Data science and Machine learning with Python

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

Regression

Regression analysis is one of the most important fields in statistics and machine learning. There are many regression methods available. Linear regression is one of them.

What Is Regression?

Regression searches for relationships among variables.

For example, you can observe several employees of some company and try to understand how their salaries depend on the **features**, such as experience, level of education, role, city they work in, and so on.

This is a regression problem where data related to each employee represent one **observation**. The presumption is that the experience, education, role, and city are the independent features, while the salary depends on them.

Similarly, you can try to establish a mathematical dependence of the prices of houses on their areas, numbers of bedrooms, distances to the city center, and so on.

Generally, in regression analysis, you usually consider some phenomenon of interest and have a number of observations. Each observation has two or more features. Following the assumption that (at least) one of the features depends on the others, you try to establish a relation among them.

In other words, **you need to find a function that maps some features or variables to others sufficiently well**.

The dependent features are called the **dependent variables**, **outputs**, or **responses**.

The independent features are called the **independent variables**, **inputs**, or **predictors**.

Regression problems usually have one continuous and unbounded dependent variable. The inputs, however, can be continuous, discrete, or even categorical data such as gender, nationality, brand, and so on.

It is a common practice to denote the outputs with y and inputs with x. If there are two or more independent variables, they can be represented as the vector $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of inputs.

When Do You Need Regression?

Typically, you need regression to answer whether and how some phenomenon influences the other or **how several variables are related**. For example, you can use it to determine *if* and *to what extent* the experience or gender impact salaries.

Regression is also useful when you want **to forecast a response** using a new set of predictors. For example, you could try to predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and number of residents in that household.

Regression is used in many different fields: economy, computer science, social sciences, and so on. Its importance rises every day with the availability of large amounts of data and increased awareness of the practical value of data.

Types of Regression

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Stepwise Regression

Linear Regression

Linear regression is probably one of the most important and widely used regression techniques. It's among the simplest regression methods. One of its main advantages is the ease of interpreting results.

Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables.

Linear regression is a basic predictive analytics technique that uses historical data to predict an output variable. It is popular for predictive modelling because it is easily understood and can be explained using plain English.

Linear regression models have many real-world applications in an array of industries such as economics (e.g. predicting growth), business (e.g. predicting product sales, employee performance), social science (e.g. predicting political leanings from gender or race), healthcare (e.g. predicting blood pressure levels from weight, disease onset from biological factors), and more.

Where is Linear Regression Used?

1. Evaluating Trends and Sales Estimates

Linear regressions can be used in business to evaluate trends and make estimates or forecasts. **For example,** if a company's sales have increased steadily every month for the past few years, conducting a linear analysis on the sales data with monthly sales on the y-axis and time on the x-axis would produce a line that that depicts the upward trend in sales. After creating the trend line, the company could use the slope of the line to forecast sales in future months.

2. Analyzing the Impact of Price Changes

Linear regression can also be used to analyze the effect of pricing on consumer behaviour.

For example, if a company changes the price on a certain product several times, it can record the quantity it sells for each price level and then performs a linear regression with quantity sold as the dependent variable and price as the explanatory variable. The result would be a line that depicts the extent to which consumers reduce their consumption of the product as prices increase, which could help guide future pricing decisions.

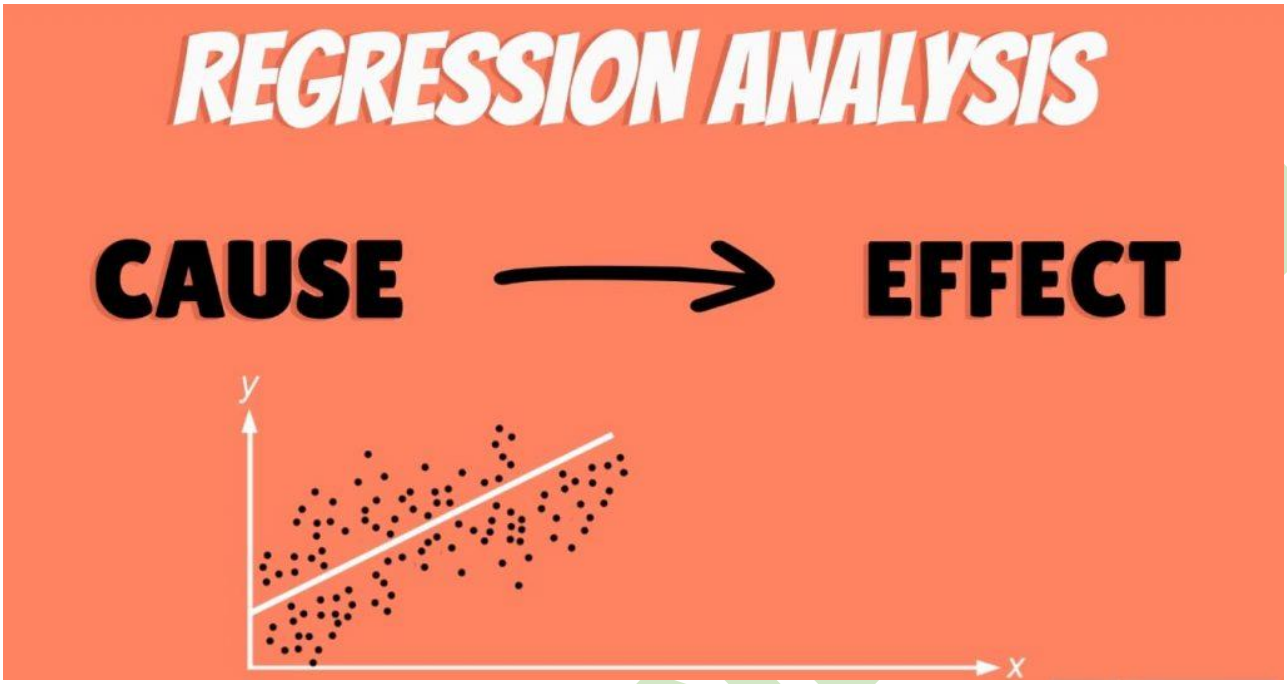
3. Assessing Risk

Linear regression can be used to analyze risk.

