Analyzing Not Just for Correlation But for Causation

ReInforcedCare™

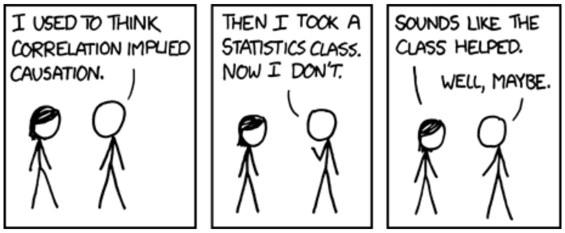
Objective: This article is aimed at readers who would like to improve their ability to interpret and evaluate research findings, especially findings that, relying on statistics, are presented as evidence of cause-and-effect relationships. Readers should come away better able to recognize the choices, often subjective choices, that such research entails. Readers will learn how findings about causation are obtained and will hopefully be better equipped to critique them.

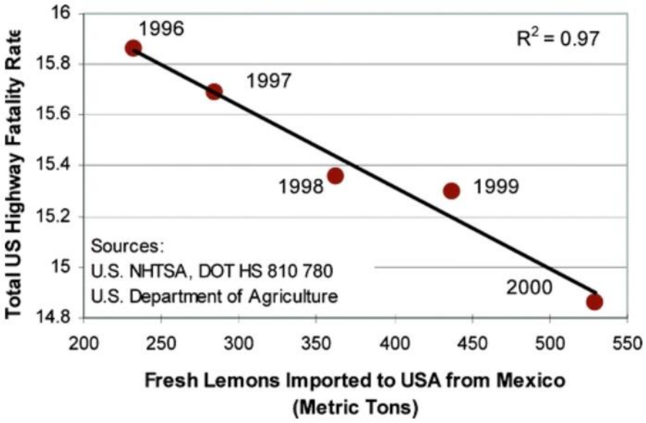
Introduction

Often we are interested not just in predicting an outcome or finding good indicators of it, but in causality: the dynamics of cause and effect. We want to know which are the causal factors that, if changed, will bring about a change in the outcome. And in performing data analysis we look for statistical methods that can shed light on this.

Randomized clinical trials are the definitive method for uncovering cause and effect. But so often we find ourselves as analysts or consumers of *observational* rather than experimental data. In these frequent instances, we must be careful not to equate correlation (or any type of link or relationship) with causation. This is wryly demonstrated by a number of perhaps familiar sources. There are maxims such as "The more fire engines on the scene, the greater the fire's damage."

Randall Munroe's classic cartoon from his xkcd series emphasized the point beautifully.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.



Sources:
U.S. NHTSA, DOT HS 810 780
U.S. Department of Agriculture

So did Derek Lowe's data graphic inexorably linking Mexican lemon imports and US traffic fatalities.

If we are serious about trying to understand causation while avoiding these kinds of errors, there is a hierarchy of analyses we might conduct. They range from the most basic and superficial, or one might say most naïve, to the most sophisticated and best able to address questions of causality. None is perfect, and all involve some choices on the part of the analyst. We will look at one set of four methods that span this spectrum and that are part of the toolkit of many analysts. Roughly in increasing order of complexity, they are

Correlation………Partial Correlation………Multiple Regression………Path Analysis.

Please note that a study need not use any of these methods in order to be valid, nor does the use of the latter two guarantee the soundness of findings.

Case Study

Throughout this article, for demonstration purposes we employ a fictional example, connecting **affluence** and **longevity**. In our fictional example, more affluent Americans tend to live longer. Hypothetically, this might be because greater **affluence** means…

