Necessary Condition Analysis: more value from data

By ***Jan Dul***

Data is often surprisingly short on real answers for managers, particularly if a problem is multi-causal. Sometimes we can measure all the relevant factors in a complex system to the fourth decimal place and yet still have no clear sense of which factors matter.

As a business school professor, I had long been frustrated that we know so much but can make so little use of what we know. Then about seven years ago, I had an idea for getting around this problem. It seemed to me that if we looked at the data from another angle, we could get much more useful answers. The solution lay not in sorting contributory factors, the normal result of a regression analysis, but in selecting only those factors that are absolutely essential to avoid failure: the necessary but not sufficient conditions.

A necessary but not sufficient condition is a factor that can't be left out – a "gotta have" as Americans say. It's a constraint that does not ensure success if it is present but guarantees failure if it is missing: a car stops moving if the fuel tank is empty; a brokerage collapses if trust is gone; contracts and trust are both necessary for a successful collaboration between firms (Sumo 2015).

Necessary Condition Analysis (NCA) transforms our ability to look at data from situations like this. Instead of expressing our findings in shades of grey as we would normally have to – this *could* happen; this is *likely* to result – NCA enables the analyst to look at the data and say, if you don't do this, you won't succeed. Or, if this metric does not pass that particular threshold, you will fail – full stop. It's a powerful statement that can give you more definitive, actionable answers than conventional statistics allows.

"A necessary but not sufficient condition is a factor that can't be left out – a "gotta have" as Americans say."

Three examples

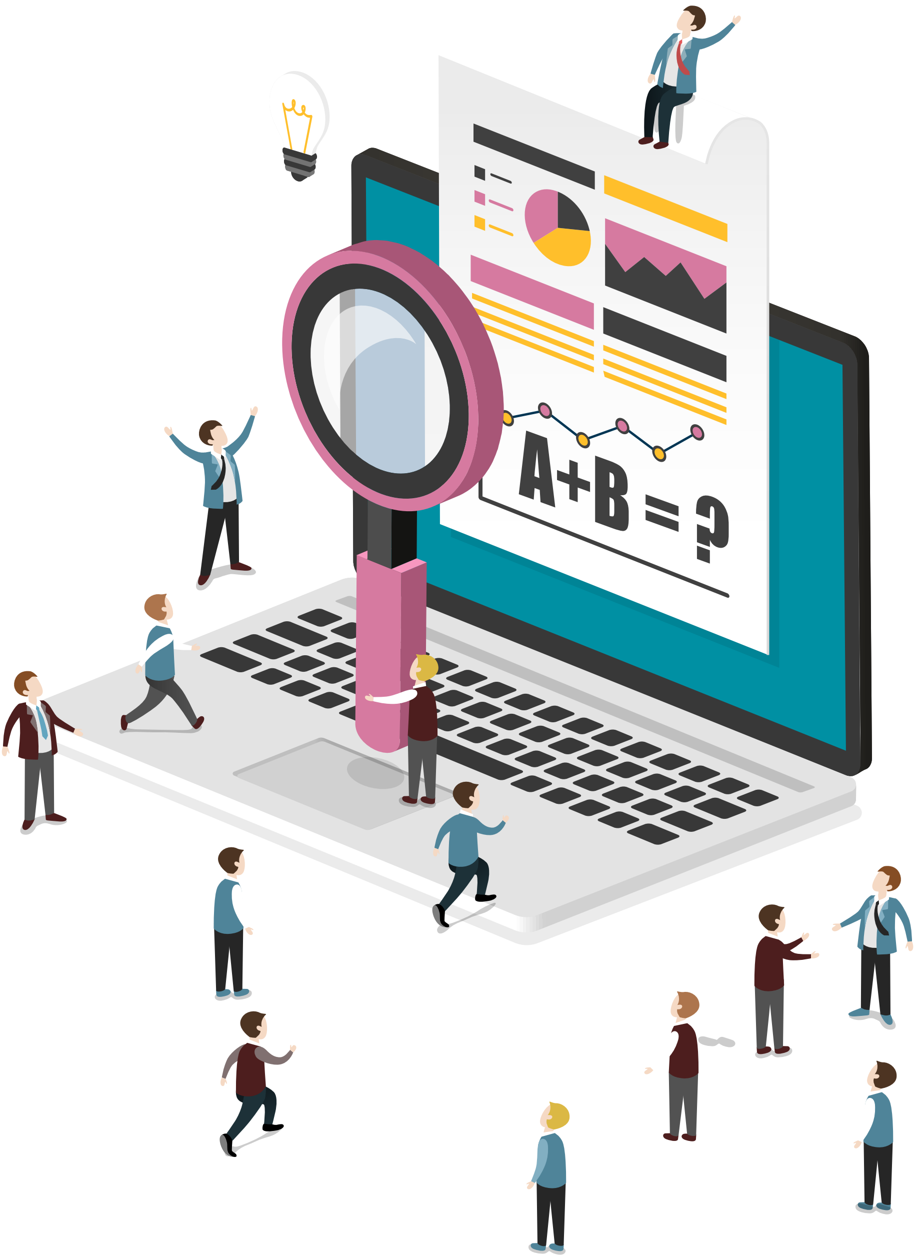
Three examples illustrate how NCA can be used. Consider the Quantitative Graduate Record Exam scores of 342 students applying for admission to the Berkeley Sociology Graduate Program in 2009 (Goertz & Mahoney, 2012; Vaisey, 2009). A traditional data analysis would show a high correlation between GRE score and admission. The data would be used to justify the conclusion that success is more likely when the GRE score is at least 620.

