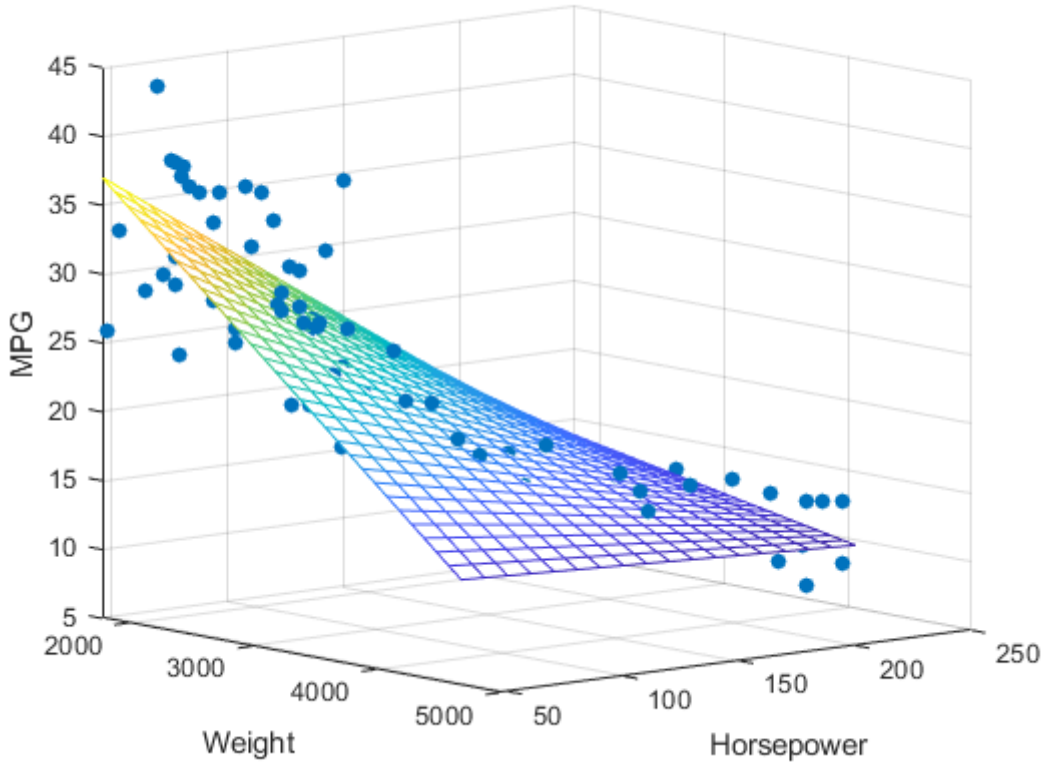
Basics of linear regression

AKM Adib

Nov 6, 2018 · 5 min read



Linear regression. You may have come across that term while you were taking a course on statistics or going through YouTube videos on machine learning. Linear regression is one of the most popular machine learning algorithm, easiest to learn and the oldest types of machine learning algorithm. This is the foundation for learning more complex machine learning algorithms and learning this algorithm is a good way to dive deep into machine learning.

What is linear regression?

Linear regression is simply a linear method to model the relationship between your independent variables and your dependent variables. This means that let's say if we

have a scatter plot with some points on it, the objective for linear regression is to make a line that is able to be as close as possible to all the points.

