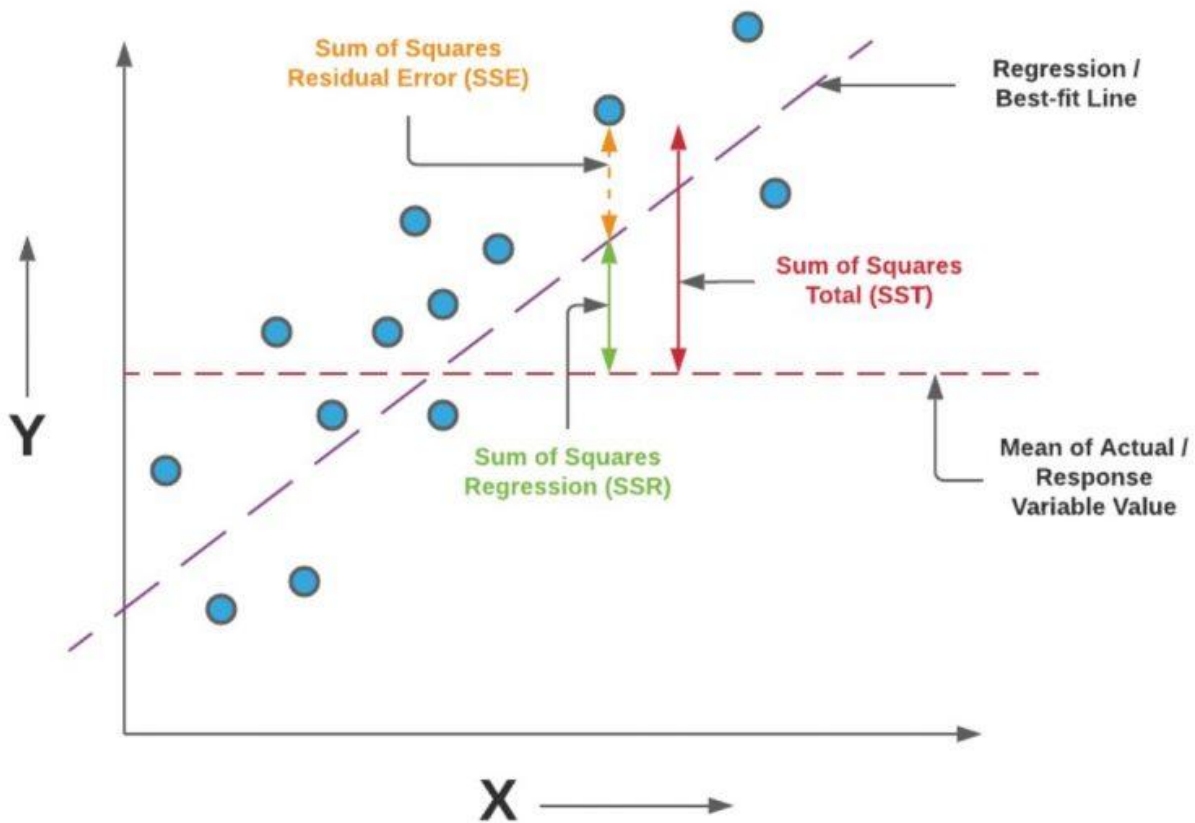


Study Guide: Regression Analysis Study Guide



What is Linear Regression?

- A linear regression is a linear approximation of a causal relationship between two or more variables
- Regressions models are highly valuable, as they are one of the most common ways to make inferences and predictions

Linear Regression Process:

1. Get sample data
2. Come up with a model that explains the data
3. Make predictions for the whole population based on the model you have developed

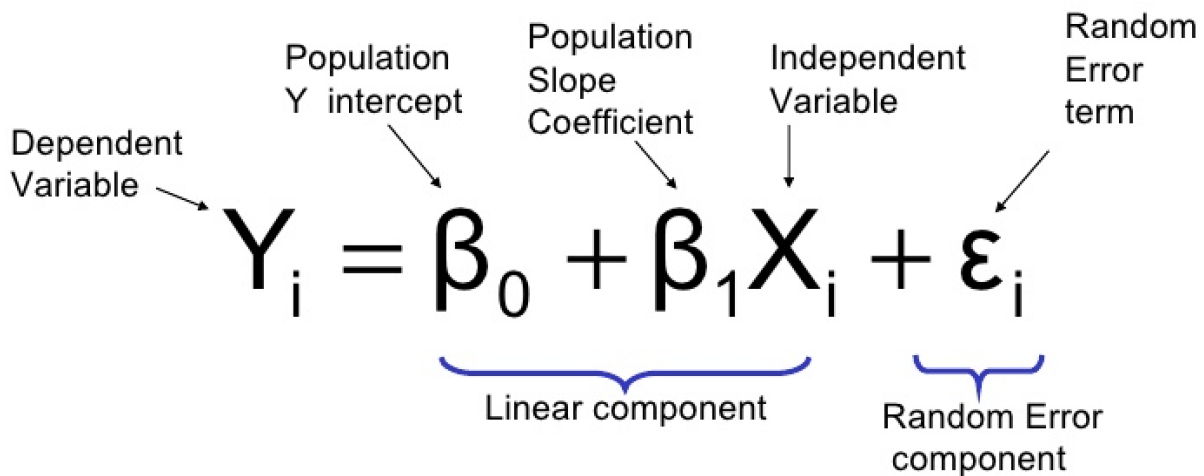
There is a dependent variable labeled Y, being predicted, and an independent variable labeled X_1, X_2, \dots, X_k . These are the predictors.

