Multivariate and multiple NCA

Supplement to Dul, J. (2020) Conducting Necessary Condition Analysis. Sage publications.

Jan Dul

Version 1.1.0; June 3, 2021

Suggested reference:
Dul, J. ([year], [date]). Multivariate and multiple NCA. Supplement to 'Dul, J. (2020) Conducting Necessary Condition Analysis, Sage Publications'. Retrieved from http://erim.eur.nl/nca

In this supplement I explain why I have introduced in the book the terms *multiple NCA* and *multiple bivariate NCA* for performing a regular necessary condition analysis with more than one condition, whereas in the original 2016 NCA article (Dul, 2016) I have used the term *multivariate NCA* for the same analysis. The main reason is to avoid confusion about the interpretation of 'multivariate'.

Multivariate in statistics

The term *multivariate* is somewhat loosely used in statistics. The Cambridge Dictionary of Statistics (Everitt and Skrondal, 2010, pp. 293-294) gives a broad definition, stating that a *multivariate analysis* is a 'generic term for the many methods of analysis important in investigating multivariate data', and that *multivariate data* are 'data for which each observation consists of values for more than one random variable.' Following this definition, a univariate analysis is an analysis with one variable, and a bivariate analysis is a special case of multivariate analysis with two variables.

Multivariate in the 2016 NCA article

In the 2016 article (Dul, 2016), I introduced the term *bivariate NCA* for an analysis with one condition (and one outcome) and the term *'multivariate NCA'* for an analysis with more than one condition (and one outcome). This terminology fits the above definition of multivariate. In the article I present multivariate NCA as an analysis of several conditions *one by one*. The ceiling line for each condition is analysed separately in its two-dimensional plane (e.g. X_1Y, X_2,Y, X_3Y, etc.). For example, with two conditions and assuming straight ceiling lines the ceiling line for X_1 is $Y = a_1 + b_1*X_1$ (in the X_1Y plane) and the ceiling line for X_2 is $Y = X_2$: $a_2 + b_2*X_2$ (in the X_2Y plane). Because each ceiling line is analysed separately the scatter plots of X_1Y and X_2Y can be used to conduct NCA. In NCA's bottleneck table, the two separate analyses and ceiling lines are considered in combination when answering the question: "What levels of X_1 and X_2 are necessary for a given level of Y"? or "For given levels of X_1 and X_2 what is the maximum possible level of Y".

Multiple in the 2020 NCA book

In the book, I have introduced the term *multiple* NCA (or *multiple bivariate* NCA) instead of *multivariate* NCA when several ceiling lines are analysed. This should avoid confusion with the

term ‘multivariate analysis’. ‘Multivariate NCA is often interpreted as a single analysis with several variables *at the same time*. For example, in regression analysis a ‘multivariate analysis’ is understood as an analysis of multiple variables at the same time in one model (e.g., *regression* surface: $Y = a + b_1*X_1 + b_2*X_2 + ….b_i*X_i + \varepsilon$). However, such combined analysis is not used in NCA. NCA does *not* analyse a multi-dimensional ceiling in the multidimensional space (e.g., *ceiling* surface $Y = a + b_1*X_1 + b_2*X_2 + ….b_i*X_i$). NCA analyses single ceiling lines in multiple two dimensional planes (e.g., ceiling line 1: $Y = a_1 + b_1*X_1$ and ceiling line 2: $Y = a_2 + b_2*X_2$, ceiling line i: $Y = a_i + b_i*X_i$) . As shown in Figure 1, with two conditions there is a 3D ceiling surface in the three-dimensional space (X_1,X_2,Y, Figure 1, left) and NCA analyses the corresponding two ceiling lines in the two two-dimensional planes (X_1Y and X_2Y, Figure 1 right).

