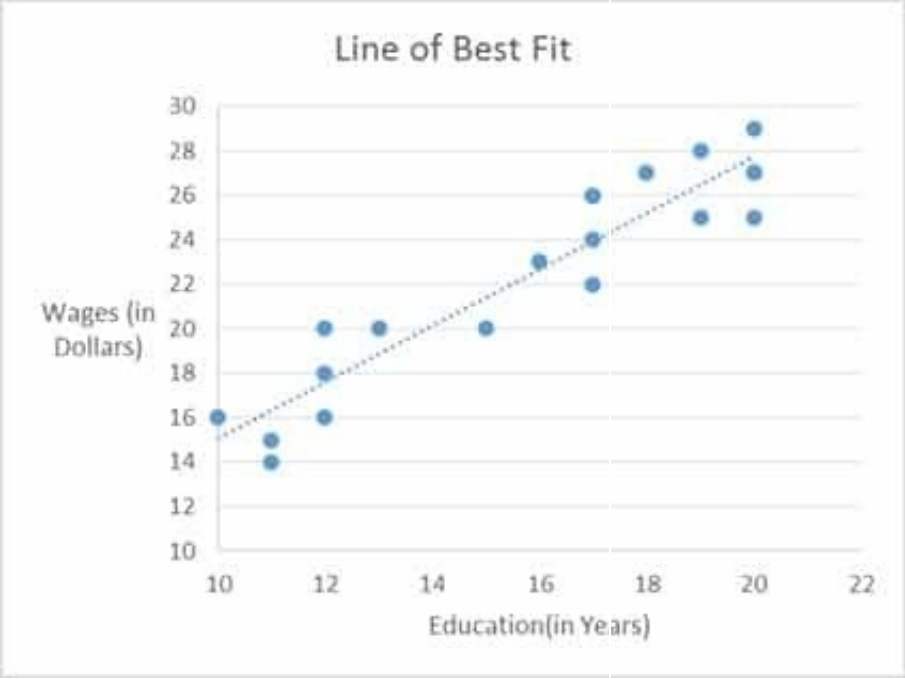
A beginners guide to Linear Regression - Learn and grow with Analytics

Notebook:	Tester and downloads
Created:	7/17/2017 19:36
URL:	https://equiskill.com/a-beginners-guide-to-linear-regression/

A beginners guide to Linear Regression

- Posted by Amit Upadhyay
- Categories Analytics, Business Analytics, Linear Regression
- Date April 23, 2016
- Comments 0 comment



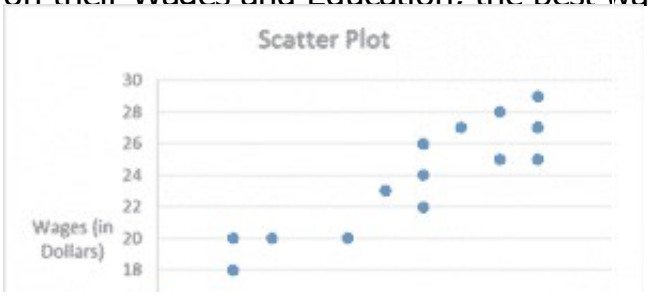
Introduction to Linear Regression When you think of Regression, think prediction. A regression uses the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable. **Types of Regressions** A regression models the past relationship between 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; when two or more independent variables are used, it is called a multiple regression. The general formula for simple and multiple linear regression is given as: Simple linear regression: Wages(dependent variable) = (Y-Intercept) + Education(Independent Variable) $Y = \beta_0 + \beta_1 X$ Multiple regression equation: Wages(dependent variable) = (Y-Intercept) + (Education) + (age) + (Gender) + (Work Experience) + (Sector) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$ So the best way to know the relationship between independent and dependent variable is by scatter plot.

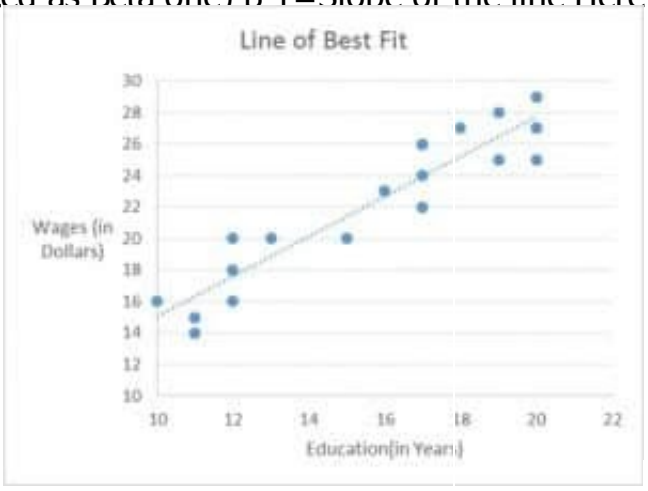
Scatter plot Consider an above example of wages and education:

Sno	Education(in yrs)	Wages (in Dollars)
1	19	28
2	16	23
3	20	27
4	17	24
5	20	25
6	11	14
7	11	15
8	19	25
9	15	20
10	13	20
11	12	20
12	20	29
13	20	27
14	12	16
15	17	26
16	18	27
17	12	18
18	10	16
19	17	22
20	17	26

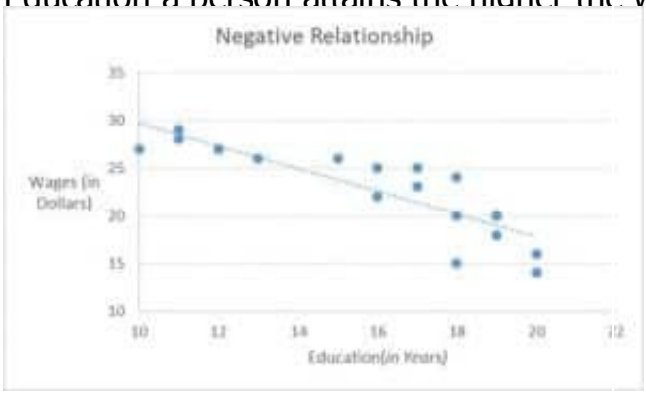
Let us consider data of 20 professionals of their years of education and Wages in dollars per hour. **Note : Make sure the collected data is a representation of the population.** In statistics we must ensure that our sample of individuals must represents our population. That means we must ensure the 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.