However, the *necessary but not sufficient* interpretation of these same data points show that a score of at least 620 was almost always a necessary condition for admission, the one exception being a student admitted based on a faculty member's explicit testimony as to the student's quantitative abilities (Vaisey, personal communication, July 2, 2014). Scoring 620 or over did not guarantee success – in the end, only a few of the applicants with that score (14 per cent) were admitted – but scoring below 620 practically guaranteed failure.

Traditionally, admissions officers might advise students to 'score high on GRE to increase your chances for success', but if they advised them based on an NCA they could say flat out that the applicant won't make it without the specific high GRE score – a much more useful statement to the would-be applicant.

Many kinds of questions can be resolved in a similar way. For instance, consider the case of a pool of salespeople evaluated through the Hogan Personality Inventory (HPI), a widely used tool to assess employee personality in order to predict organisational performance (Hogan & Hogan, 2007).

For Level 4, a high level of sales ability, it is necessary to have an ambition level of at least 30, as rated by Hogan.



About 20 per cent of the sales representatives in the sample do not reach this level of ambition, which suggests that people with ambition levels below 30 will certainly fail to reach a sales ability level of four of higher. The other sales representatives who do have an ambition level of at least 30 (about 80 per cent), have the potential to reach a high level of sales ability.

In the end, only 10 per cent of the sales representatives in the sample reach this level of sales ability. In other words, an ambition level of at least 30 is necessary but not sufficient to achieve a sales ability level of at least four. This means that an ambition level below 30 is a bar to achieving a high level of sales ability (Hogan, 2007).

This technique can be useful in multi-causal analysis as well. Consider what happens in an analysis of sales rep personalities that not only includes ambition but sociability, interpersonal sensitivity, and learning. A traditional multiple regression analysis indicates that the four personality traits together hardly predict performance, but NCA makes it clear that all four plots, including the plot for learning, indicate that each personality trait is a necessary but not sufficient condition for above-average sales ability (Hogan, 2007).

Is this really new?

When I explain my NCA methodology to practitioners, they are typically less than thrilled. They say – correctly – this sounds very logical and consistent with the way we actually think and work. Hasn't anybody done this before?

Necessary Condition Analysis: more value from data *(continued)*

By ***Jan Dul***

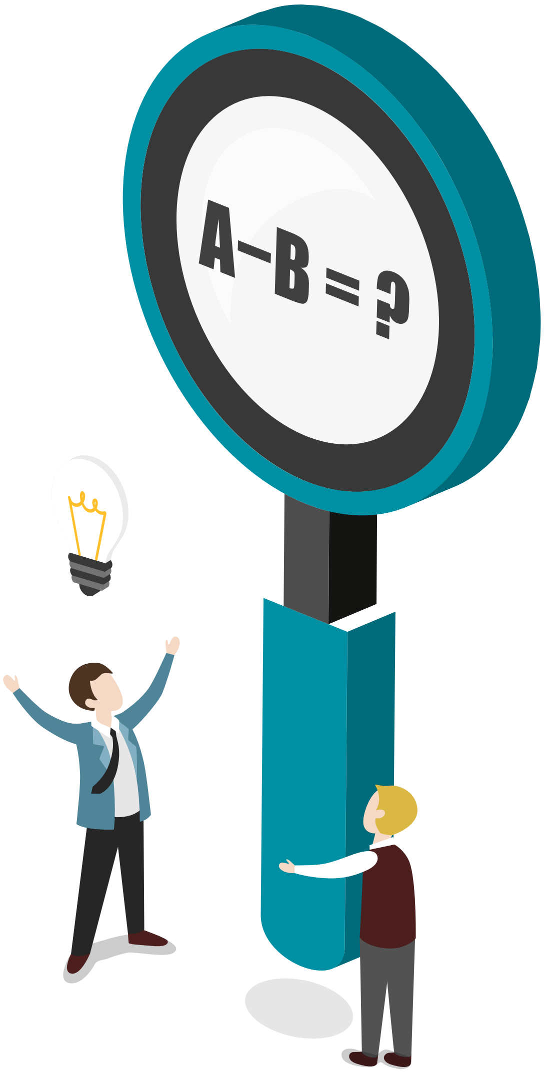
For a long time, I wondered myself whether someone had. The method seems so intuitive to me I thought somebody must have used it already, but I still haven't found evidence of it. Surprisingly little has been done to turn the data in a way that allows us to sort exclusively for necessary but sufficient conditions.