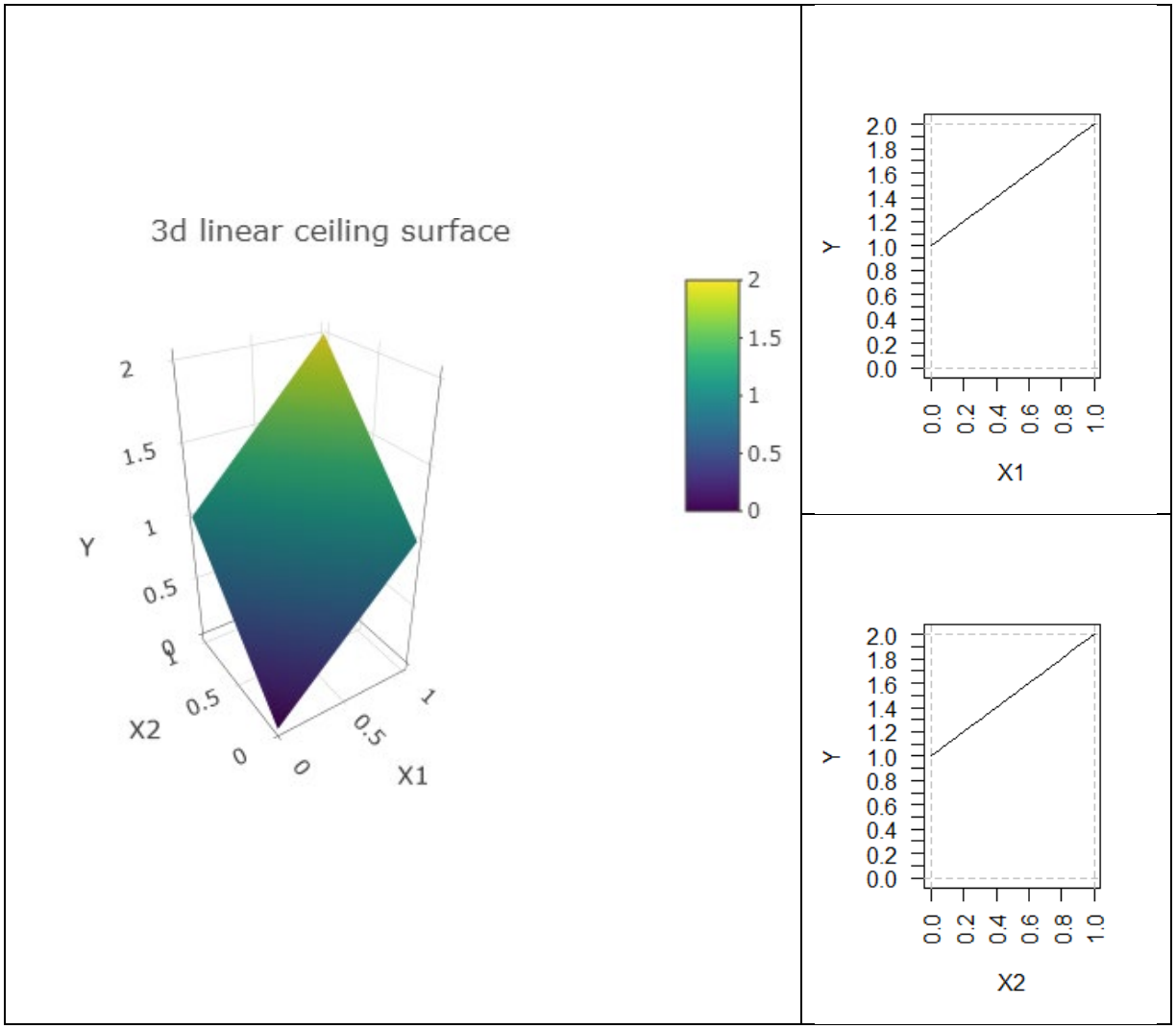

Figure 1. A linear multivariate three-dimensional ceiling ($Y = X_1 + X_2$) and its projection on the two two-dimensional XY planes (ceiling lines in X_1Y plane and in X_2Y plane).

The relationship between multidimensional ceiling and ceiling line?

