



Last time Stephen Hampshire looked at the tricks our minds use to leap to (mostly correct) judgements about the causal processes going on in the world around us. In this article Stephen addresses the techniques that exist for applying a more rigorous analysis to causal processes, starting with what we actually mean by causation.

Part 2 patterns from the tapestry

What is causation?

It is actually surprisingly difficult to arrive at a satisfactory definition. To keep it very brief we can say that a causal link exists if, other things being equal¹, when the cause happens so does the effect. Or, alternatively, the presence of the cause makes the occurrence of the effect more likely. Normally we can also assume that the cause should precede the effect in time, unless we're in a science fiction film or you're talking to a philosopher or a quantum physicist.

Types of causes

Not all causes are the same. **Necessary** causes mean that the effect cannot happen without them, but they do not necessarily mean that the effect will happen. For example the presence of oxygen is a necessary cause of a fire, but there isn't always a fire when oxygen is around.

Sufficient causes are those that guarantee the effect if they happen. However a sufficient cause isn't necessarily the only cause, so other things could also have made the effect happen. An electrical fault may be a sufficient

cause for a fire, but it is not a necessary cause since there are lots of other potential causes (cigarettes, candles etc). In many cases a number of causes must combine in order to achieve sufficiency - understanding these **interactions** is a key part of successful causal modelling.

It's also important to consider the impact of potential causes that have a negative relationship with the effect - causes that **prevent** or **suppress** the effect.

Philosophies of causation

The necessary cause was the first to occur to philosophers looking for a definition. Hume first proposed the idea that a cause can be defined as something that, if it hadn't happened, then neither would the effect. This, known as counterfactual causation, is useful as far as goes, but it simply doesn't work in a lot of cases. If causality was a nice deterministic relation like this then a cause would always lead to its effect, but smoking doesn't always lead to lung cancer. Does that mean I should go back to 40 a day? Clearly something doesn't add up.

Nowadays the most popular way of dealing with causal thinking is probabilistic

causation - in other words the presence of a cause, such as smoking, increases the probability of an effect rather than necessarily leading to it. This can capture the idea of multiple potential causes, with effects sometimes not following from causes, and the interactions between causes to achieve sufficiency. Depending on your philosophical outlook, probabilities reflect either the inherently uncertain nature of causal relationships or our shallow knowledge about systems that are fundamentally deterministic but very complex. Probabilistic thinking is formalised in the techniques that form the basis of most causal modelling in practice, such as Structural Equation Modelling.

A probabilistic approach to causation lends itself perfectly to these techniques. It reflects the fact that causal links can be of different strengths, and it is the relative strength of different potential drivers that causal research is normally trying to investigate. Most statistical models are mainly used to provide estimates of **effect size**, as this is what will aid in decision making. In other words, **if I do X how much more likely is Y? Or if we increase X by 10% what will happen to Y?** Sometimes not enough care is

A: This is not just a phrase, the ceteris paribus assumption is fundamental to establishing the nature of any complex causal relationship



taken to assume that the model itself is valid, which can lead to some erroneous answers.

You may be wondering what we mean by probabilities. If the thought of coins flipping and the lottery are giving you cold sweats don't worry, it's really pretty easy. Probabilities are based simply on how often things happen. If 20% of non-smokers get lung cancer and 70% of smokers get lung cancer then you have a 20% chance of getting lung cancer if you don't smoke and a 70% chance of getting lung cancer if you do. Of course these probabilities can be "wrong" in the individual case², but they are true on average. More powerfully it may be possible to show that every step up in terms of the number of cigarettes smoked daily results in a X% increase in the probability of getting lung cancer.

The language of probabilities can get very complex, but fortunately the graphic techniques that have been developed to illustrate and aid causal thinking tend to be much easier to interpret. If all we had were equations then even a mildly complex causal structure would be very difficult to envisage:

$$Z = B_5V + B_6Y + B_7X$$

$$Y = B_1V + B_2W + B_3X$$

$$X = B_4W$$

...where the various Bs represent effect

sizes and the other letters are variables. The same information can be neatly and intuitively represented at the bottom of the page. It's a bit like the difference between a map and several different sets of directions.

Statisticians have been very wary of causal claims, preferring to use the term "association". Because of this techniques for the formal expression of causal thought are only just being developed. The reluctance to consider causality explicitly has hampered researchers in terms of statistical techniques and has resulted in much research, paradoxically, that has placed inappropriate reliance on statistical techniques that aren't up to the job rather than an emphasis on solid study design. The status quo in social science modelling is a bad compromise under attack from not only traditional statisticians but also experts in cutting-edge causal reasoning.