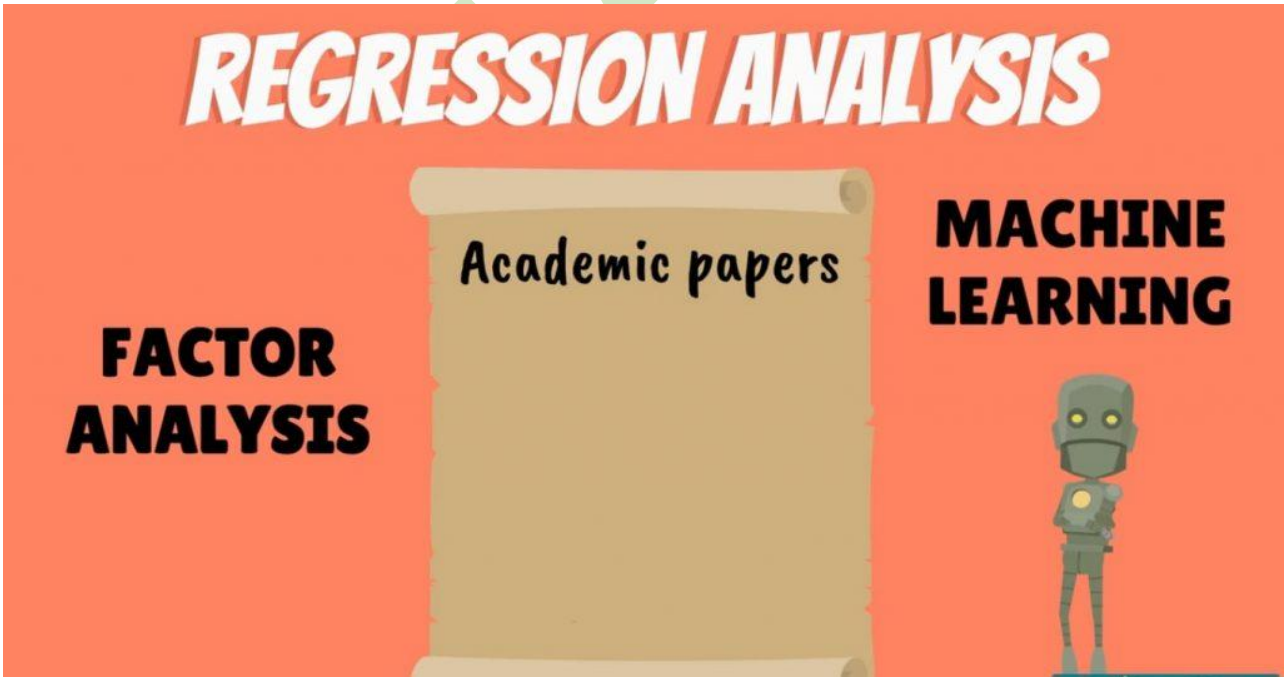
For example

A health insurance company might conduct a linear regression plotting number of claims per customer against age and discover that older customers tend to make more health insurance claims. The results of such an analysis might guide important business decisions made to account for risk.

Regression analysis is one of the most widely used methods for prediction. It is applied whenever we have a causal relationship between variables.



A large portion of the predictive modelling that occurs in practice is carried out through **regression analysis**. There are also many academic papers based on it. And it becomes extremely powerful when combined with techniques like *factor analysis*. Moreover, fundamentals of **regression analysis** are used in **machine learning**.

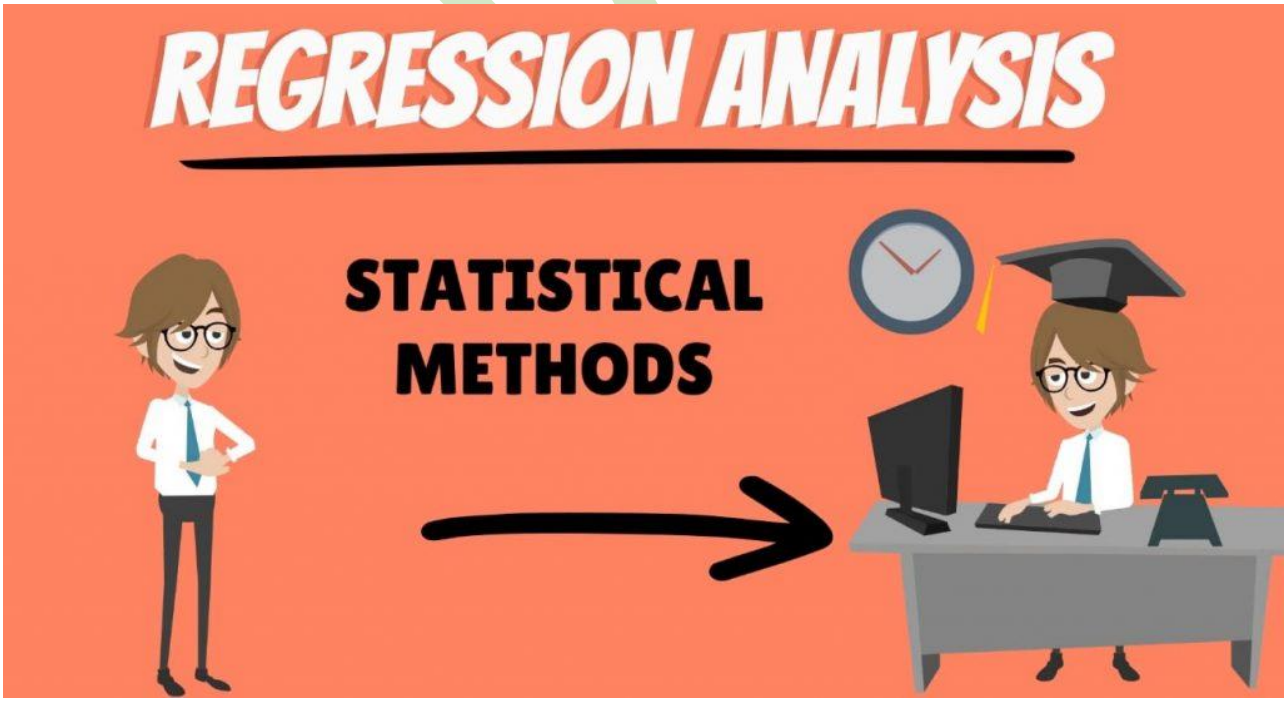


Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

Therefore, it is easy to see why **regressions** are a must for data science. The general point is the following.

"The amount of money you spend depends on the amount of money you earn."