The ceiling line is a line on top of the data in a XY scatter plot (in a two-dimensional XY plane). The ceiling line represents the maximum possible value of Y for a given X. A multidimensional ceiling in the three-dimensional space is a ‘blanket’ on top of the data. The three-dimensional

ceiling represents the maximum possible Y for a given combination of X_1 and X_2. In a multidimensional space with more than two conditions, the multidimensional ceiling represents the maximum possible Y for a given combination of all X's, which cannot be graphically imagined. NCA's ceiling lines are the projections of a multidimensional ceiling on the respective two-dimensional XY planes. Figure 2 shows a general three-dimensional non-linear ceiling surface ('the surface of a mountain'). The projections of the surface on the XY planes result in a non-linear ceiling lines.

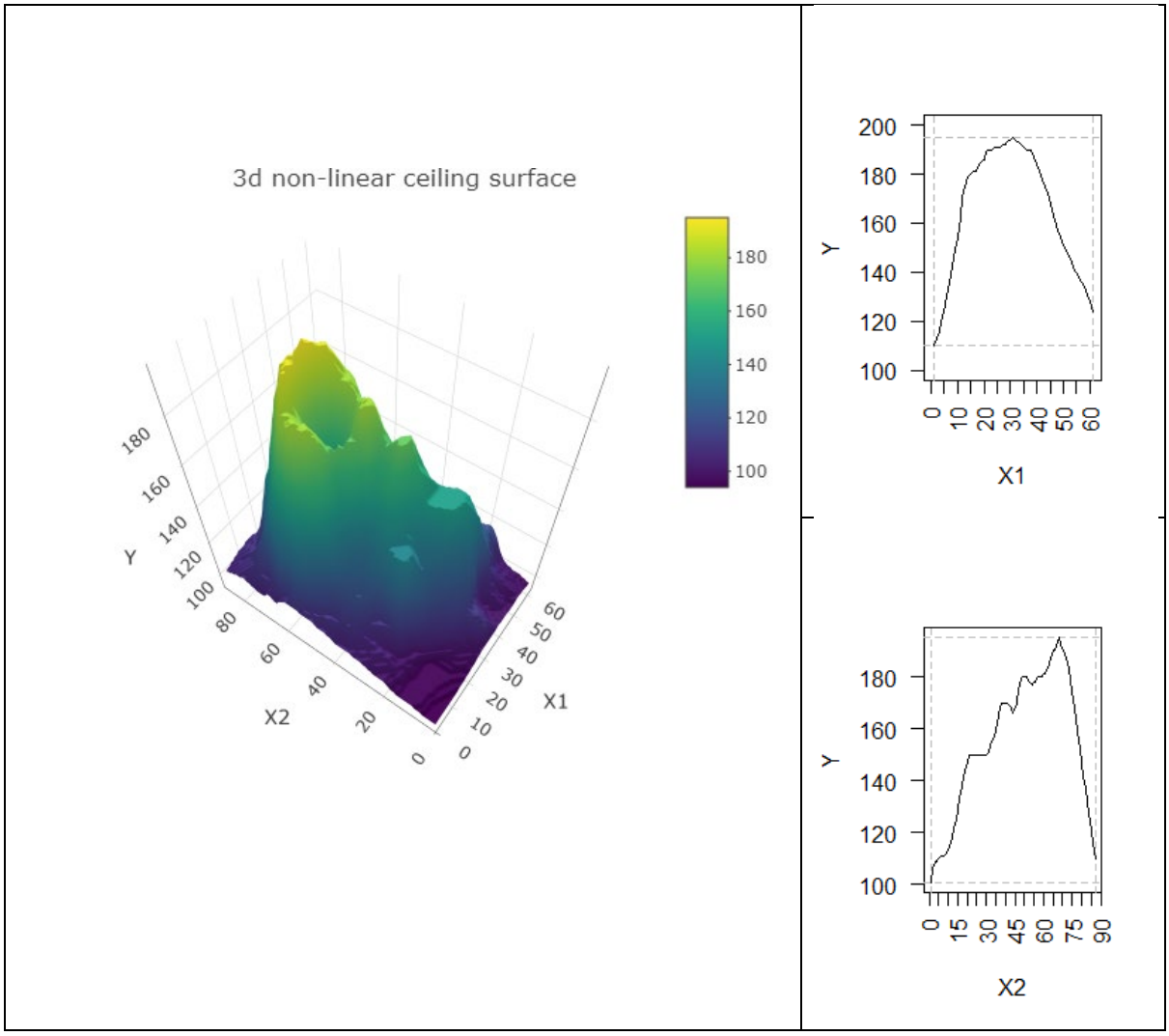

Figure 2. A nonlinear three-dimensional ceiling (mountain surface) and its projection ("shadow") on two two-dimensional XY planes (non-linear ceiling lines).