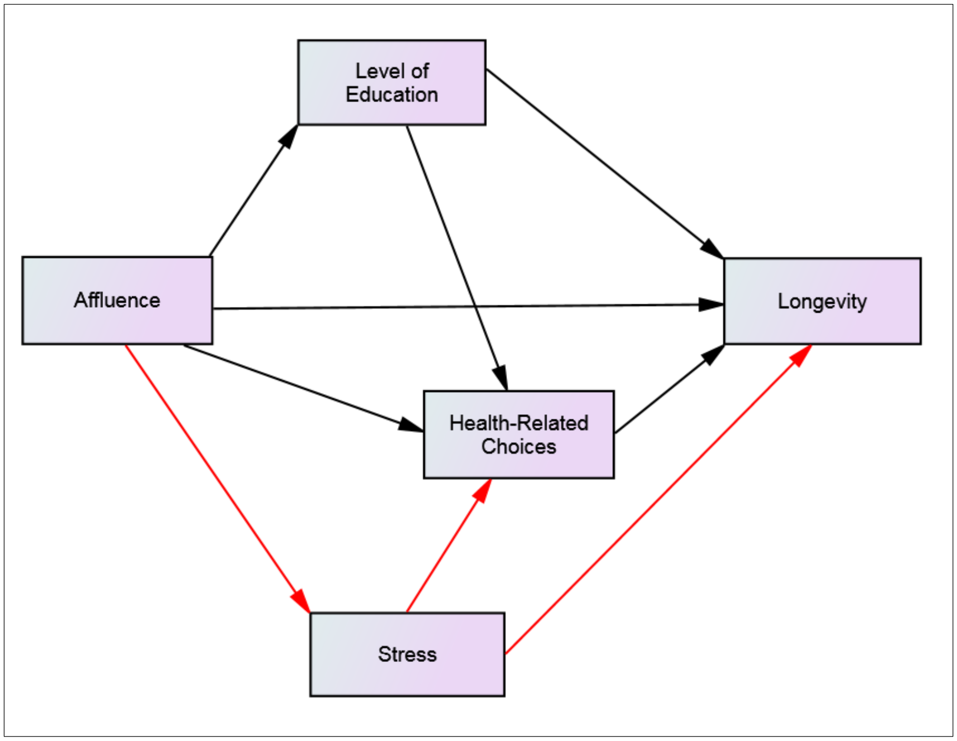
- Lower levels of **stress**…
  - which itself tends to lead *directly* to greater **longevity**.
  - which also *indirectly* leads to a longer life by fostering better **health-related choices**.
- Higher **level of education**…
  - which has its own indirect effect, tending to steer people toward better **health-related choices**, which then enhances **longevity**.

With five different variables in our sights, it helps to form a visual model of our imagined relationships. Such a model might initially look like Figure 1 below.[1]

- The black arrows show what we assign as (believe to be) positive associations: "more of this means more of that," as in greater **affluence** means higher **level of education**.
- Red is used for a negative or inverse relationship: we assume that more **stress** tends to *decrease* **longevity**.

[1] At this point readers may object that the **affluence**-**level of education** arrow should be bi-directional, particularly with the spate of reports in recent years about the greater earning power that results from a college degree. We include just the one-directional arrow for the sake of simplicity.

Figure 1. Initial Path Diagram (Fictional)



Note that in this view of the causal paths, **health-related choices** forms a hub of activity because it not only exerts a direct effect, but mediates three indirect effects. That is, we are making subjective choices in supposing that at least part of the reason why **affluence**, **level of education**, and **stress** matter for **longevity** is that they all affect **health-related choices**.

Suppose we had valid, reliable measurements of each of these five variables. And suppose these measurements were all on a less-to-more scale such as from 1-10 or 1-100. What analyses might we try so as to assess these relationships and our ideas about causes and effects? We would like to know not only which arrows truly belong, but whether the links are positive or negative and then how strong each path is.

We start with the most basic method. For complex questions related to cause and effect, an analyst may well need to make use of a more sophisticated method. Then again, if an exploration is in an early stage, or if causes and effects are known or not of interest, then a basic method may be the best choice indeed.

Correlation

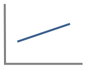
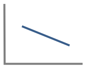
A correlation coefficient tells whether, on the surface, a relationship between two variables is negative or positive, and how strong it is. In theory, the result, denoted by "*r*", ranges from -1 (strongest possible negative relationship) to 0 (no relationship) to 1 (strongest possible positive relationship). Often when we are studying a group of factors like those diagrammed above, we examine correlations in a matrix like the one below.