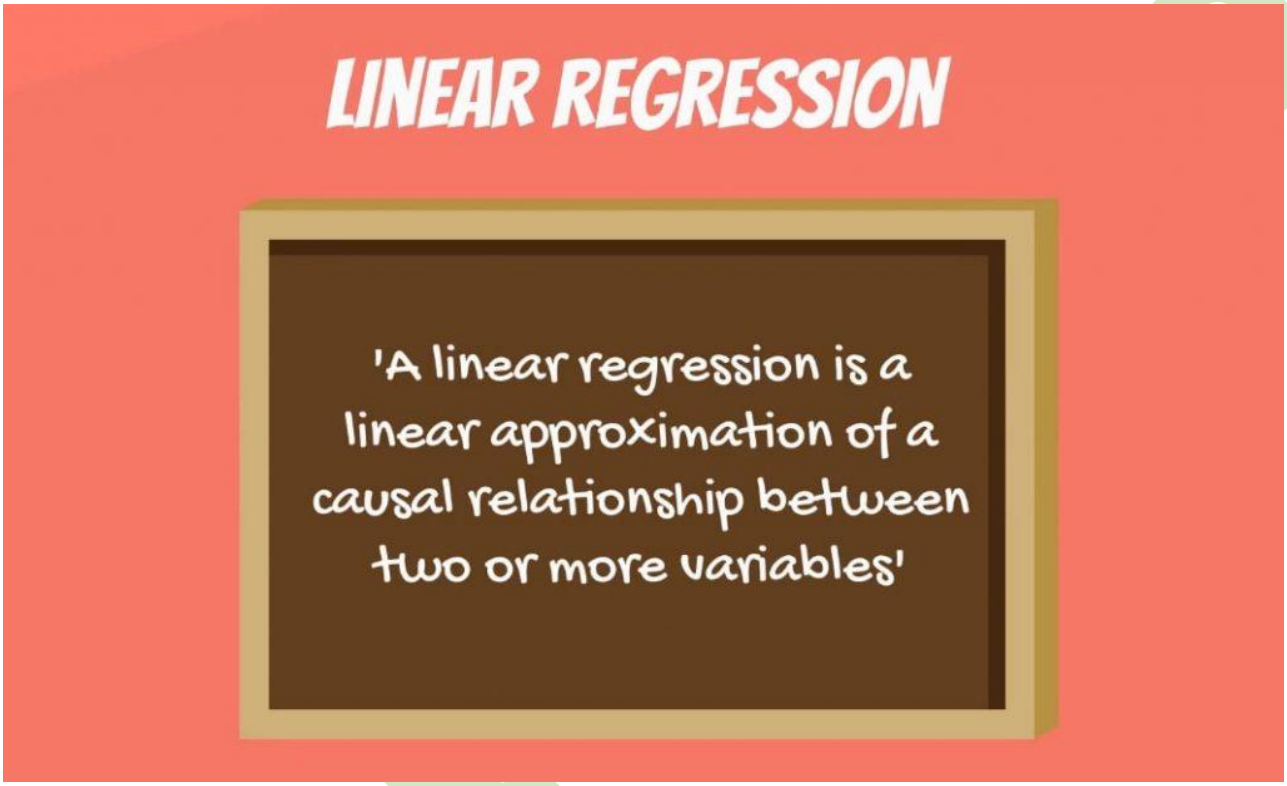
In the same way, the amount of time you spend reading our tutorials is affected by your motivation to learn additional statistical methods.



You can quantify these relationships and many others using **regression analysis**.

What is a Linear Regression

Let's start with some dry theory. A **linear regression** is a linear approximation of a causal relationship between two or more variables.

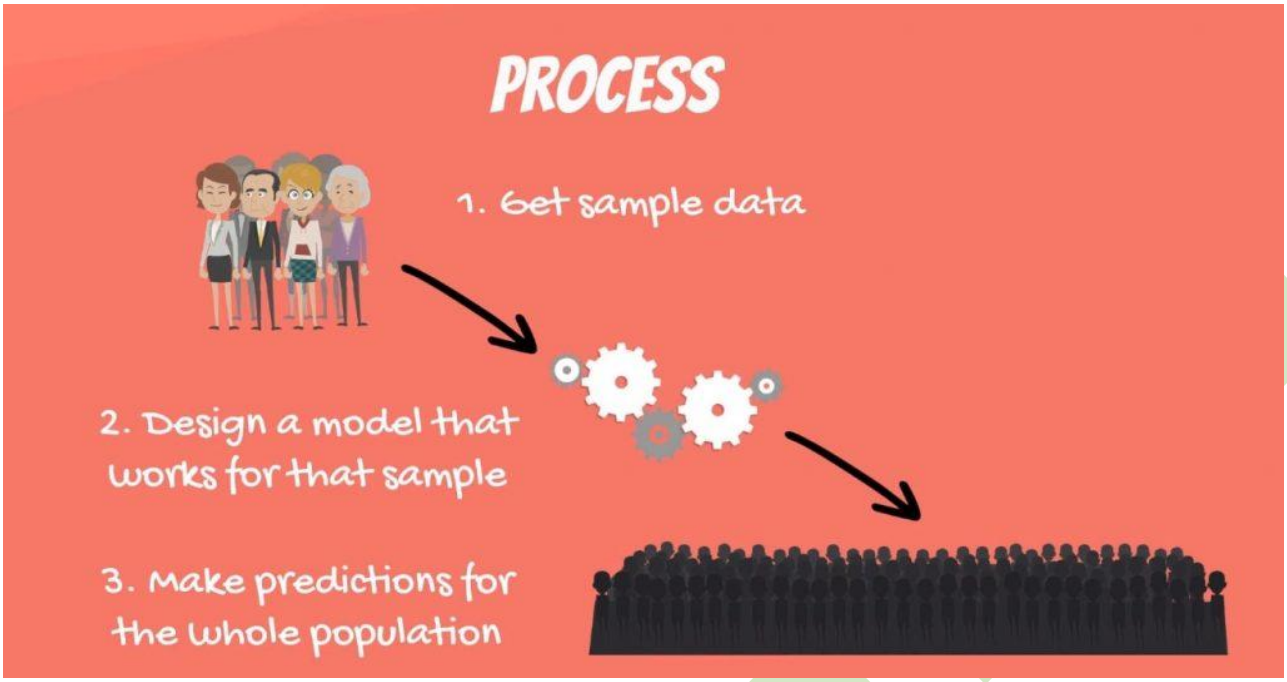


Regression models are highly valuable, as they are one of the most common ways to make inferences and predictions.

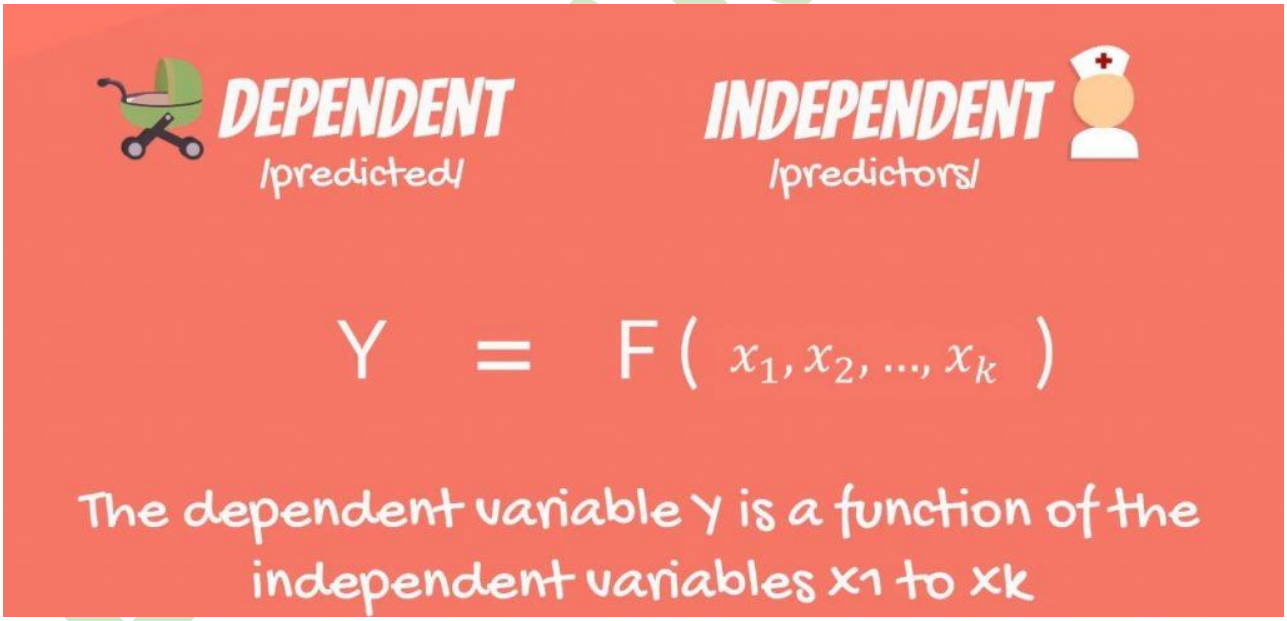
The Process of Creating a Linear Regression