Building causal research studies

So much for the theory, where does that leave us in terms of making decisions in the real world? Formal causal research is still young, and often seems primitive compared to our own ability to reach common sense judgements. Judea Pearl, one of the leading thinkers on causality, has commented: "We all understand that the rooster's crow does not cause the sun to rise, but even this simple fact cannot

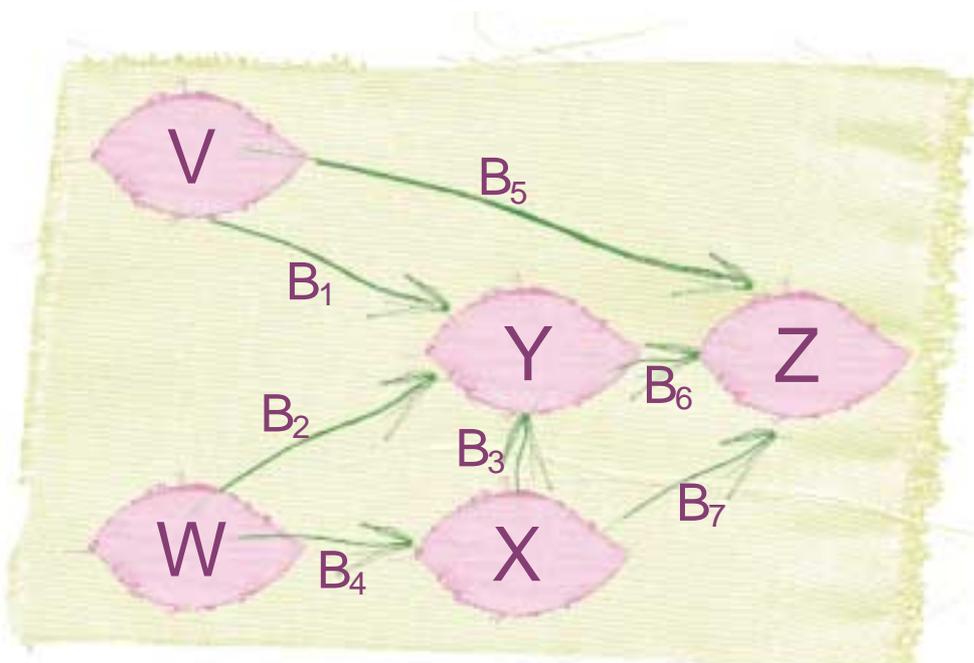
easily be translated into a mathematical equation." [1]

The late 19th century discovery of correlation as a technique for measuring the strength of relationship between two variables represented the first time we were able to objectively assess the strength of such links. Notice that there is no suggestion of causation, only a relationship (and even that may be spurious). Traditionally, and many statisticians still argue this, the only way to examine causation with statistics is by means of a randomised experiment.

Experiments

Randomised experiments are simple in theory, but often difficult or impossible in practice. The principle is that in order to study the effect of the variable we are interested in we must make sure that that is the only thing that varies between a test group and a control group (ceteris paribus again). We can do this by randomly allocating people to one group or the other.

Drug trials work in this way - we split our "guinea pigs" in half at random and then give one group the drug to be tested and the other group a placebo or a current drug. Participants don't know which group they are in. Ideally we would also avoid the researcher being aware which group was which - a "double blind" experiment.



Nonetheless many studies do reach causal conclusions without using randomised experiments, including, for example, the initial studies linking smoking and lung cancer. Ethical considerations make it impossible to randomly allocate people to undergo something which is suspected of being a serious health hazard.

When first published the smoking studies

B: Stephen Jay Gould's "The Median isn't the Message" makes this point very powerfully after discovering he had a predicted lifespan of eight months

[http://www.cancerguide.org/median_not_msg.html]



were attacked by the tobacco industry, as you might expect. What seems more surprising now was that they were also attacked by a number of prominent statisticians on the basis that the supposed causal link might be spurious, resulting from an unmeasured variable that caused both smoking and lung cancer. It was suggested that there might be a genetic predisposition both to enjoy smoking and to be susceptible to lung cancer.

This is a perfectly plausible hypothesis. It is accepted that other types of cancer have a genetic factor in their causes, and it is also suspected that there is a genetic predisposition to enjoy smoking. So how was it disproved? Not with regression or factor analysis but by conducting large-scale studies with identical twins where only one twin smoked - capitalising on a natural experiment that would rule out a genetic factor. A remarkably difficult exercise, but the only sure way to rule out the alternative hypothesis.

Statisticians, unlike many research practitioners, are very wary about the use of certain techniques (notably regression and factor analysis) in a causal context. The misuse of such techniques is probably what makes them uncomfortable with the use of causal terminology in any context other than randomised experiments. The result is described by Pearl:

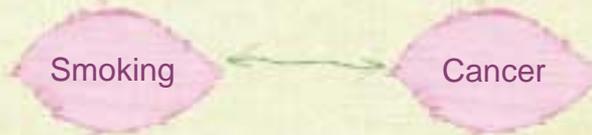
"...in the bulk of the quantitative statistical literature, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations." [2]

The origin of all science is the desire to know causes, and the origin of all false science and imposture is the desire to accept false causes rather than none; or, which is the same thing, in the unwillingness to acknowledge our own ignorance.