Table 1. Correlation (*r*) Matrix (Fictional)

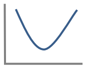
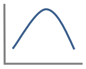
	Affluence	Level of Education	Health-Related Choices	Stress	Longevity
Affluence	*	.37	.44	-.66	.45
Level of Education	.37	*	.32	.06	.35
Health-Related Choices	.44	.32	*	-.58	.67
Stress	-.66	.06	-.58	*	-.50
Longevity	.45	.35	.67	-.50	*

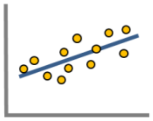
- Again we use red to indicate negative associations.
- The coefficient connecting **level of education** and **stress** is printed in grey. This is because for this exercise we are treating this correlation as essentially zero, indicating no relationship. We assign it no arrow in our path diagram. (Another subjective choice.)

Analyzing via correlation entails a notable limitation: it only considers associations that are linear, such as "more of this means more of that" or "more of this means less of that."

 or 

But what if such a pattern applies only up to a point? In that case the best fit line between one variable and another is going to be some sort of curve such as

 or  .



If such a curve fits best, then whatever number we obtain for *r* may be quite misleading. For this reason, good analysts routinely check scatterplots to see just what shapes relationships are taking, and they choose a method besides correlation if those relationships are nonlinear. (At right we have a very linear pattern.)

From Table 1 we can see initial indications that some links, such as between **affluence** and **stress** (-.66), are much stronger than others, such as between **affluence** and **level of education** (.37). But these results give us no idea which way the causal arrows might run. They also offer no help in distinguishing among different kinds of effects. We have assumed (Fig. 1) that **affluence** enhances **longevity** at least in part because **affluence** enhances **level of education**. As to this the correlations themselves are mute.

Partial Correlation

Virtually every day a researcher finds herself asking, “Can this relationship be *explained away* by something else?” Sisters’ and brothers’ height are surely correlated, but this can be explained away by father’s height. Are there analogous dynamics at work in our data? And so we turn to partial correlation, a technique that underlies many other statistical methods including those described below. It helps us to isolate a relationship, to purify it from the effects of confounding, “lurking,” or “nuisance” variables (like father’s height). Partial correlation answers questions such as “what is the correlation between **affluence** and **longevity** if **level of education** is held constant? (I.e., if it is controlled, adjusted for, or “partialled out,” or if its “influence is removed.”)

Suppose we found that, when we statistically controlled for **level of education**, the *r* between **affluence** and **longevity** became weaker, dropping from .45 to .20. This would tell us that if it weren’t for the mediating effect of **level of education**, **affluence** would be less important to our outcome. **Level of education** would partly explain away the **affluence**-**longevity** connection.

Partial correlation has innumerable uses and can be quite a bit more informative than simple correlation. But with many relationships to assess, as with the case of five variables being considered, it can become quite a task to compute each relationship in turn while controlling all the other relevant factors. After all, we have identified nine paths worth investigating.

In addition, just determining which factors to control can be anything but straightforward. We often want to isolate two variables of interest from the confounding effects of other nuisance variables. But we have to take care not to control for variables “downstream” of our outcome. It wouldn’t make sense to control for anything that *resulted from* **longevity**, because we’d be forfeiting away some of the very information we seek. Similarly, we would want to avoid controlling for any other factors that might really be *alternate indicators* of the same things we are analyzing. Doing so would falsely water down our connections. Thus partial correlation, in any of its forms, should not be applied mechanically.[2]

[2] In *The Logic of Causal Order*, James A. Davis offers extremely helpful advice on this topic.

- Control for variables established earlier in time, since a later event can never cause an earlier one.
- Control for more objective variables, since subjective ones (happiness, political preference) will seldom determine objective ones (income, zip code).
- Control for more stable, generative variables (social class, primary diagnosis) rather than those that tend to be ephemeral or reactive (toothpaste brand preference).