Why NCA analyses ceiling lines and not multidimensional ceilings?

NCA is not designed as a multivariate technique for analysing the multidimensional ceiling but as a method of analysing the bivariate ceiling lines. This has several related reasons.

The first reason is the fundamental choice to focus on *single factors* that are necessary for an outcome. This makes NCA different from conventional analyses. Conventional approaches such as regression analysis study multi-causal phenomena by considering a combination of factors and their interplay *as a whole* (e.g., a regression equation with several factors). A multivariate analysis is the realistic option for predicting the *presence* of an outcome, because

normally only a combination of factors and not a single factor can produce the outcome. However, for predicting the *absence* of an outcome is it realistic to study a single factor (the necessary condition or 'bottleneck') that can prevent the outcome to exist when the level is too low. Therefore, it is possible and useful to study the necessity of single factors for an outcome.

The second reason that in NCA makes statements about single factors that are necessary *independently* of other factors. Thus, the necessity of a single factor does not depend on the level of other factors. This allows for a "pure" and straightforward interpretation of necessity: "the factor is necessary" rather than "the factor is necessary depending on other factors", thus not always necessary. Such generic necessity statements hold in a broad variety of contexts, independently of other factors. However, the context where the necessity statement holds is usually not unlimited. The domain where the generic necessary condition is supposed to hold must be defined as the 'theoretical domain' of the necessity theory (sometime called "scope conditions"). Such specification of the theoretical domain must be part of any theory and related hypotheses (for further discussion and examples, see the book).

A third reason is that NCA wants to contribute to *parsimonious (simple) theories*: avoiding that theories become complex, not understandable and (thus) less useful. This is a general goal of theory building in applied sciences. NCA is an elegant way of reducing complexity, in particular in situations where it is hard or impossible to predict the outcome (e.g., when the explained variance of regression models is low) or when a good prediction is only possible with a very complex model.

The fourth reason is that *practical recommendations* from identified single necessary conditions are immediately clear and useful: "always satisfy all necessary conditions" otherwise there is guaranteed failure of the outcome. The absence of a necessary condition cannot be compensated by other factors.

The final reason is that the search for single necessary conditions is more *efficient* with a bivariate analysis (ceiling line) than with a multivariate analysis (multivariate ceiling). In a multivariate analysis the multivariate ceiling line must be estimated first, followed by taking the projections of the ceiling on the XY planes. Modeling and estimating a multivariate ceiling may be complex (e.g., Figure 2). [Estimating a multivariate ceiling is done in 'frontier analysis': predicting the maximum outcome for a given combination of factors, which is used for example in benchmarking applications to describe how far a case is distant from the maximum possible outcome]. NCA has a different goal than estimating the maximum outcome for different values of the condition. NCA estimates the maximum possible outcome for a single condition and therefore can focus directly on the XY planes and the ceiling line. [NCA uses techniques from frontier analysis, for example the Free Disposal Hull for the CE-FDH and CR-FHD ceiling lines, applies them in the two-dimensional plane, interprets the results in terms of necessity, and focuses on the empty space above the ceiling (prediction of absence of the outcome), rather than the space with cases below the ceiling].

In conclusion: to avoid misinterpretation of the term ‘multivariate NCA’ I suggest to use ‘multiple NCA’.

References

Dul, J. (2016) Necessary Condition Analysis (NCA): Logic and methodology of “necessary but not sufficient” causality. Organizational Research Methods, 19(1), 10-52.

Everitt, B.S. and Skrondal, A. (2010). The Cambridge Dictionary of Statistics. 4th edition. Cambridge University Press. 480 pp.

Hidalgo B, Goodman M. (2013) Multivariate or multivariable regression? American Journal of Public Health, 103(1), 39–40.