

Observations are independent from each other. Y (X+1) should not be correlated to Y (X)

## **Tools to Build Linear Regression**

**Excel-** “Data Analysis” tool pack in Excel has a tool for building multiple linear regression models

**R-** Function “lm” or “glm” are frequently used for building linear regression models

**SAS**- Proc REG in SAS achieves the same objective.

## **Key Metrics and Interpretation-**

There are several metrics generated in the multiple linear regression output. Key ones are-

**R^2-** This tells the percent of variance in the dependent variable that can be explained by the model and the independent variables in the model. R^2= Explained Variation / Total Variation. The range for this metric is 0 to 1 or from 0% to 100%. If the R^2 is 0% that means the model explain 0% of the variation in dependent variable. On the other hand, 100% signifies a perfect model, i.e. explains 100% of variations. R^2 should be as to 1 as possible for a good model.

**F Statistics and Related ‘p’ or significance value**

The F test measures the lift in the model with predictor variables versus a model with only intercept. ‘p’ value is giving the significance of rejecting the null hypothesis that all model coefficients for predictor variables are zero. F stat should be as high as possible and the associated p value of significance should be as low as possible for a good model. For example, a p value of 0.002 means that we are 1-0.002 or 99.8% confident that some coefficients of the independent variables are non-zero in the model. In other words, some independent variables have good explaining power for the dependent variable.

**Coefficients Estimate**

Coefficient estimates are the multiplicative term for each of the independent factor to derive the regression equation. In the example below, we are modelling for Sales of an ecommerce company. The equation for this can be derived as-

Sales= 624.69704+ 0.18184\*Marketing Budget- 0.556408\*Price

**Coefficients:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) 624.69704 5.29087 118.071 <2e-16 \*\*\*

Marketing Budget 0.18184 0.97023 0.187 0.851

Price -0.556408 0.06547 -84.982 <2e-16 \*\*\*

Std. Error can be used for constructing the confidence interval around the coefficient estimates.

T value is the T-test stat for the null hypothesis that the coefficient for this independent factor is zero. Associated p value is shown for the same T-test. T value should be as high and p value should be as low as possible for a good model. A p value of greater than 0.05 signifies that the variable is not a good predictor. Similarly an absolute (T value) of less than 2 shows a weak predictor. In above case for example, Number of Vehicles has a p value of 0.851 and T value of 0.187 , and hence this is not a good predictor and should be removed from the regression equation. So the equation should like this- Sales= 624.69+ 0.18\*Marketing Budget- 0.56\*Price (with rounding).

The interpretation of the coefficient of the independent factor is really simple. This shows how a unit level change in the independent factor will change the dependent factor. For example, if the Price of an item goes up $ 1, it will cause a $ -0.56 drop in sales.

**Summary:**

1) The Regression line is the “Line of Best Fit”

2) β1 is slope of the line. A 1unit increase in X will lead to β1 increase in Y

3) β0 is the value of Y when X is equals to Zero

4) β1>0 means that there is an positive relationship with X and Y

5) The Estimated regression can be used to make the prediction for Y given X. Example with 12 years of education gives wage of $17 per hour

6) The Residuals are the actual value of Y minus the predicted value

7) The Residuals terms contains all the factors(other than X) that impact Y