The process goes like this.

1. First, you get sample data;
2. Then, you can design a model that explains the data;
3. Finally, you use the model you've developed to make a prediction for the whole population.



There is a dependent variable, labeled *Y*, being predicted, and independent variables, labeled *x1*, *x2*, and so forth. These are the predictors. *Y* is a function of the *X* variables, and the **regression model** is a linear approximation of this function.



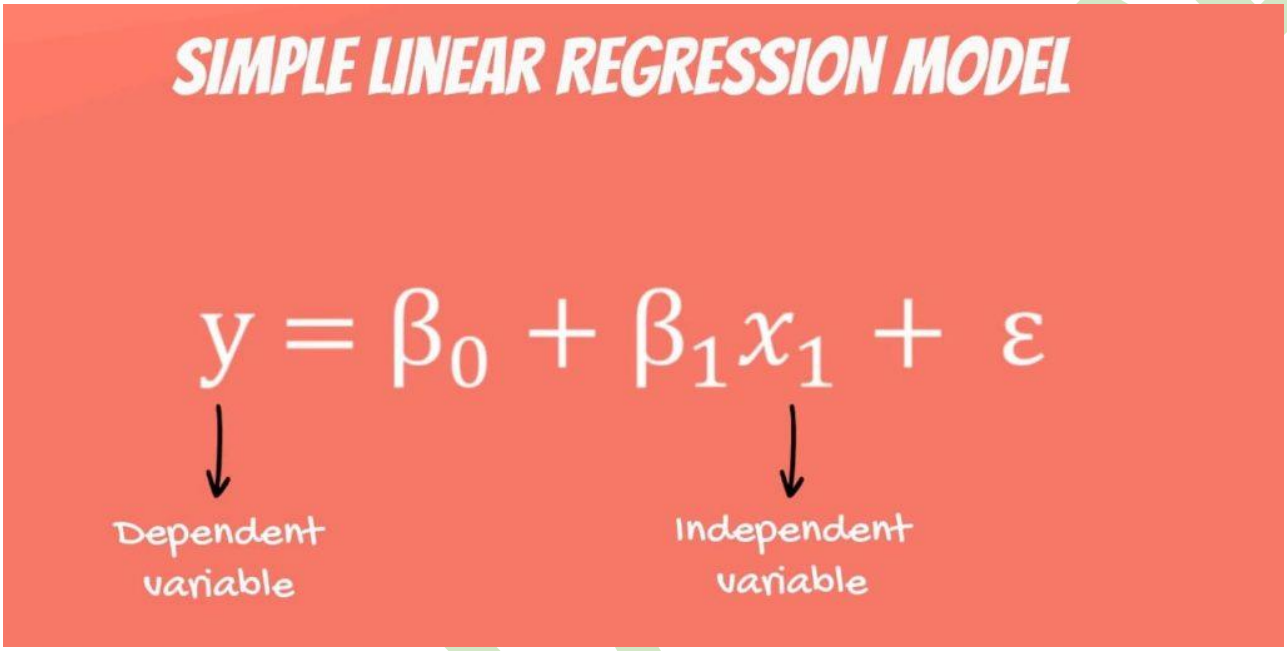
Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

The Simple Linear Regression

The easiest **regression model** is the **simple linear regression**:

$$Y = \beta_0 + \beta_1 * x_1 + \varepsilon.$$

Let's see what these values mean. *Y* is the variable we are trying to predict and is called the *dependent variable*. *X* is an *independent variable*.

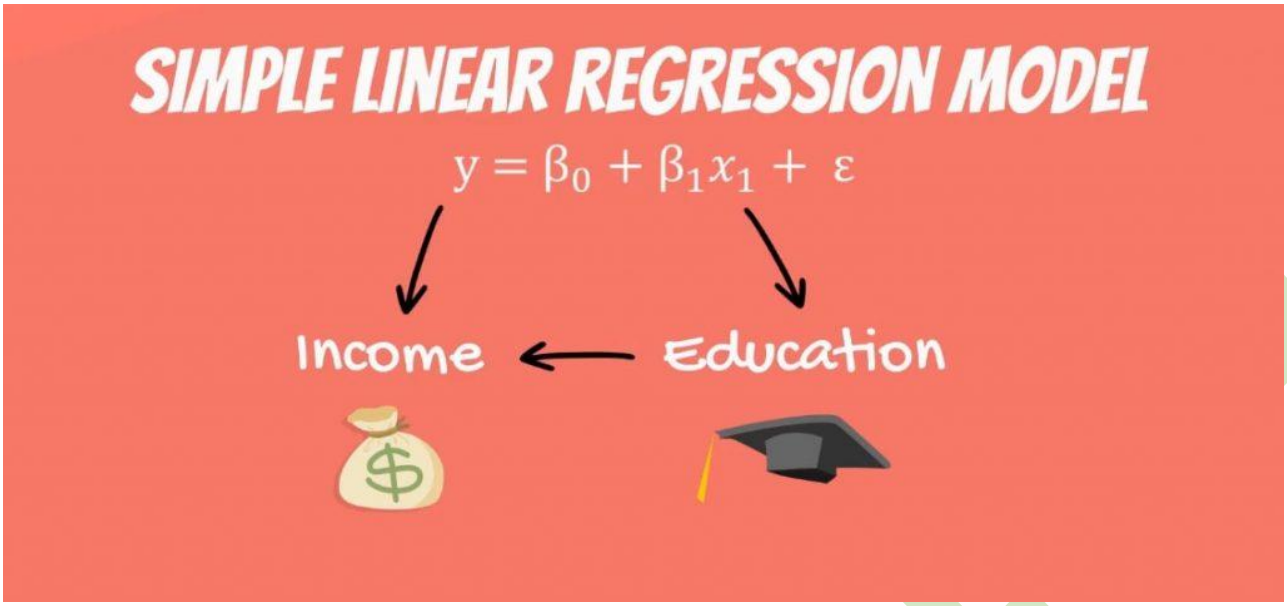


When using **regression analysis**, we want to predict the value of *Y*, provided we have the value of *X*.

