NCA and regression

Supplement to Dul, J. (2020) Conducting Necessary Condition Analysis. Sage publications.
Jan Dul
Version 1.1.0; June 4, 2021

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

Introduction

Regression is the mother of all data analyses in the social sciences. It was invented more than 100 years ago when Francis Galton (1886) quantified the pattern in the scores of parental height and child height (see Figure 1 with the original graph).

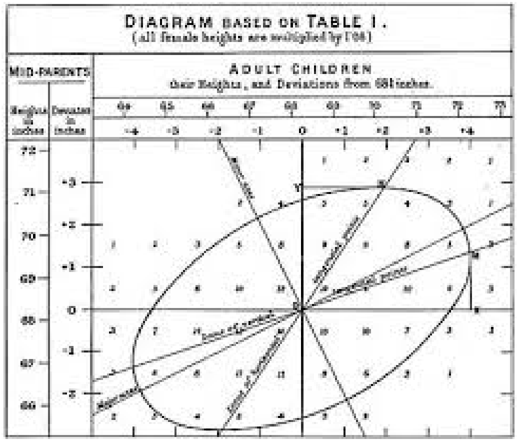

Figure 1. Francis Galton's (1886) graph with data on Parent height (X) and Child height (Y).

In Figure 2 Galton's data are shown in two XY scatter plots.

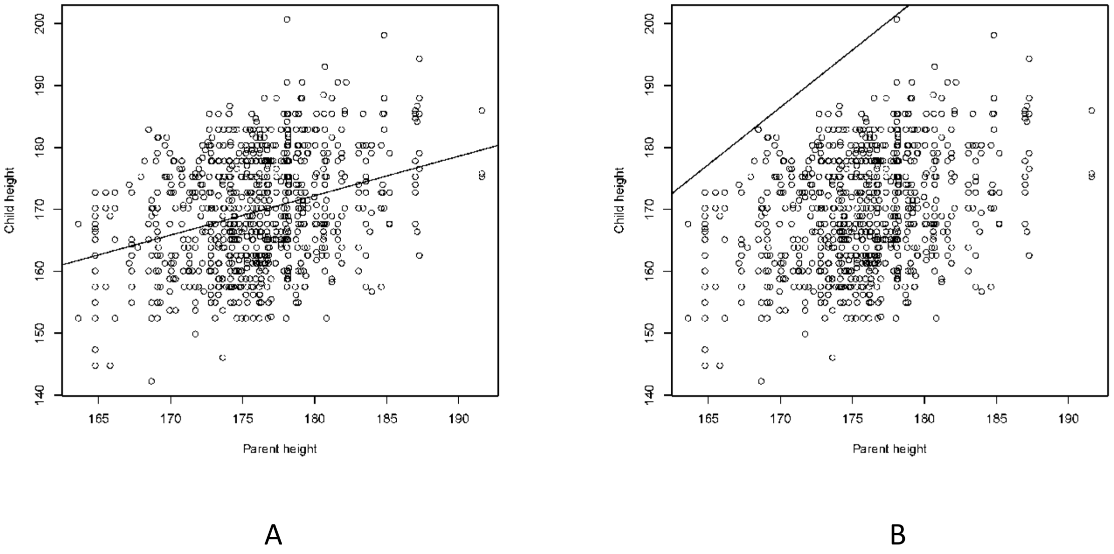

Figure 2 Scatter plots of the relationship between Parent height (X) and Child height (Y) (after Galton (1886). A. With a regression line. B. With a ceiling line.

Galton drew lines though the middle of the data for describing the average trend between Parental height and Child height: the regression line (Figure 2A). For example with a Parent height of 175 cm, the estimated average Child height is about 170 cm. Galton could also have drawn a line on top of the data for describing the necessity of Parent height for Child height: the ceiling line (Figure 2B). For example, with a Parent height of 175 cm, the estimated maximum Child height is about 195 cm. But Galton didn't draw a ceiling line, and the social sciences have adopted the average trend line as the basis for many data analysis approaches. Regression analysis has developed over the years and many variants exist. The main variant is Ordinary Least Squares (OLS) regression with which I compare NCA.

Logic and theory

OLS regression uses additive, average effect logic. The regression line (Figure 2A) predicts the average Y for a given X. Because the cases are scattered, for a given X also higher and lower values of Y than the average value of Y are possible. With one X (simple OLS regression), Y is predicted by the regression equation is $Y = \beta_0 + \beta_1 X + \varepsilon(X)$, where β_0 is the intercept of the regression line, β_1 is the slope of the regression line, and $\varepsilon(X)$ is the error term representing the scatter around the regression line for a given X. The slope of the regression line (regression coefficient) is estimated by minimizing the squared vertical distances between the observed Y-values and the regression line ('least squares'). The error term includes the effect of all other factors that can contribute to the outcome Y. For the parent-child data, the regression equation is $Y = 57.5 + 0.64 X + \varepsilon(X)$. OLS regression assumes that on average $\varepsilon(X) = 0$. Thus, when X (Parent height) is 175 cm, the estimated *average* Child height is about 170. In contrast NCA's ceiling line is defined by $Yc = -129 + 1.85 X$. Thus, when X (Parent height) is 175 cm, the estimated *maximum* Child height is about 195 cm. Normally, in NCA the ceiling line is interpreted inversely (e.g., in the bottleneck table): Xc = (Yc + 129)/1.85 indicating, assuming a non-decreasing ceiling line, that a minimum level of X = Xc is necessary (but not sufficient) for a desired level of Y =Yc. When parents wish to have a child of 200 cm it is necessary (but not sufficient) that their Parent height is at least about 177 cm.

