Correlation versus causation: Interpret regression results

In previous videos, you learned that correlation is not causation. In this reading, you will continue to explore the differences between correlation and causation so that you will be prepared to report regression results responsibly, honestly, and effectively.

What is correlation?

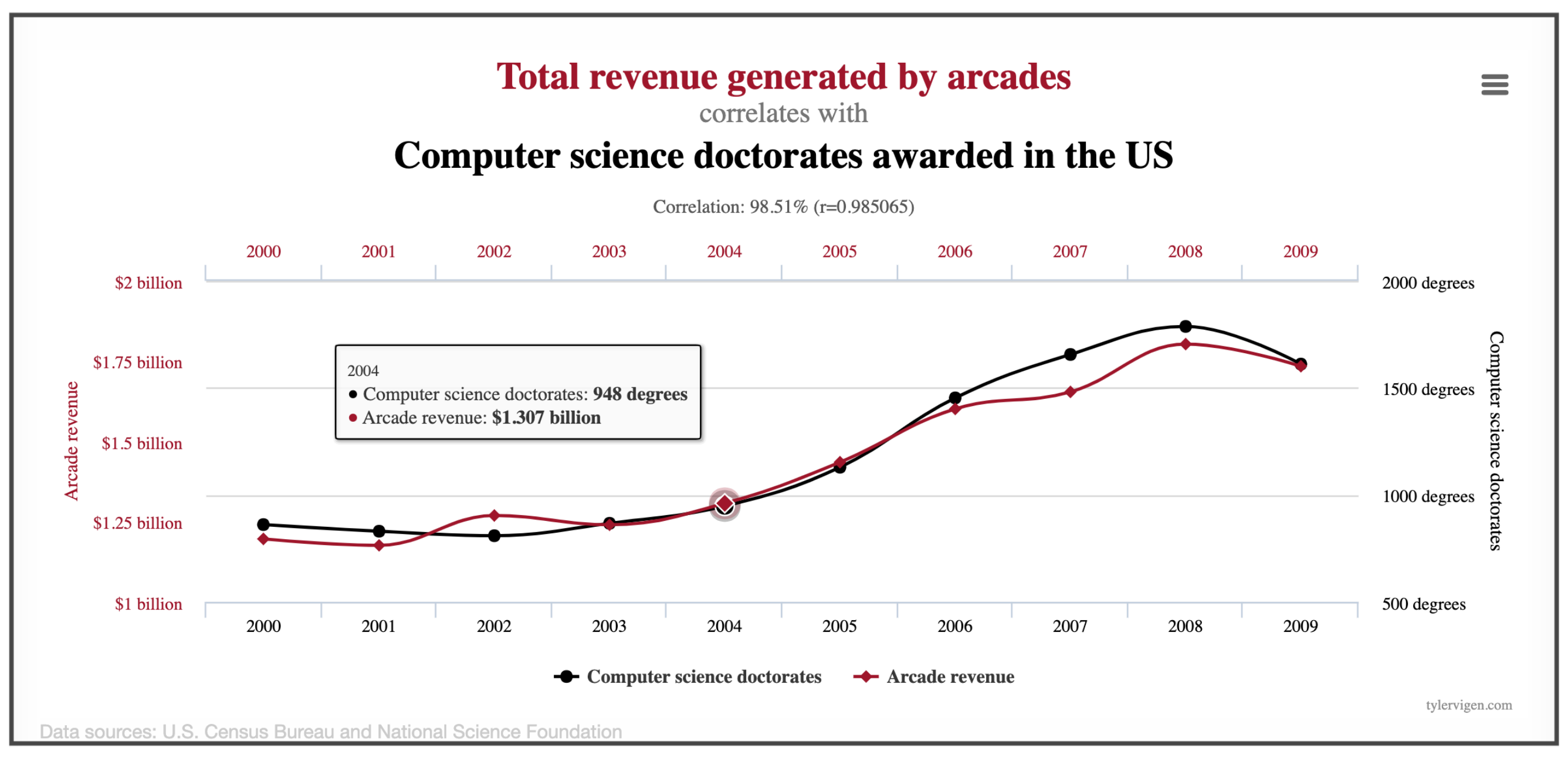
You might recall that there are two main kinds of correlation: positive and negative correlation.

