Harvard Business Review

A Refresher on Regression Analysis

by Amy Gallo

https://hbr.org/2015/11/a-refresher-on-regression-analysis

November 04, 2015

You probably know by now that whenever possible you should be making data-driven decisions at work. But do you know how to parse through all of the data available to you? One of the most important types of data analysis is **regression**.

To better understand this method and how companies use it, I talked with Tom Redman, author of *Data Driven: Profiting from Your Most Important Business Asset.* He also advises organizations on their data and data analysis.

What is regression analysis?

Redman offers this example scenario: Suppose you're a sales manager trying to predict next month's numbers. You know that dozens, perhaps even hundreds of factors from the weather to a competitor's promotion to the rumor of a new and improved model can impact the number. Perhaps people in your organization even have a theory about what will have the biggest effect on sales. "Trust me. The more rain we have, the more we sell." "Six weeks after the competitor's promotion, sales jump."

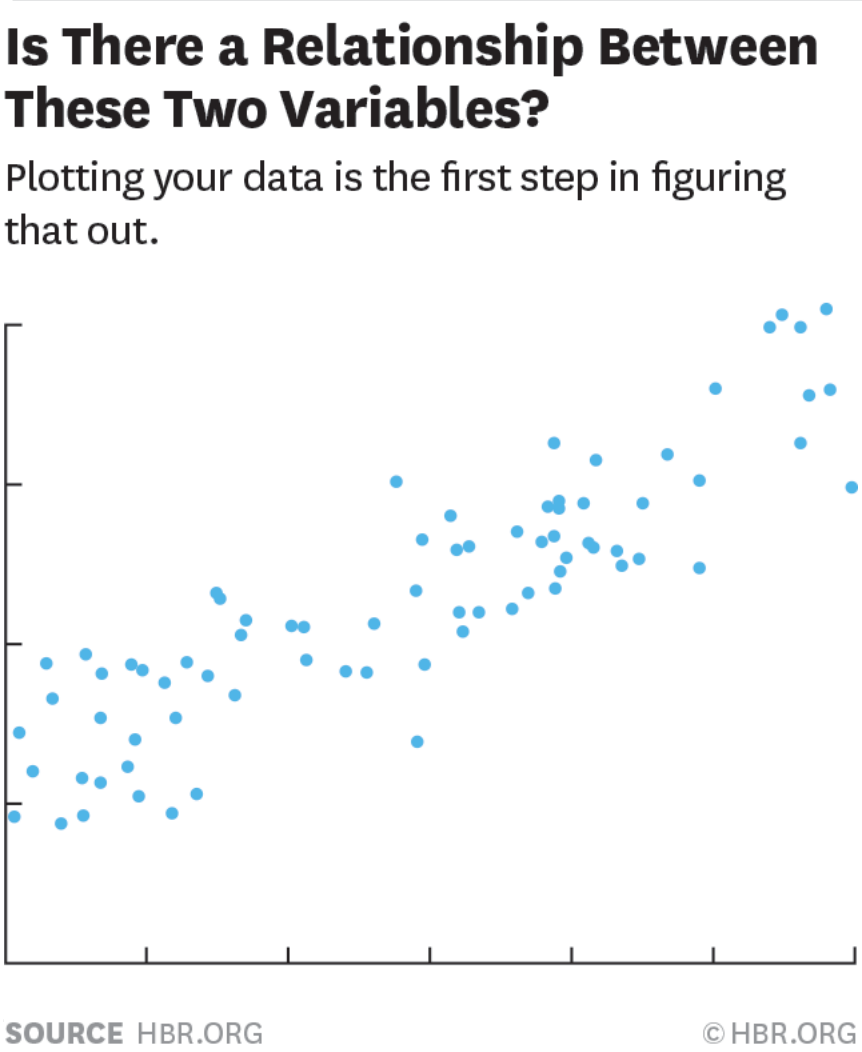
Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors?

In regression analysis, those factors are called variables. You have your **dependent variable** — the main factor that you're trying to understand or predict. In Redman's example above, the dependent variable is monthly sales. And then you have your **independent variables** — the factors you suspect have an impact on your dependent variable.

How does it work?

In order to conduct a regression analysis, you gather the data on the variables in question. You take all of your monthly sales numbers for, say, the past three years and any data on the independent variables you're interested in. So, in this case, let's say you find out the average monthly rainfall for the past three years as well.