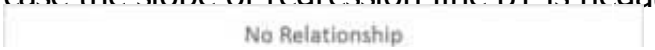
This Scatter plot allowed us to accommodate all the individuals with their wages and years in Education. Now to know the relationship between our variables or the pattern between them we use the **line of best fit.** The line of Best fit is the line which represents the general pattern of the sample. **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

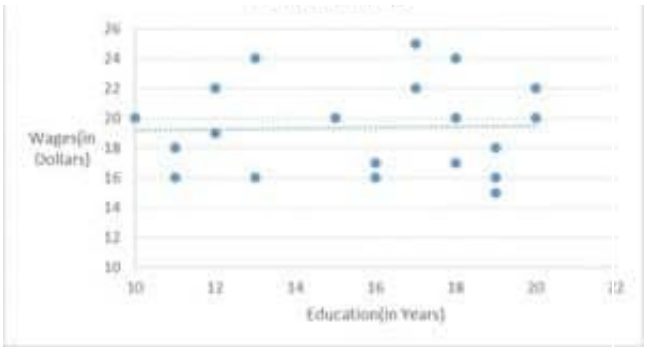


So **Y= β 0 + β 1X** Wages = β 0 + β1(Education) so if β1 >0 it has a **positive relationship.** The above Shows the positive relationship between Wages and Education. The more Education a person attains the higher the wage it gets.

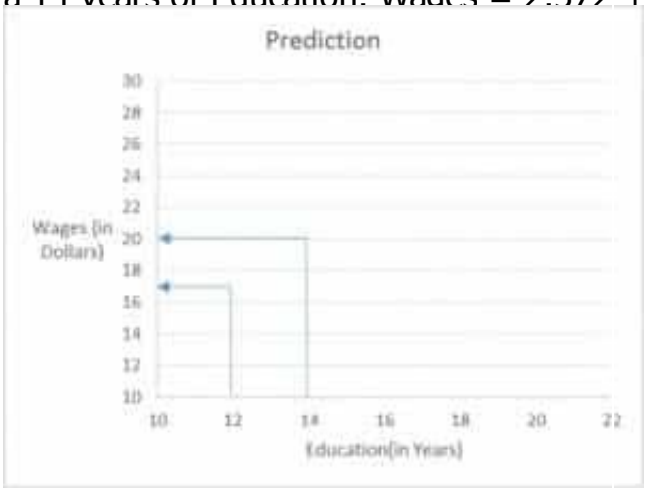


If β1 <0 it has a **negative relationship.** The regression line is in a downward direction. There is an negative relationship between the Wages and Education. It has a general trend that the more educated is any individual the less pay they would get. In this case the slope of regression line β1 is negative.

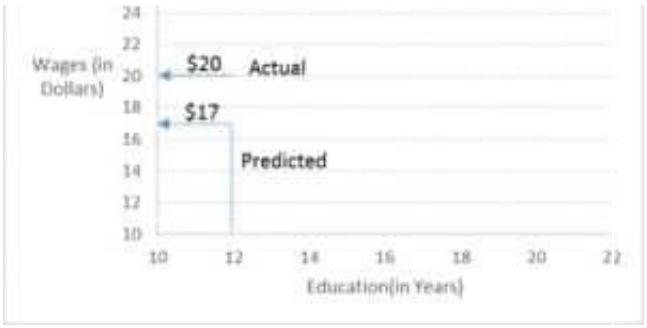




If β1 =0 it has a **No relationship.** The regression line is in a Straight direction. There may be no relationship between Wages and Education. The Slope of the regression line β1 is zero. **Estimation of regression line** Let suppose we get an estimated regression line as: **Y=2.372 + 1.267x** Means: 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



Inference or What we can infer from our prediction and the data 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 does not 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) 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

-
-
-

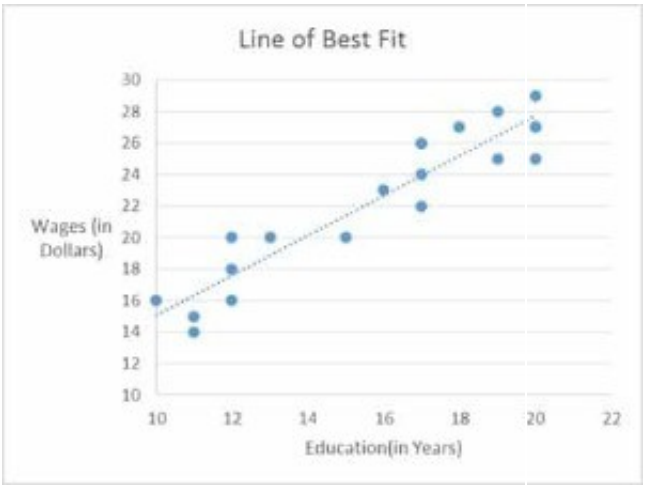
Amit Upadhyay

Previous post

Theory of Linear Regression

April 23, 2016

You may also like

- 

  Theory of Linear Regression

  29 March, 2016