- **Positive correlation** is a relationship between two variables that tend to increase or decrease together.
- **Negative correlation** is an inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa.

To generalize, **correlation** measures the way two variables tend to change together. There is a metric called the Pearson correlation coefficient that ranges from -1 to 1 that can measure the relationship between two variables.

Note that correlation is just observational. Two variables can be correlated–they tend to change together without one variable causing the other variable to change. In fact, there is an entire website and book, *Spurious Correlations*, devoted to documenting interesting and unexpected correlations between variables.

For example, here is a graph illustrating the correlation between total revenue generated by arcades and computer science doctorates awarded in the United States. Over time, computer science doctorates and arcade revenue increase at about the same rate. So, the graph definitely shows a correlation between the two variables, but it's pretty hard to argue that one causes the other.



Data sources: U.S. Census Bureau and National Science Foundation

It's difficult to claim causation

Previously you learned that **causation** describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way. Although this is an intuitive definition, proving causation requires a lot of particular circumstances to be met.

To argue for causation between variables, in general, you must run a **randomized controlled experiment**. The following are some key components of a proper randomized controlled experiment:

- You must control for every factor in the experiment.
- You must have a control group under certain conditions.
- You must have at least one treatment group under certain conditions.
- The difference(s) between the control and treatment groups must be observable and measurable.

Setting up a randomized controlled experiment is quite laborious and intensive. There are a number of requirements and factors not included in this reading, but there is a lot of information online and in academic research that you can explore on your own. Understanding the basics of causal claims allows you to responsibly report the results of your data analysis.

Correlation leads to interesting insights

When working as a data professional, you often do not have complete control of how the data is collected. You or your team might not be able to run a randomized controlled experiment. But, even if you cannot make causal claims, correlational research can still yield interesting results that have meaningful business implications.

Scenario 1: Optimizing athletic performance

Suppose a runner is training for a race. There are many ways to track health data—from built-in apps to paid-for apps. But, there are also so many factors that can contribute to the runner's performance—how much water they drank, how sore their muscles are on a given day, the weather, how much sleep they got, what equipment they are using on race day, and the clothes they are wearing. It's very hard to claim that any one factor would make or break their race time. But, over time, one might observe patterns in how sleep, soreness, water, clothing, and other factors tend to correlate with performance. This is why athletes can be so particular about brands of equipment, their diet, and pre-race day routines.

Claims you can make (correlation)

- When the runner drinks more water the day before a race, they tend to have more stamina.
- When the runner doesn't run long distances the week before a race, they tend to feel better on race day.

Claims you cannot make (causation)

- Drinking more water the day before a race causes the runner to run faster.
- Not running long distances the week before a race causes the runner to run faster.

Scenario 2: Improving food quality

Perhaps you're a new chef at a restaurant or you're cooking for yourself or family. Every time you make a dish, there are a lot of variables: what pan was used, when the ingredients were purchased, if the ingredients are in season, and how hungry everyone was. Any one of these factors can change how "good" the dish is. But, this data is valuable regardless of causal claims. Over time, you can hone your cooking skills for this particular dish to ensure better food quality.

These are just two examples where gathering data to understand correlation between factors can drastically improve outcomes. The same principles can be applied on a larger scale, with big data, and in different industries, depending on the desired outcome you want to optimize.

Claims you can make (correlation)

- When I use fresher ingredients, the final dish tends to taste better.
- When I am very hungry, the final dish tends to taste better.

Claims you cannot make (causation)

- Using fresher ingredients makes the dish taste better.
- Being hungrier makes the dish taste better.

Key takeaways

- Claiming causation requires specific circumstances that are often not within your control.
- Correlation analyses are an incredibly useful tool for data professionals, and can provide interesting insights and actionable next steps.

Mark as completed