Then you plot all of that information on a chart that looks like this:



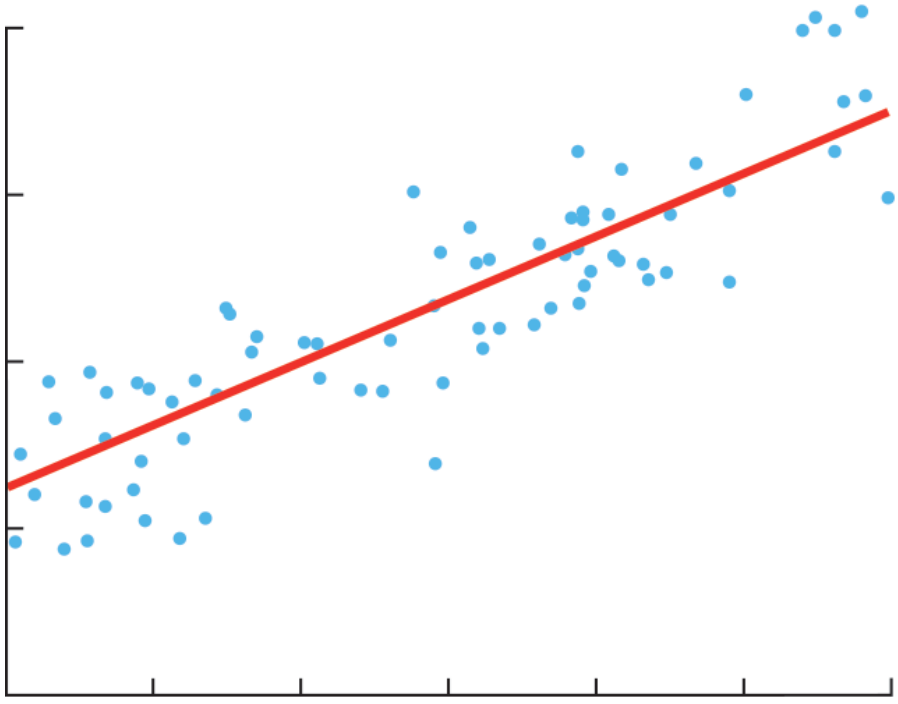
The y-axis is the amount of sales (the dependent variable, the criterion variable, which is the thing you're interested in) and the x-axis is the total rainfall (the independent variable, the predictor). Each blue dot represents one month's data—how much it rained that month and how many sales you made that same month.

Glancing at these data, you probably notice that sales are higher on days when it rains a lot. That's interesting to know, but by how much? If it rains 3 inches, do you know how much you'll sell? What about if it rains 4 inches?

Now imagine drawing a line through the chart above, one that runs roughly through the middle of all the data points. This line will help you answer, with some degree of certainty, how much you typically sell when it rains a certain amount.

Building a Regression Model

The line summarizes the relationship between x and y.



SOURCE HBR.ORG

This is called the regression line and it's drawn to show the line that best fits the data. In other words, explains Redman, "The red line is the best explanation of the relationship between the independent variable and dependent variable."

In addition to drawing the line, your statistics program also outputs an equation that explains the slope of the line and looks something like this:

$$y = 5x + 200 + \text{error term}$$

Ignore the error term for now. It refers to the fact that regression isn't perfectly precise. Just focus on the model:

$$y = 5x + 200$$

What this formula is telling you is that if there is no "x" then Y = 200. So, historically, when it didn't rain at all, you made an average of 200 sales and you can expect to do the same going forward assuming other variables stay the same. And in the past, for every additional inch of rain, you made an average of five more sales. "For every increment that x goes up one, y goes up by five," says Redman.

Now let's return to the **error term**. You might be tempted to say that rain has a big impact on sales if for every inch you get five more sales, but whether this variable is worth your attention will depend on the error term. A regression line always has an error term because, in real life, independent variables are never perfect predictors of the dependent variables. Rather the line is an estimate based on the available data. So the error term tells you how certain you can be about the formula. The larger it is, the less certain the regression line.

The above example uses only one variable to predict the factor of interest — in this case rain to predict sales. Typically you start a regression analysis wanting to understand the impact of several independent variables. So you might include not just rain but also data about a competitor's promotion. "You keep doing this until the error term is very small," says Redman. "You're trying to get the line that fits best with your data." While there can be dangers to trying to include too many variables in a regression analysis, skilled analysts can minimize those risks. And considering the impact of multiple variables at once is one of the biggest advantages of regression.