But on this issue in particular, deep study of statistics and research methods pays great rewards, and we find that the learning process never really ends. In works such as *Multiple Regression in Behavioral Research* Elazar Pedhazur points out the many pitfalls that have befallen even seasoned and oft-published researchers. He shows

Interpreters of research findings who are not themselves analysts can often be just as effective as analysts in finding improvements to the choice of variables to control.

Multiple Regression

Through regression – with multiple predictors – we have the chance to assess the relationship between our ultimate *outcome* – here, **longevity** – and each predictor in our model, while simultaneously controlling for each of the others. Alternatively, we can exert our statistical control selectively, over subsets of variables, as we examine links with that outcome. Regression is a powerful tool both for explanation (as discussed here) and prediction. For data analysts around the world facing complex problems involving inputs and outputs it has been a tool of choice for the last 60 years.

With regression we might learn that, with controls applied across the board – that is, holding everything constant except one predictor at a time being assessed – the one with the strongest relationship to **longevity** is **health-related choices**. We might further determine a quantity that would have been difficult to come by with previous methods, namely, the strength of this association in concrete terms. E.g., that an increase of 10% on the (fictional) **health-related choices** score leads on average to a 6% longer life, or that a 20% increase in **stress**, to an 8% shorter life.

Still, there may be gaps. For example, we know there is overlap in the way that **affluence**, **stress** and **health-related choices** account for **longevity**. All these things are correlated, as we see in Table 1. To articulate all of the direct and indirect paths we have diagrammed, to quantify their strength, goes beyond what regression can sometimes accomplish. Even with its many features and its greater sophistication than simple or partial correlation, we may need to try many iterations of regression analysis, and still could be left with questions.

Path Analysis[3]

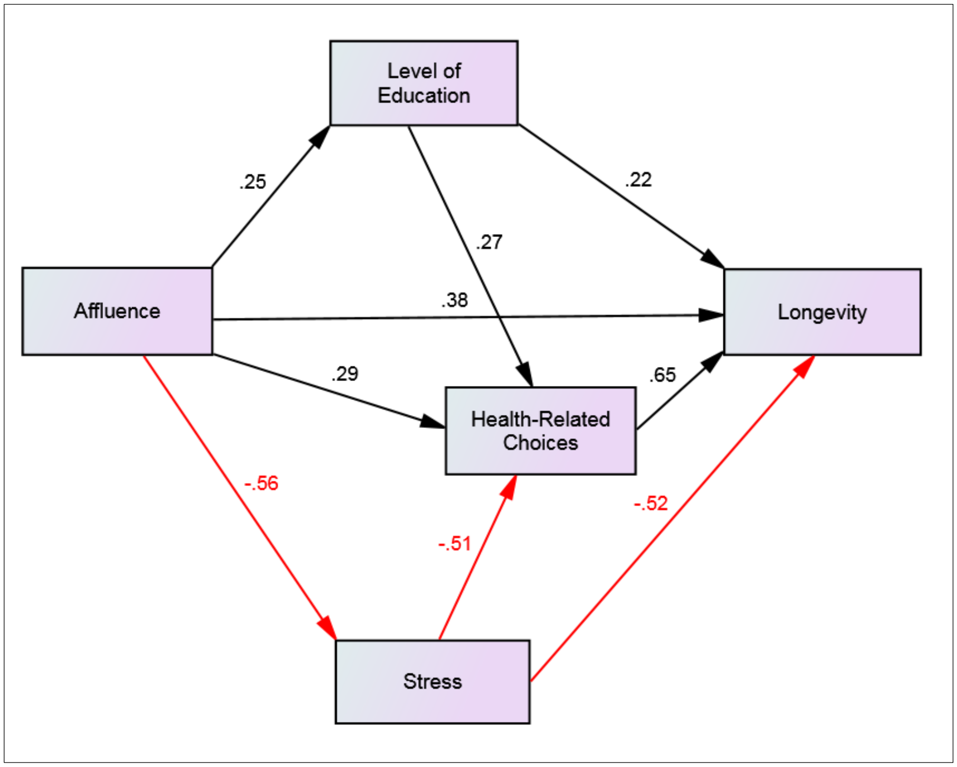
We cannot settle every question – that would be unrealistic. But if, as above, we specify a limited number of hypotheses about causes and effects, including possibly some indirect ones, then using path analysis we can quantify these using a model that may have great explanatory power.