William Hazlitt

Smoking and lung cancer, alternative hypotheses:

The observed correlation



was consistent with two competing hypotheses (plus others!). More work was needed to support the hypothesised link.



The reason it is so difficult to restrain social scientists from using regression is that it seems to do exactly what we need - it determines the unique effect of a number of causes on an outcome variable, assessing statistically the effect each cause would have if all the other causes were held constant. When experiments are not feasible, models are the best thing we've got to work with, and multiple regression is what most people plump for most of the time.

David Freedman is one of the most prominent critics of the automatic and thoughtless use of regression models in causal studies. He points out that the main distinguishing feature between compelling work like Snow's cholera study (covered last time) and less persuasive studies is not the techniques

used but the "...investment of skill, intelligence, and hard work..." [3]. He is suspicious, justifiably, of the ability of regression models to statistically control for the complexities of a real world causal relationship.

Interestingly his negative view of the standard regression approach to causal studies is echoed by some of the leading lights in the development of causal modelling techniques: *"In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled [i.e. non-experimental] studies."* [4]

Clark Glymour, an expert on causal modelling in psychology, is particularly damning of the use of such methods. Discussing reactions to "The Bell Curve" which used standard tools of social science research to reach some very uncomfortable conclusions he comments:

"The unstated problem for many commentators is how to reject the particular conclusions of The Bell Curve without also rejecting the larger enterprises of statistical social



science....The hard issue is whether the methods of large parts of social science are bogus, phony, pseudoscientific. They are. The other hard issue is whether there are better methods....There are." [5]

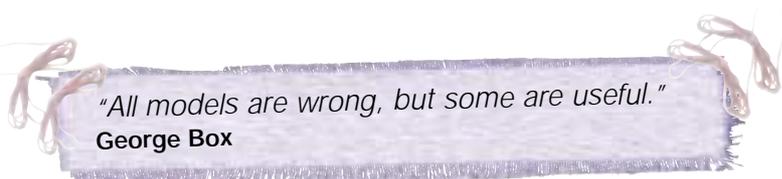
Drawing reliable causal conclusions

So the uncomfortable truth is that the familiar tools used by the majority of causal studies are fatally flawed. The substitute should either be much more robust new techniques (and thinking) or carefully designed experiments. Failing that we must treat any prospective causal findings very cautiously indeed. In a classic discussion, Sir Austin Bradford Hill clearly outlined his criteria for moving from association to causation, including:

- **Strength**
- **Consistency**
- **Specificity**
- **Temporality**
- **Dose-response curve**
- **Plausibility**
- **Coherence**

His conclusion is worth repeating in full, as it underlines the fact that the key to drawing reliable conclusions lies not in the statistical techniques we use, but in the way we think about what the results can and cannot tell us:

"I do not believe...that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we



"All models are wrong, but some are useful."
George Box

can accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do...is help us make up our minds on the fundamental question - is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?

No formal tests of significance can answer these questions. Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us in the likely magnitude of those effects. Beyond that they contribute nothing to the 'proof' of our hypothesis." [6]

This has several key applications to the building of causal models:

1. "Fit" statistics tells us only that there is no significant evidence against our model
2. Statistical significance means that the size of an association is too big not to exist in the population, but that's all
3. Every model has alternatives which need to be considered. Many of these are mathematically **indistinguishable** - why have we preferred one over the others?

Conclusion

Reaching causal conclusions from social data is, at best, a difficult problem. Most studies use inadequate techniques that are vilified by experts across the spectrum of statistics. In practice it is not always possible to use techniques that are more reliable, but careful thought needs to be given to studies in order to eliminate possible sources of error (for example it is easy to remove the possibility that a causal link should be reversed if the cause happens before the effect).

Where regression models are used for

causal inference the best reassurance will be if their predictions are confirmed in practice. If they are able to predict the outcome of an intervention then it is likely they have accurately caught a real causal relationship. It is this proven ability of a model to be useful that inspires confidence in it. It may still not be exactly right, but it is at least doing us some good! [S]

Bibliography

- [1] Judea Pearl, "Causality"
- [2] Judea Pearl, "Causal Inference in Statistics: A Gentle Introduction"
- [3] David Freedman, "From Association to Causation: Some Remarks on the History of Statistics"
- [4] Peter Spirtes, Clark Glymour & Richard Scheines, "Causation, Prediction, & Search"
- [5] Clark Glymour, "The Mind's Arrows"
- [6] Austin Bradford Hill, "The Environment and Disease: Association or Causation?"



Stephen Hampshire
 Development Manager
 The Leadership Factor

If you have any thoughts about this article you can contact Stephen at:

stephenhampshire@leadershipfactor.com