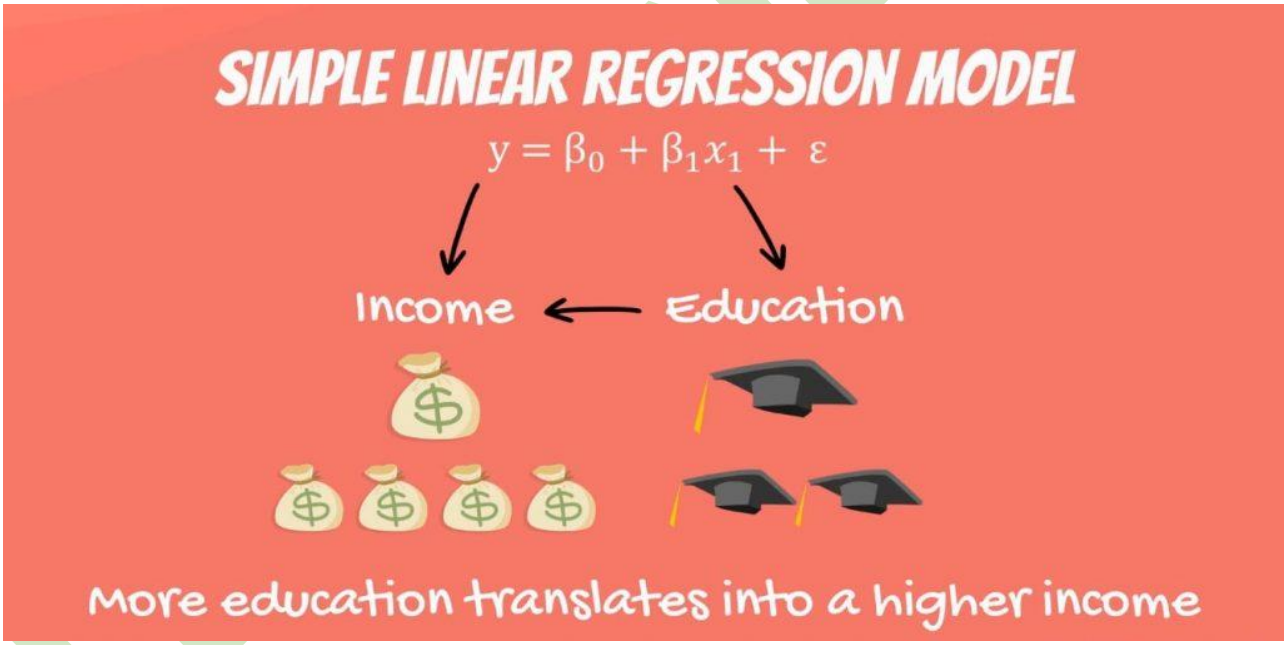
But to have a **regression**, *Y* must depend on *X* in some way. Whenever there is a change in *X*, such change must translate to a change in *Y*.

Providing a Linear Regression Example

Think about the following equation: the income a person receives depends on the number of years of education that person has received. The *dependent variable* is income, while the *independent variable* is years of education.

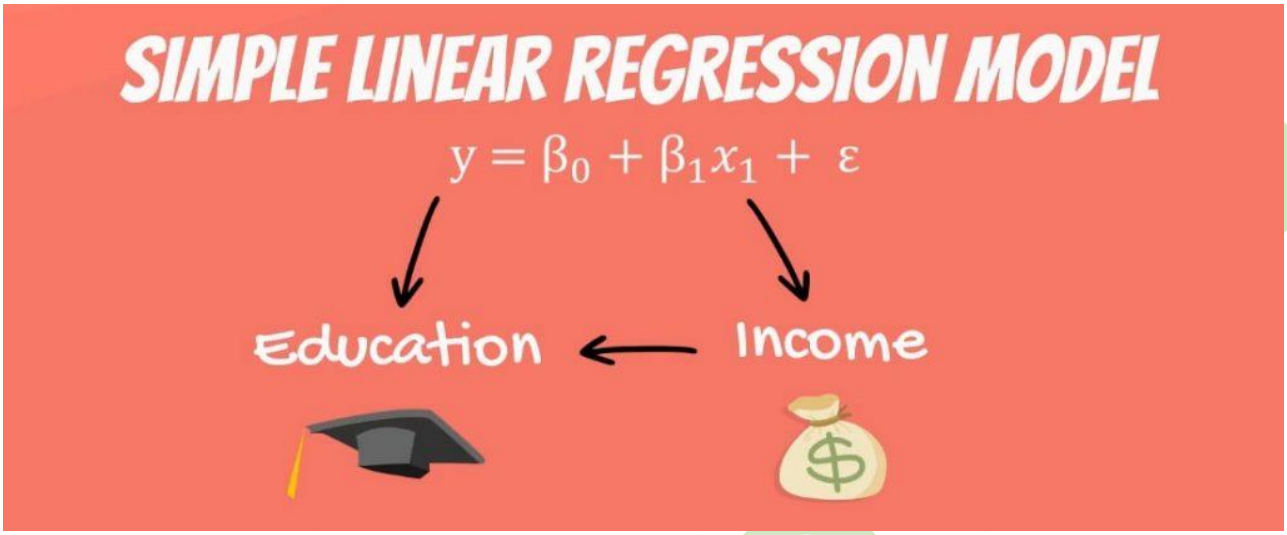


There is a causal relationship between the two. The more education you get, the higher the income you are likely to receive. This relationship is so trivial that it is probably the reason you are reading this tutorial, right now. You want to get a higher income, so you are increasing your education.

