Correlation vs Causation

Correlation is a statistical technique that tells us how strongly the pair of variables are linearly related and change together. It does not tell us why and how behind the relationship but it just says the relationship exists.

Causation takes a step further than correlation. It says any change in the value of one variable will cause a change in the value of another variable, which means one variable makes another happen. It is also referred to as cause and effect.

This segment is about the use of regression to infer causal relations and answer causal questions. The causal questions are important and they arise in many real-world problems, especially in determining the efficacy of medical treatments, social programs, and government policies, and business applications.

For example,

1. Does a particular drug cure an illness?
2. Do more lenient gun laws increase murder rates and other types of crime?
3. Does the introduction of a new product raise the profits of a company?

Regression uncovers correlation or association between outcome variables and regressors. However, correlation or association does not necessarily imply a causal relation, unless certain assumptions hold. The association between outcome variable Y and a regressor D formally means that D predicts Y, namely the coefficient in the linear regression of Y on D is not zero, or, that, more generally, the conditional expectation of Y given D is not constant. This is what the regression analysis gives us.

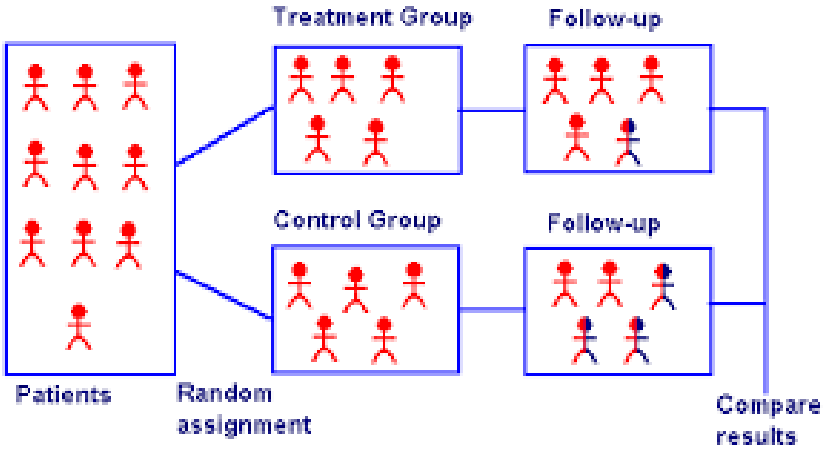
So why does not the association between Y and D necessarily imply that D causes Y?

Here is a contrived example. Suppose we conduct an observational study on people's use of painkillers and pain. We record the outcome variable Y, which is whether a person is in pain, and a regressor variable D, which is whether the person has taken a pain-killer medicine. It is likely that people take these pills when they are in fact in pain, and so we are likely to find that D indeed predicts Y, so that the regression coefficient of Y on D is positive. So Y and D are associated, but D does not cause Y, as we know

from the clinical trials in medicine that establish that painkillers actually work to remove pain.

Another example, Ice cream sales are correlated with homicides in New York (Study). As the sales of ice cream rise and fall, so do the number of homicides. Does the consumption of ice cream cause the death of people? No. Two things are correlated doesn't mean one causes the other. Correlation does not mean causality or in our example, ice cream is not causing the death of people.

An ideal way to establish a causal relation from the regression analysis is to rely on data from a randomized control trial (RCT). In RCT, we have the so-called random assignment or exogeneity, namely the treatment variable D is assigned or generated randomly, that is, independently of potential outcomes. Specifically, in randomized control trials or experiments, we have a group of treated participants and the untreated ones, the latter called the control group, and there are no systematic differences between the two groups when the trial starts.



After the trial is completed, we record the outcomes of interest for participants in the two groups, and carry out the regression analysis to estimate the regression equation:

$$Y = \alpha + \beta D + u, \quad \mathrm{E}[u(1, D)] = 0$$

Here the regression coefficient β indeed measures the causal impact of the treatment on average outcomes; and so β is called the average treatment effect (ATE).

Randomized control trials are also widely used in business applications under the name of AB testing, to evaluate whether new products and services raise profits.

Under random assignment, we don't need to use any additional regressors or controls. However, we may use additional controls, to improve the precision of estimating the average treatment effect β. This is called the covariate adjustment method. Specifically, we may set up a linear or a partially linear model:

$$Y = \alpha + \beta D + g(Z) + \epsilon, \quad \mathrm{E}(\epsilon \mid D, Z) = 0$$

And can carry out inference on β using the inference methods that we have learned in this module. The randomized control trials represent the golden standard in proving that medical treatments and social programs work, and for this reason, they are very widely used in medicine and in policy analysis.

But what if we don't have access to the data from the randomized control trials? What if we work with a data set that is purely observational? Under what conditions can we claim that we have established a causal relation? A sufficient condition is that D, the variable of interest, is generated as if randomly assigned, conditional on the set of controls Z. This is called the conditional random assignment or conditional exogeneity. That is, conditional on Z, variation in D is as if it were generated through some experiment, as in a randomized control trial. In this case, we can also apply the partially linear regression model:

$$Y = \beta D + g(Z) + \epsilon, \quad \mathrm{E}(\epsilon|D, Z) = 0$$

and β indeed measures the causal effect of D on the average outcome, controlling for Z. We then apply the inference method that we have learned to estimate β and construct confidence intervals.

Under conditional random assignment/conditional exogeneity controls must be included to ensure that we measure the causal effect and not something else. Under pure random assignment, we do measure the causal effect and not something else whether or not we include the controls.

So, under conditional random assignment/conditional exogeneity regression doesn't cover causal effects. For example, recall our case study of the impact of gun ownership

on predicted gun homicide rates. We did find there that increases in gun ownership rates lead to higher predicted gun homicide rates, after controlling for the demographic and economic characteristics of various counties. If we believe that the variation in gun ownership rates across countries was as good as randomly assigned, after controlling for these characteristics, then we should conclude that the predictive effect is a causal effect. If we don't believe that this variation is as good as randomly assigned, even after controlling for all these characteristics, then we should not conclude that the predictive effect we've found is a causal effect.

Similarly, in another case study, we studied the impact of being female on the predicted wage. We did find that females get paid on average 2 dollars per hour less than men, controlling for education, experience, and geographical location. If we believe that the variation of gender conditional on these controls is as good as randomly assigned, then we should conclude that the predictive effect is a causal effect, which is the discrimination effect in this context. However, if we don't believe that this variation is as good as randomly assigned, then we should not claim that this effect is due to discrimination.

As you can see, making causal claims from observational data is difficult, and the success really depends on being convincing at persuading ourselves and others that the assumption of conditional random assignment or conditional exogeneity holds here. The burden of persuasion lies entirely on the data scientist if indeed she or he is willing to make the causal claim.

To see some amazing examples of correlation vs causation you can refer to the following link.