How do companies use it?

Regression analysis is the "go-to method in analytics," says Redman. And smart companies use it to make decisions about all sorts of business issues. "As managers, we want to figure out how we can impact sales or employee retention or recruiting the best people. It helps us figure out what we can do."

Most companies use regression analysis to explain a phenomenon they want to understand (e.g. why did customer service calls drop last month?); predict things about the future (e.g. what will sales look like over the next six months?); or to decide what to do (e.g. should we go with this promotion or a different one?).

A note about "correlation is not causation"

Whenever you work with regression analysis or any other analysis that tries to explain the impact of one factor on another, you need to remember the important adage: Correlation is not causation. This is critical and here's why: It's easy to say that there is a correlation between rain and monthly sales. The regression shows that they are indeed related. But it's an entirely different thing to say that rain *caused* the sales. Unless you're selling umbrellas, it might be difficult to prove that there is cause and effect.

Sometimes factors are correlated that are so obviously not connected by cause and effect but more often in business, it's not so obvious. When you see a correlation from a regression analysis, you can't make assumptions, says Redman. Instead, "You have to go out and see what's happening in the real world. What's the physical mechanism that's causing the relationship?" Go out an observe consumers buying your product in the rain, talk to them, and find out, what is actually causing them to make the purchase. "A lot of people skip this step and I think it's because they're lazy. The goal is not to figure out what is going on in the data but to figure out is what is going on in the world. You have to go out and pound the pavement," he says.

Redman wrote about his own experiment and analysis in trying to lose weight and the connection between his travel and weight gain. He noticed that when he traveled, he ate more and exercised less. So was his weight gain caused by travel? Not necessarily. "It was nice to quantify what was happening but travel isn't the cause. It may be related," he says, but it's not like his being on the road put those extra pounds on. He had to understand more about what was happening during his trips. "I'm often in new environments so

maybe I'm eating more because I'm nervous?" He needed to look more closely at the correlation. And this is his advice to managers. Use the data to guide more experiments, not to make conclusions about cause and effect.

What mistakes do people make when working with regression analysis?

As a consumer of regression analysis, there are several things you need to keep in mind.

First, don't tell your data analyst to go out and figure out what is affecting sales. "The way most analyses go haywire is the manager hasn't narrowed the focus on what he or she is looking for," says Redman. It's your job to identify the factors that you suspect are having an impact and ask your analyst to look at those. "If you tell a data scientist to go on a fishing expedition, or to tell you something you don't know, then you deserve what you get, which is bad analysis," he says. In other words, don't ask your analysts to look at every variable they can possibly get their hands on all at once. If you do, you're likely to find relationships that don't really exist. It's the same principle as flipping a coin: do it enough times, you'll eventually *think* you see something interesting, like a bunch of heads all in a row.

Also keep in mind whether or not you can do anything about the independent variable you're considering. You can't change how much it rains so how important is it to understand that? "We can't do anything about weather or our competitor's promotion but we can affect our own promotions or add features, for example," says Redman. Always ask yourself what you will do with the data. What actions will you take? What decisions will you make?

Second, "analyses are very sensitive to bad data" so be careful about the data you collect and how you collect it, and know whether you can trust it. "All the data doesn't have to be correct or perfect," explains Redman but consider what you will be doing with the analysis. If the decisions you'll make as a result don't have a huge impact on your business, then it's OK if the data is "kind of leaky." But "if you're trying to decide whether to build 8 or 10 of something and each one costs $1 million to build, then it's a bigger deal," he says.

Redman says that some managers who are new to understanding regression analysis make the mistake of ignoring the error term. This is dangerous because they're making the relationship between something more certain than it is. "Oftentimes the results spit out of a computer and managers think, 'That's great, let's use this going forward.'" But remember that the results are always uncertain.

As Redman points out, "If the regression explains 90% of the relationship, that's great. But if it explains 10%, and you act like it's 90%, that's not good." The point of the analysis is to quantify the certainty that something will happen. "It's not telling you how rain will influence your sales, but it's telling you the probability that rain may influence your sales."

The last mistake that Redman warns against is forgetting about the data.

Ask yourself whether the results fit with your understanding of the situation. And if you see something that doesn't make sense ask whether the data was right or whether there is indeed a large error term.