Analysts in the social sciences often make theoretical statements that include the phrase "necessary but not sufficient", but the tool they use to demonstrate their hypothesis is multiple linear regression (Y= a + b1X1+ b2X2 + b3X3 + ...), a technique that only works in demonstrating additive causes, a series of factors that add up to an outcome, not a series of factors that all have to be present to produce an outcome.

Researchers in technology and medicine have thought a little more about necessary but not sufficient conditions. The kinds of complex systems doctors and engineers work on often operate with a number of non-negotiable factors.

Diseases, for instance, often have necessary but not sufficient conditions, such as the HIV virus that is always a precursor to AIDS, or the virus that accompanies cervical cancer – factors that must be present if a disease is to develop. However, even in these disciplines, people think more often in terms of trying to predict outcomes.

I think the reason people don't think this way is that psychologically we are much more attuned to looking for a recipe to produce an outcome: you generally don't look up a recipe about



"The method seems so intuitive to me I thought somebody must have used it already, but I still haven't found evidence of it."

how *not* to bake a cake. Most of the time, we want a positive prescription.

Using this method is not difficult – it can even be done visually, by drawing lines through a standard scatterplot X-Y graph (see Panel). I have also developed a software package for people who need more precision, and even this is simple to use for anyone familiar with statistical software. Whichever method the analyst uses, however, the most important aspect is the initial concept that instead of trying to find for a recipe to create an outcome, you look for the factors that will stop you.

A special tool

Of course, NCA is not the best way to solve every problem. Regression analysis and the other traditional analytics tools remain very useful for a variety of purposes. NCA is instead a very useful tool for analysts to add to their toolbox if they understand what it can and cannot do.

It cannot solve the commonly observed problem that 'observational data cannot ensure causality'. In other words, a data pattern that is consistent with the causal hypothesis is not evidence of a causal connection, only that a connection is possible, when it is theoretically justified.

It is easy to plot storks versus human newborns in a given region (Box, Hunter, & Hunter, 1978) and find what looks like a necessary but not sufficient condition, but is not. Nor can NCA overcome the difficulties of limited sample sizes, and indeed may be somewhat more sensitive to measurement error than other approaches.

What NCA does very well is prove negative statements: 'if you don't have X you will not get Y.' If the data suggest a definitive consequence of an absence of something, NCA is probably the right tool to demonstrate it.

For many problems, the combination of being able to suggest a positive, using traditional analysis, and a negative, using NCA, makes it much easier to win more substantial and practical insights than through traditional methods alone.

Evangelising

These days, I am spending a lot of time on further developing the method, and discussing and spreading the use of NCA. I have given up some administrative duties at RSM in order to focus on NCA. I am refining it to make the concept more precise, adding the ability, for example, to calculate confidence intervals, and I am also doing a lot of evangelising, speaking at conferences and talking to groups about how my method works.

I'm convinced that a lot of people could benefit from this technique, in more ways than we imagine now. We talk a lot these days about living in the era of big data, but big data is only worthwhile if we use it to generate big insights. ■

Jan Dul *is Professor of Technology and Human Factors, Rotterdam School of Management, Erasmus University. For further information on the method, the software, or its applicability, the reader is invited to contact Professor Dul at* EMAIL *jdul@rsm.nl*

Spotting the necessary but not sufficient

The starting point for finding a necessary but sufficient condition using NCA is a scatter plot of data that plots X (the determinant and potential necessary condition) against Y (the outcome) for each case.

If you see a largely empty zone in the upper left-hand corner (with the convention that the X-axis is "horizontal" and the Y-axis is "vertical" and that values increase "upwards" and "to the right"), you may have found a necessary condition of X for Y. At this point, draw a ceiling line between the empty zone without observations and the full zone with observations.

Whether this condition matters depends on comparing the size of the empty zone with the entire area of the graph. The larger the empty zone compared to the rest of the observations, the larger the size of the necessary condition.

Determining the best position for the line is not always easy. I have identified eight ways to do it in my academic paper on NCA, *Necessary Condition Analysis (NCA): Logic and methodology of 'necessary but not sufficient' causality*, which will be published in an upcoming edition of *Organizational Research Methods*.