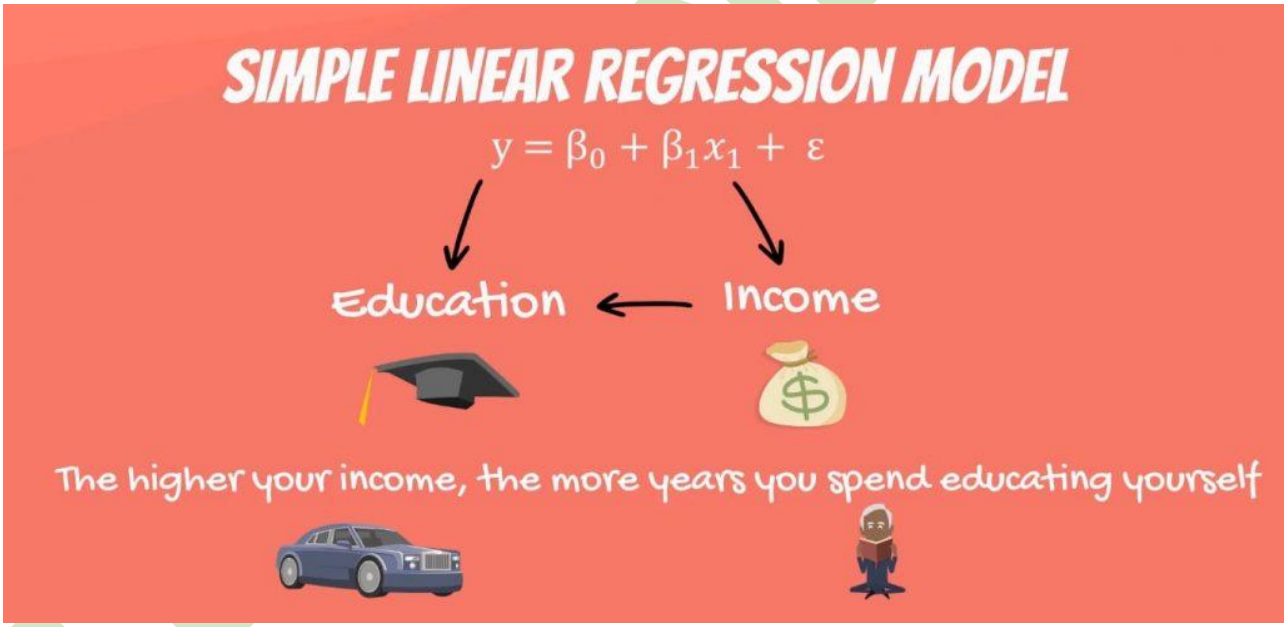


Is the Reverse Relationship Possible?

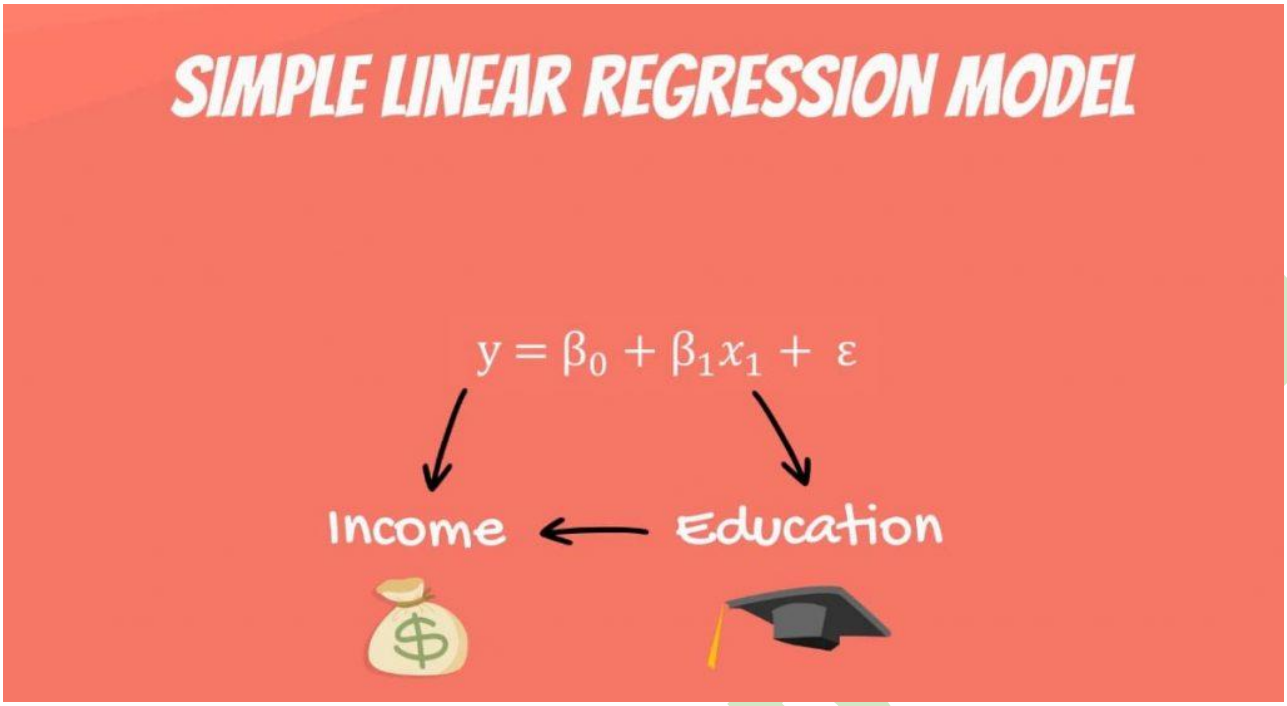
Now, let's pause for a second and think about the reverse relationship. What if education depends on income.



This would mean the higher your income, the more years you spend educating yourself.