Y is a function of the X variables, and the regression model is a linear approximation of this function $Y = F(X_1, X_2, \dots, X_k)$

The easiest regression model is the simple linear regression.

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Let's see what these values mean.



The diagram shows the equation $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with the following labels and arrows:

- Dependent Variable** points to Y_i .
- Population Y intercept** points to β_0 .
- Population Slope Coefficient** points to β_1 .
- Independent Variable** points to X_i .
- Random Error term** points to ϵ_i .

Below the equation, two blue brackets indicate components:

- A bracket under $\beta_0 + \beta_1 X_i$ is labeled **Linear component**.
- A bracket under ϵ_i is labeled **Random Error component**.

- **Y** is the variable we are trying to predict, and is called the **dependent variable**. **x** is the 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 causal way.

Whenever there is a change in X , such change must translate into a change in Y .

- β_1 is the coefficient that stands before the independent variable.
- β_0 - is a constant.
- ϵ - This represents the error of estimation. On average, across all observations, the error is 0.

The original formula was written with Greek letters. What does this tell us? It was the population formula.

But we know statistics is all about sample data. In practice, we use the linear regression equation.

$$\hat{y} = \beta_0 + \beta_1 X_1$$

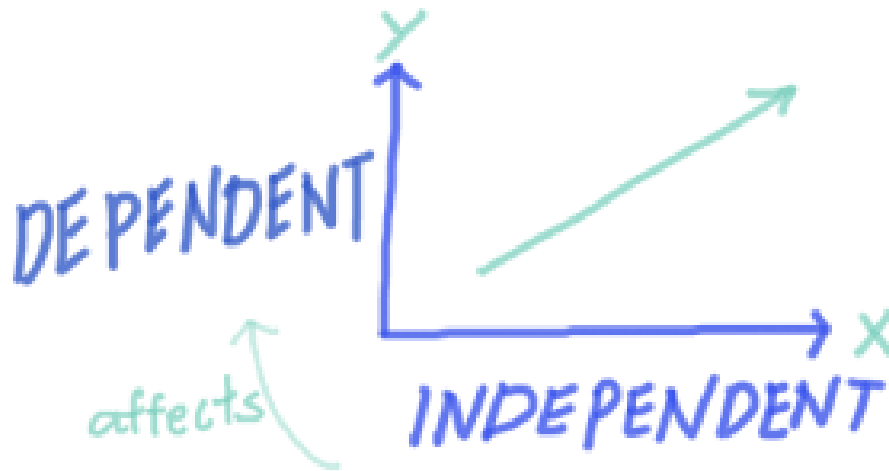
You heard it right. The y here is referred to as \hat{y} . Whenever we have a hat symbol, it is an estimated or a predicted value.

β_0 is the estimate of the regression constant β_0 , while β_1 is the estimate of β_1 , and x is the sample data for the independent variable.

The Linear Regression Model at Work

There's an X variable and a Y variable in this case.

The independent variable is on the x -axis, and the dependent variable is on the y -axis.



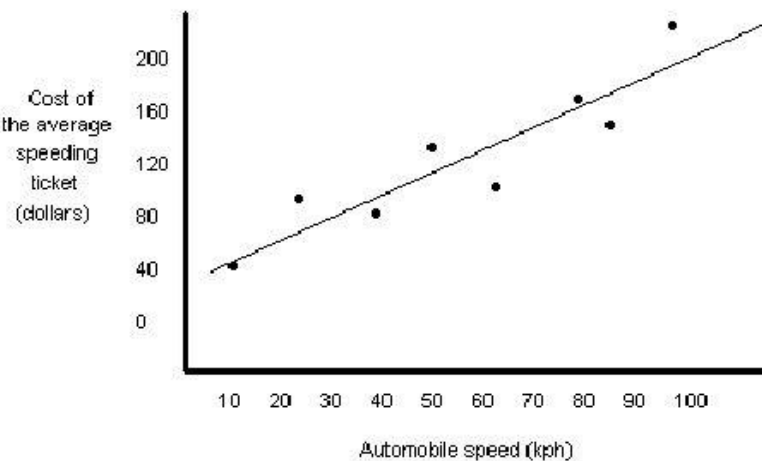
And we try to form a relationship between these two variables, and draw a line, in this case a straight line.

Independent vs Dependent variables on a graph

Look at the graph on the right

- Which is the independent variable?
- Which is the dependent variable?