Path analysis, like regression, is a world unto itself and has been the subject of many illuminating guidebooks. Again we explore just some highlights. We use a path diagram that now has been filled in with results drawn from our fictional data.

how improper or careless choices of control variables can lead, among other things, to findings that "have an air of fantasy about them."

[3] This type of analysis might also fit the term *structural equation modeling* (SEM), among others.

Figure 2. Path Diagram with Coefficients Derived from Data (Fictional)



Each *path coefficient* in the diagram is a type of *r,* one that describes its arrow while taking into account (controlling for) every other arrow shown. Each of these numbers can be interpreted like a partial correlation, describing each relationship when the other variables are held constant. (Remember that we *chose* not to draw the arrow from **level of education** to **stress**, and so this link is not among those controlled.)

The fictional diagram "confirms" that we were right about the hypothesized positive and negative signs of our black and red arrows, respectively. **Stress** indeed shows not just negative initial correlations, but negative path coefficients.

We were also arguably right in that each of our expected links shows at least moderate strength, with the weakest being .22 (not that .22 constitutes any special cutoff point; this will also vary by research topic). **Health-related choices** shows the strongest connection in the diagram, with a direct effect on **longevity** of .65.

The path from **level of education** to **longevity** is surprisingly weak. This *direct* connection turns out to be weaker than our original *r* value in Table 1 (.22 instead of .35), and weaker than those leading to **longevity** from the other three predictors (.38, .65, and -.52).

However, as expected, there is evidence that **level of education** plays an additional role, as expected, via an *indirect* effect through **health-related choices**. One of this method's attractions is that we can measure the strength of this indirect route. We simply multiply the coefficients for the two paths: .27 * .65 = .18. This indirect effect is nearly as strong as **level of education**'s direct effect of .22 (see inset).

One takeaway is that, since **health-related choices** appear to depend so much on upstream factors, efforts to improve these choices will have a greater effect on **longevity** if these other factors are also addressed – where possible.

We can use such findings, together with our substantive knowledge, to guide action. If our goal is to enhance population health, raising levels of **affluence**, although desirable, would hardly be feasible in the short term. The same may be true for **level of education**. But we may have more leverage when it comes to reducing **stress** levels or to raising the quality of individuals' **health-related choices**.

The diagram also allows us to see just how large a *total effect* each predictor shows, based on the sum of its direct and (downstream) indirect effects. Here we'll ignore negative signs, as we care only about each effect's strength.

For **level of education**, the total is .40; for **health-related choices**, .65; for **stress**, .85; and for **affluence**, the most generative of the four, .91.

If these were real data, the last two figures would be improbably high; a value of 1.0 would mean that a factor determined **longevity** completely.

Conclusion

"All models are wrong; some models are useful."

One of the most oft-quoted lines in the statistical literature is this half-serious one attributed to George Box (1919-2013). We hope that with this account of a hierarchy of tools for understanding causation, we have shared a useful model with you. We welcome your comments and questions at info@reinforcedcare.com. Check back with us for an example of causal modeling using real-world data on hospital readmission.

Recommended Reading

Keller, Dana K. (2005). *The Tao of Statistics: A Path to Understanding (with No Math)*. Thousand Oaks, CA: Sage.

Explores the landscape of statistical methods and specifies the kind of question each method helps to answer. Combines one-page summaries with Taoist-inspired cartoons. Enjoyable to browse or can be read in 45 minutes.

Davis, James A. (1985). *The Logic of Causal Order*. Thousand Oaks, CA: Sage.

Pages 5-15 are especially recommended as an introduction to studying causality with statistics. Shows in clear lay terms why the niftiest statistical methods cannot substitute for genuine thought about real-world relationships.

Prepared by:
Roland B. Stark, M.Ed., Senior Research Analyst
Laurie Courtney, Clinical Director
ReInforced Care, Inc.