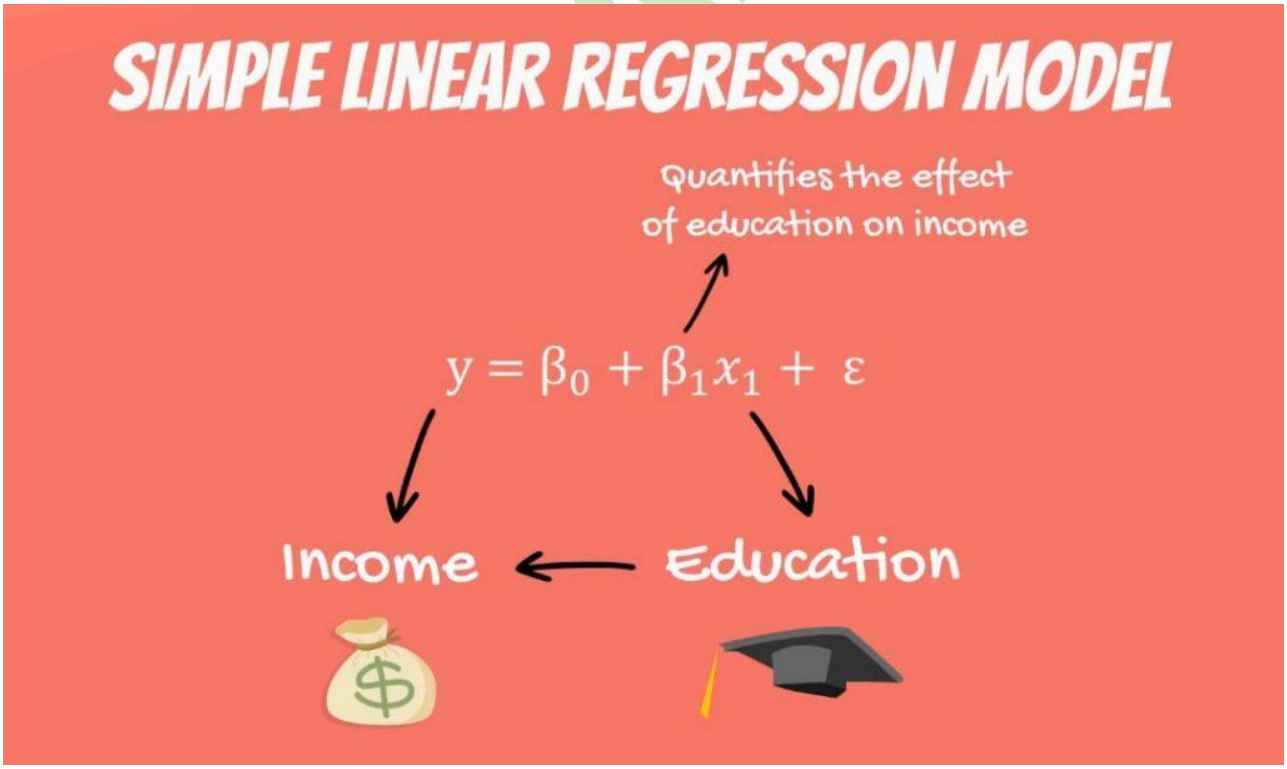


Let's go back to the original **linear regression** example. Income is a function of education. The more years you study, the higher the income you will receive. This sounds about right.

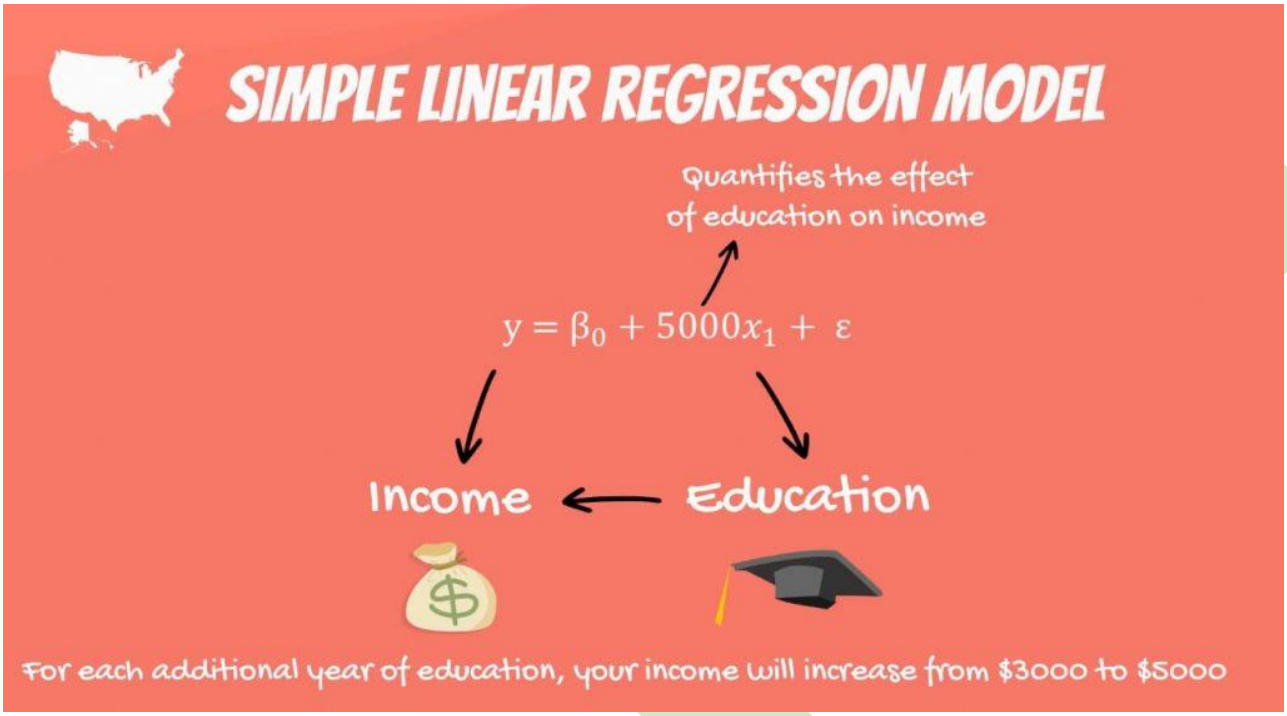


The Coefficients

What we haven't mentioned, so far, is that, in our model, there are coefficients. β_1 is the coefficient that stands before the independent variable. It quantifies the effect of education on income.



If β_1 is 50, then for each additional year of education, your income would grow by $50. In the USA, the number is much bigger, somewhere around 3 to 5 thousand dollars.

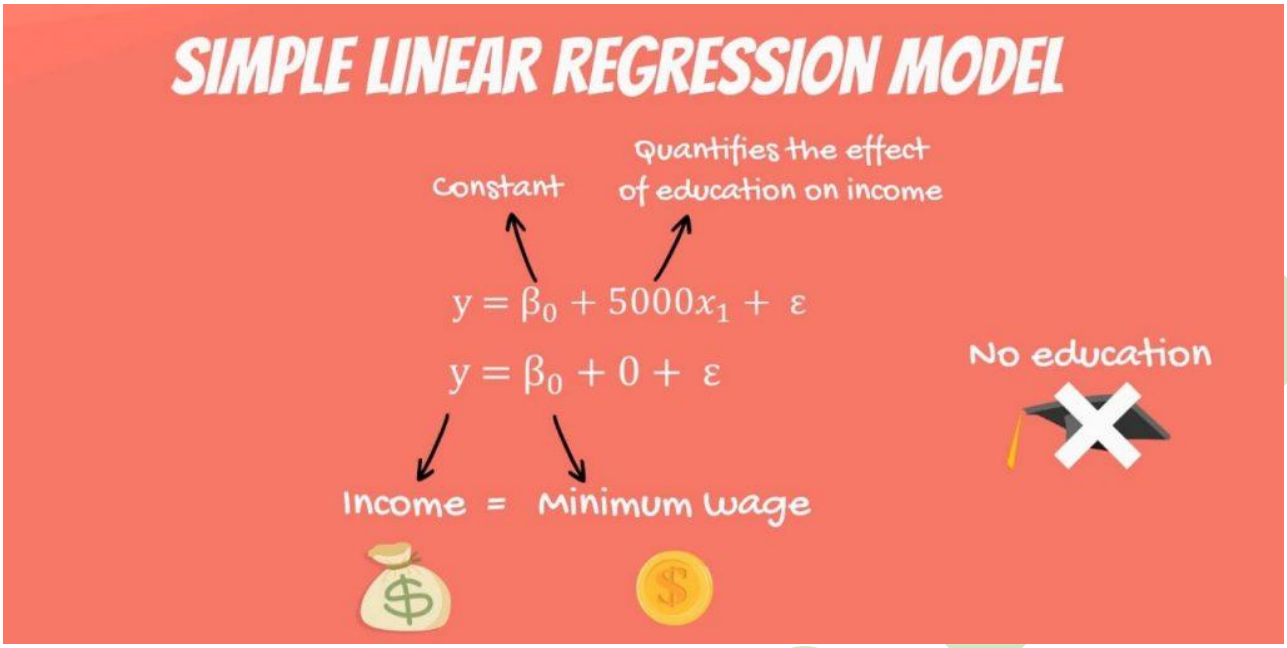


The Constant

The other two components are the constant β_0 and the error – epsilon(ε).