The Dependence of Traffic Ticket Cost on Automobile Speed

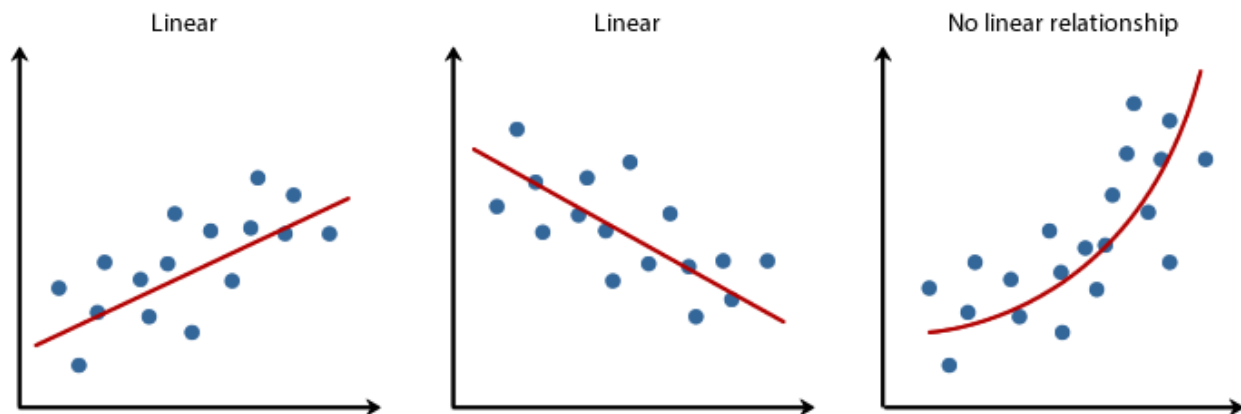


What we try to understand is as the independent variable is moving, or changing, what happens to the dependent variable? Does it go up, or does it go down? How does it change?

If they move in the same direction, if the independent variable increases, and the dependent variable increases as well, we say there's a **positive relationship**.

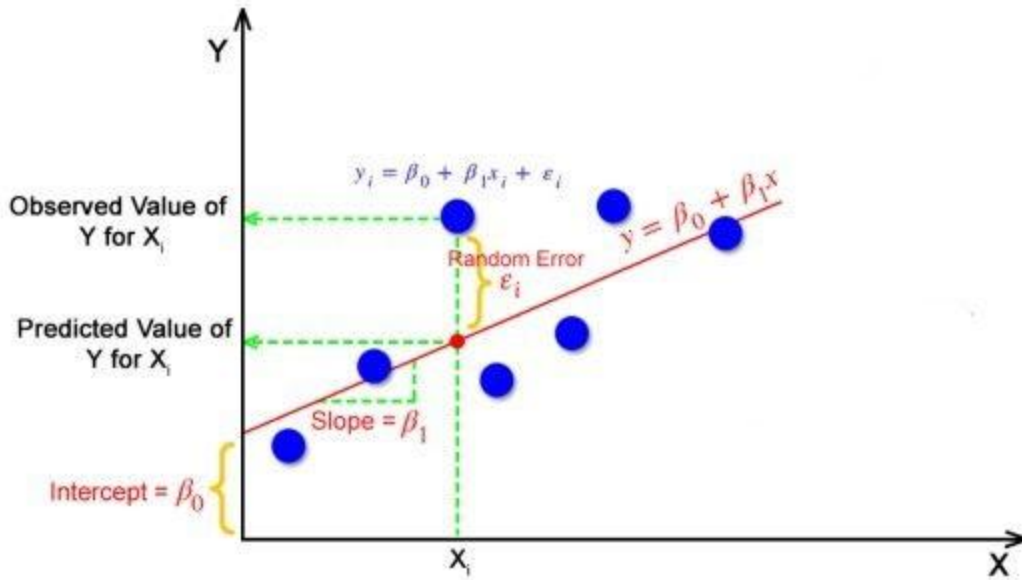
If, on the other hand, as the independent variable increases, and the dependent variable decreases, we say there's a **negative relationship**. In this case the line would go downward.

In linear regression, we try to make a straight line. You can also do curved lines, but for this topic, it's all straight lines.



Copyright 2014. Laerd Statistics.

To conduct regression, you need to take observations or data points, and then try to find a line that will fit, a straight line that fits through all these different points. This is called **the regression line**, and it's based upon the least squares method.



And in the end, you want to minimize the difference between the estimated value and the actual value, you want to **minimize the errors**.

When you have a positive relationship in a regression (when one variable goes up the other variable goes up as well making a positive slope), we write the equations like this:

$$\hat{y} = \beta_0 + \beta_1 X_1$$

When we have a negative relationship, we write the equation like this (negative slope):

$$\beta_0 - \beta_1 X_1$$

β_0 is still the y-intercept.

The slope of the line is $-\beta_1$, because it's downward sloping. **Negative relationship.**

Let's remember:

- The X is the independent variable.
- The Y is the dependent variable. The X is what we control, what we manipulate, what we change.
- The dependent variable is the outcome.