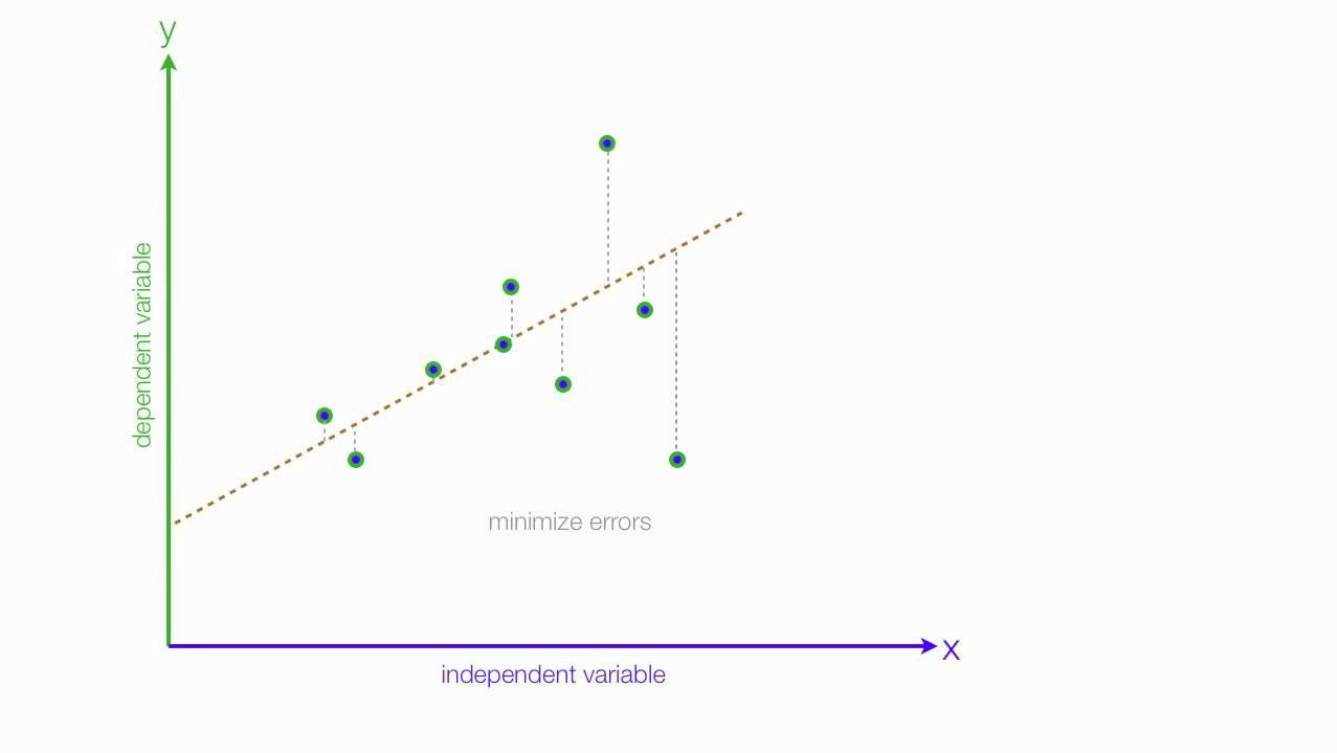


Figure 1. This is an example of what linear regression looks like and aims to achieve

The most common uses for linear regression is to predict results for a given data set. For example, we can have 3 houses which have sizes of 400, 800, and 1200 feet square respectively and the costs of these three are 100, 200, and 300 dollars. Let's say we want to buy a house that has a size of 600 feet square and we want to know the price of it. Well the price of it will most likely be in between 100 or 200 and it will most likely be 150 because since 600 feet square is in between 400 feet square and 800 feet square and they both have costs of 100 and 200 then a 600 feet square house will have a price that's exactly 150 dollars.

The general equation for the linear regression is $y = mx + b$. But in both cases, the value to your left is always the dependent variable which depends on the independent variable multiplied by the slope (m) and the added or subtracted by the value of b.

Advantages of linear regression

When we know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because it's the least complex to compared to other algorithms that also try finding the relationship between independent and dependent variable.