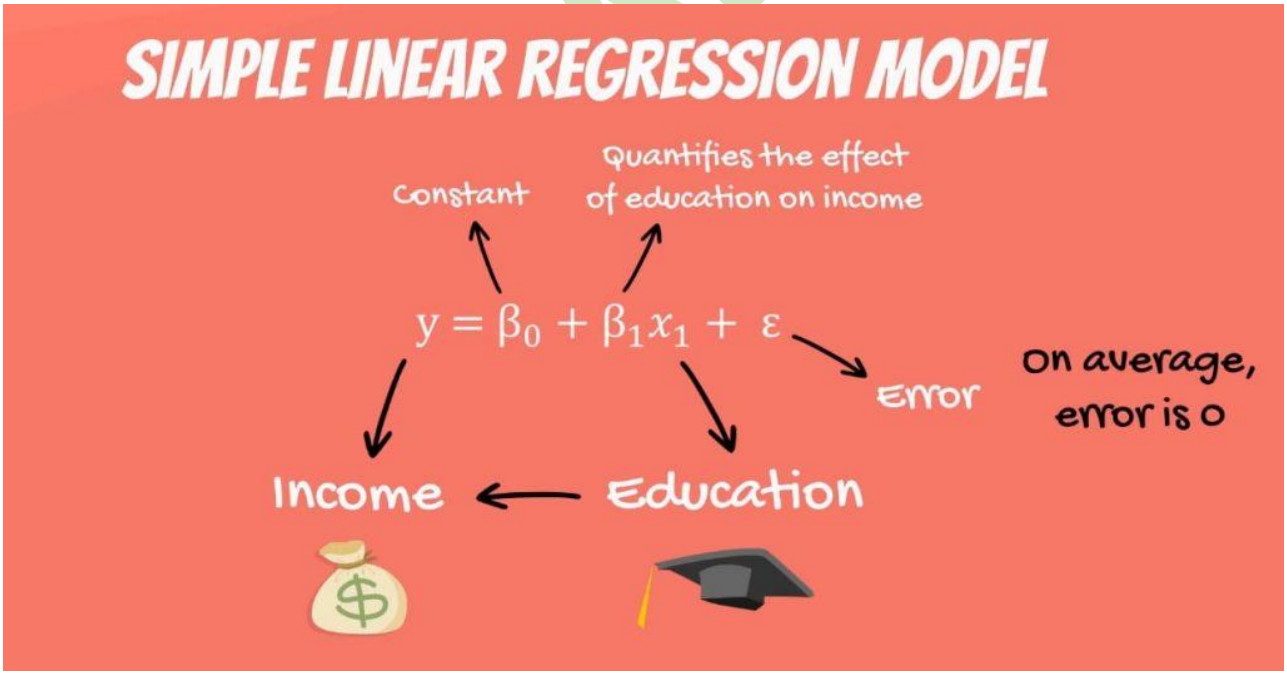
In this **linear regression** example, you can think of the constant β_0 as the minimum wage. No matter your education, if you have a job, you will get the minimum wage. This is a guaranteed amount.

So, if you never went to school and plug an education value of 0 years in the formula, what could possibly happen? Logically, the **regression** will predict that your income will be the minimum wage.



Epsilon

The last term is the epsilon(ε). This represents the error of estimation. The error is the actual difference between the observed income and the income the **regression** predicted. On average, across all observations, the error is 0.

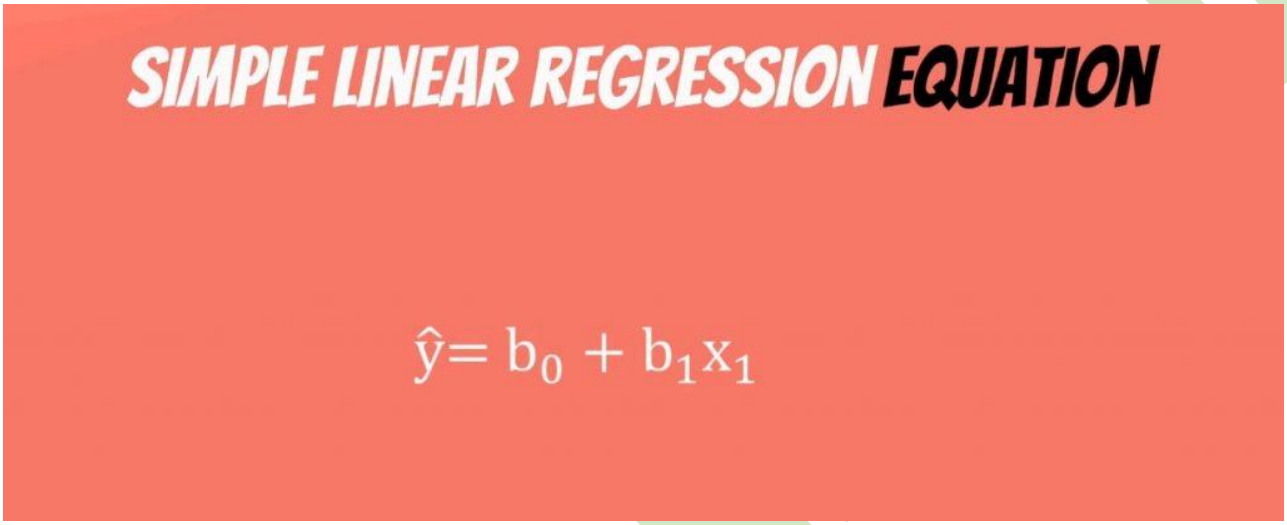


If you earn more than what the **regression** has predicted, then someone earns less than what the **regression** predicted. Everything evens out.

The Linear Regression Equation

The original formula was written with Greek letters. This tells us that it was the population formula. But don't forget that statistics (and data science) is all about sample data. In practice, we tend to use the **linear regression *equation***.

It is simply $\hat{y} = \beta_0 + \beta_1 * x$.



The $\hat{y}$ here is referred to as *y hat*. Whenever we have a hat symbol, it is an estimated or predicted value.

B_0 is the estimate of the **regression** constant β_0. Whereas, b_1 is the estimate of β_1, and x is the sample data for the *independent variable*.

-

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.