Regressions: An Economist Obsession

A basic statistical tool for distinguishing between correlation and causality

Rodney Ramcharan

READING IS AN IMPORTANT SKILL, and elementary school teachers have observed that the reading ability of their students tends to increase with their shoe size. To help boost reading skills, should policymakers offer prizes to scientists to devise methods to increase the shoe size of elementary school children? Obviously, the tendency for shoe size and reading ability to increase together does not mean that big feet cause improvements in reading skills. Older children have bigger feet, but they also have more developed brains. This natural development of children explains the simple observation that shoe size and reading ability have a tendency to increase together—that is, they are positively correlated. But clearly there is no relationship: bigger shoe size does not cause better reading ability.

In economics, correlations are common. But identifying whether the correlation between two or more variables represents a causal relationship is rarely so easy. Countries that trade more with the rest of the world also have higher income levels—but does this mean that trade raises income levels? People with more education tend to have higher earnings, but does this imply that education results in higher earnings? Knowing precise answers to these questions is important. If additional years of schooling caused higher earnings, then policymakers could reduce poverty by providing more funding for education. If an extra year of education resulted in a $20,000 a year increase in earnings, then the benefits of spending on education would be a lot larger than if an extra year of education caused only a $2 a year increase.

To help answer these types of questions, economists use a statistical tool known as regression analysis. Regressions are used to quantify the relationship between one variable and the other variables that are thought to explain it; regressions can also identify how close and well determined the relationship is. These days, running thousands of regressions has become commonplace and easy—although that was not always the case (see box)—and, in fact, it is difficult to find an empirical economic study without a regression in it. Other fields, including sociology, statistics, and psychology, rely heavily on regressions as well.

How to run a regression

To illustrate how a regression works, let's take a closer look at the problem of trying to determine the returns to education. The government collects data on people's education level and their subsequent earnings. But people go to school for a variety of reasons—some find it easier to learn than others or are just more motivated to stay in school longer. Others may be successful pursuing nonscholastic careers and may still achieve high earnings. These varied reasons for attending school may affect earnings, making it difficult to know whether the correlation between schooling and earnings represents a causal relationship or is driven by some other factor. People who find it easier to learn in school may also find it easier to learn on the job, resulting in higher earnings. Thus, the positive correlation between higher earnings and education levels may reflect innate aptitude,

THE MAGIC OF COMPUTERS

Initial conceptualizations of regression date back to the 19th century, but it was really the technological revolution in the 20th century, making desktop computers a mainstay, that catapulted regression analysis into the stratosphere. In the 1950s and 1960s, economists had to calculate regressions with electromechanical desk calculators. As recently as 1970, it could take up to 24 hours just to receive the results of one regression from a central computer lab—and that was after spending hours or days punching computer cards. One wrong punch (a misspelled control word or incorrect data value) would invalidate the whole effort.

rather than the effects of education. Before a regression is run, a theoretical model can help explain how and why one "dependent" variable is determined by one or more "independent" or "explanatory" variables. Positing that an individual's earnings depends on his or her level of education is an example of a simple model with one explanatory variable. A corresponding regression equation, assumed to be linear, would look like:

$$Y = a + bX$$

On the left-hand side is Y, our dependent variable, earnings. On the right-hand side are a, our constant (or intercept), and b, our coefficient (or slope) multiplied by X our independent (or explanatory) variable, education. The regression says in algebra that "earnings depend only on education and in a linear way"; the other explanatory factors, if there are any, are omitted.

But what if we think that the world is much more complicated and that a variety of factors might explain the impact of education on earnings? In that case, we would run a multiple–variable regression, which would look like:

$$Y = a + b1X1 + b2X2 + \ldots$$

Now, we have several X variables to help explain Y earnings—like ability, intelligence, age, education, marital status, and parental education. The b coefficients simply measure the impact of each of these variables on earnings, assuming the other variables are constant.

Regressions quantify the relationship between one variable and others that are thought to explain it.

Smarter is richer?

Let's try running a regression on the basis of the theory that hourly wages (our dependent variable) depend on the level of education (our explanatory variable). We'll assume that another possible explanatory variable—aptitude, as measured by intelligence quotient (IQ) tests—has no effect on wages separate from any effect it may have through education. We plug in all of the data on earnings and education levels. We run the regression and find:

$$Y = 5.40 + 1.06\ EDU$$

The b coefficient tells us that an additional year of education is associated with a $1.06 increase in the hourly wage. And for those with no education ($EDU = 0$), the constant indicates that the average wage is $5.40 per hour.

But what if we put IQ in the equation—that is, assume that earnings depend on both the level of education and IQ? We plug in the data on IQ test results and find:

$$Y = 5.40 + 0.83EDU + 0.001IQ$$

We learn that individuals who performed better on IQ tests also had higher hourly wages. Moreover, while the impact of education on wages remains positive, it is about 27 percent smaller than if we hadn't included IQ results (the 27 percent comes from the difference in the coefficients: 100(1.06–0.83)/0.83). The implication is that we previously overestimated the effect of education on wages because we did not take into account the influence of IQ, which is correlated with education.

Potential pitfalls

Despite their benefits, regressions are prone to pitfalls and often misused. Take the following four leading difficulties.

Omitted variables. It is necessary to have a good theoretical model to suggest variables that explain the dependent variable. In the case of a simple two-variable regression, one has to think of the other factors that might explain the dependent variable. In our example, even when IQ is included, the correlation between education and earnings may reflect yet some other factor that is not included. That is, the individuals in the sample may still be different in some "unobserved" way that explains their subsequent earnings, possibly through their education choices. Individuals from wealthy families usually have better access to education, but family wealth may also create more connections in the labor market, leading to higher earnings. Thus, parental wealth may be another variable that should be included.

Reverse causality. Many theoretical models predict bidirectional causality—that is, a dependent variable can cause changes in one or more explanatory variables. For instance, higher earnings may enable people to invest more in their own education, which, in turn, raises their earnings. This complicates the way regressions should be estimated, calling for special techniques.

Mismeasurement. Factors might be measured incorrectly. For example, aptitude is difficult to measure, and there are well-known problems with IQ tests. As a result, the regression using IQ might not properly control for aptitude, leading to inaccurate or biased correlations between education and earnings.

Too limited a focus. A regression coefficient provides information only about how small changes—not large changes—in one variable relate to changes in another. It will show how a small change in education is likely to affect earnings but it will not allow the researcher to generalize about the effect of large changes. If everyone became college educated at the same time, a newly minted college graduate would be unlikely to earn a great deal more because the total supply of college graduates would have increased dramatically. FD

RODNEY RAMCHARAN is a professor of finance at the Marshall School of Business, University of Southern California.