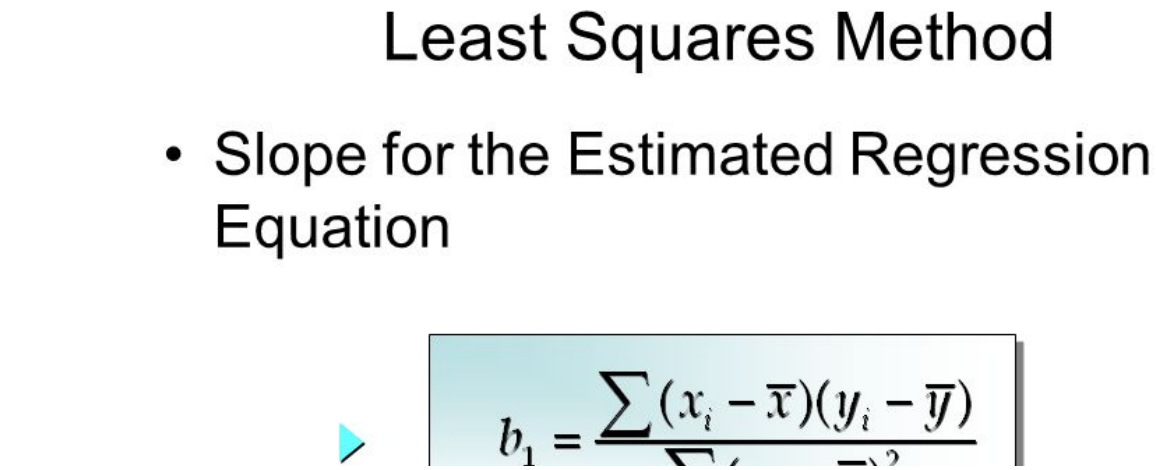
Disadvantages of linear regression

In real life, there aren't many problems in the world that exhibit a clear relationship between the independent and dependent variables. For example, let's go back to the size v cost example. Often there are many other factors that play a role in determining the cost. However, with that being said one can argue that we just need to add more independent values such as proximity to transportation, crime rate and etc. But, with even that being said there is no way that we can confirm that a 600 feet square house will cost exactly 150 dollars simply because of nothing is inevitable until it happens. I remember using ordinary least squares for a lab report on determining the relationship between the length of the pendulum and its' period. Using OLS I was 0.1 off the value calculated using the formula. 0.1 in machine learning is a big number.

Furthermore, linear regression most of the time can be only used when we deal with relationships that graphically look like a line because "linear" means according to the mathematical graphical definition is a straight line. Outliers are other that make linear regression more limited in terms of its' use because linear regression always considers the case that tends to be the most frequent. For example, if we compared a person's IQ and the score they got on the SAT and let's say the relation between this was the higher the IQ, the higher the SAT. But one cocky student with a 160 IQ got 400 on the SAT, this would be catastrophic for our model and the linear regression would ignore this.

How do we create this line of best fit?

The most common method of creating this line of best fit is by using an estimation technique called the least square method. The equation is shown below where xi and yi is a specific value in your data set and x and y with a dash above is the mean for all the x or y values.

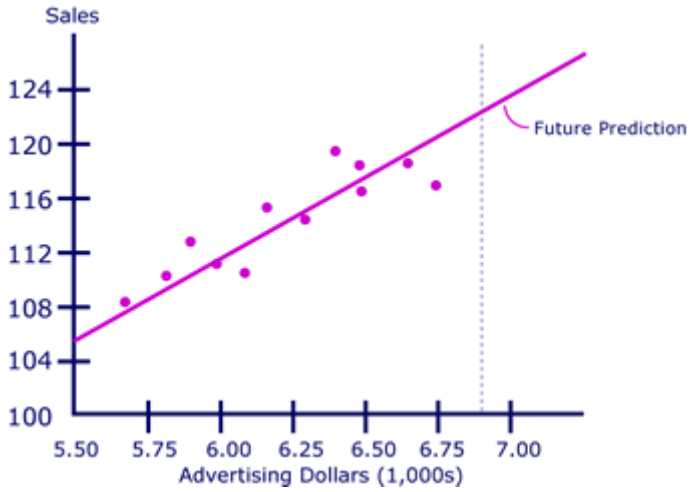


$$\sum (x_i - \bar{x})^2$$

The equation for the least square method

Applications

There are many applications to linear regression such as machine learning, trend estimation, and economics. The most common supervised learning machine learning algorithm is the linear regression because of its' simplicity and the fact that it has been around for a while. Trend estimation also extensively uses linear regression because after all regression is the prediction of results with the continuous output. Examples of situations where linear regression can be used for are predicting oil or stock prices in the future. In economics, many things are also predicted using linear regression such as labor demand and supply, consumption spending and etc.



Linear regression is really useful for trend estimation

Key Points

So this was my first article about machine learning algorithms and I hope you enjoyed it.

Here are a few takeaways:

Linear regression is one of the most popular machine learning algorithms used to predict values given a certain set of values. Linear regression is a linear method to model the relationship between your independent variables and your dependent variables.

Advantages include how simple it is and ease with implementation and disadvantages include how is' lack of practicality and how most problems in our real world aren't "linear".

You can use the least square method to create a line that would best fit the data. Some applications of linear regression can be found in machine learning, economics and in places where estimation is required.

Machine Learning Linear Regression Artificial Intelligence