To allow for doing statistical tests with OLS, it is usually assumed that the error term for a given X is normally distributed (with average value 0): cases close to the regression line for the given X are more likely than cases far from the regression line. The normal distribution is unbounded, hence very high or very low values of Y are possible, though not likely. This implies that any high value of Y is possible. Even without the assumption of the normal distribution or the error term, a fundamental assumption of OLS is that the Y value is unbounded (Berry, 1993). Thus, very large child heights (e.g., 300 cm) are theoretically possible in OLS, but unlikely. This assumption contradicts NCA's logic in which X and Y are presumed bounded. X puts a limit on Y and thus there is a border represented by the ceiling line. The limits can be empirically observed in the sample (e.g., the height of the observed tallest person in the sample is 205 cm) for defining NCA's empirical scope or can be theoretically defined (e.g., the height of the ever observed tallest person is 272 cm) for defining the NCA's theoretical scope.

Additivity is another part of regression logic. It is assumed that the terms of the regression equation are added. Next to X, the error term is always added in the regression equation. Possibly also other X's or combination of X's are added in the regression equation (multiple regression, see below). This means that the terms that make up the equation can compensate for each other. For example, when X is low, Y can still be achieved when other factors (factors in the error term or other X's) give a higher contribution to Y. The additive logic implies that for achieving a certain level of Y, no X is necessary. This additive logic contradicts NCA's logic that X is necessary: Y cannot be achieved when the necessary factor does not have the right level, and this absence of X cannot be compensated by other factors.

Results of a regression analysis are usually interpreted in terms of sufficiency. A common sufficiency-type of hypotheses is 'X increases Y' or 'X has a positive effect on Y'. Such hypothesis can be tested with regression analysis. The hypothesis is considered to be supported if the regression coefficient is positive. Often, it is then suggested that X is sufficient to produce an increase of the outcome Y. The results also suggest that a given X is not necessary for producing the outcome Y because other factors in the regression model (other X's and the error term) can compensate for the absence of a low level of X.

Data analysis

Mostly, regression models include more than one X. The black box of the error term is opened and other X's are added to the regression equation, for example: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon(X)$, where β_1 and β_2 are the regression coefficients (multiple regression). By adding more factors that contribute to Y into the equation, a larger part of the scatter is explained, hence resulting in more precise prediction of Y for given X's. R^2 is the amount of explained variance of a regression model and can have values between 0 and 1. By adding more factors, better predictions of the outcome can be achieved and higher values of R^2.

Another reason to add more factors is that not including factors that correlate with X and Y results in biased estimations of the regression coefficients ('omitted variable bias'). Hence, the common standard of regression is not the simple OLS regression with one factor, but multiple regression with many factors. Also other regression-based approaches such as Structural Equation Modelling and Partial Least Squares include many factors. By adding more relevant factors, the prediction of Y becomes better and the risk of omitted variable bias is reduced. Adding factors in the equation is not just adding new factors (X). Some factors may be combined such as squaring a factor (X^2) to represent a non-linear effect of X on Y, or taking the product of two factors (X_1 * X_2) to represent the interaction between these factors. Such combinations of factors is added as a separate terms into the regression equation. Box 1 shows an example of the prediction of an average outcome with 25 terms of single and combined factors in the regression equation. The 25 terms in the model explain 27% of the variance ($R^2 = 0.27$). Thus, the error term (representing the not included factors) represent the other 73% percent (unexplained variance). Single terms predict only a small part of the outcome. For example, 'subsidiary initiative taking' (term 18) is responsible for 2% to the explained variance.

Box 1 Example of predicting an outcome with many additive factors in a regression model

Bouquet and Birkinshaw's (2008) study on multinational enterprises (MNE's) to predict how subsidiary companies gain attention from their headquarters (Y) is one of the most cited papers in the Academy of Management Journal. They use a multiple regression model with 25 terms (X's and combination of X's) and an 'error' term ε. With the regression model the average outcome (average attention) for a group of cases (or for the theoretical 'the average case') for given values of the terms can be estimated. The error term represents all unknown factors that have a positive or negative effect on the outcome but are not included in the model, assuming that the average effect of the error term is zero. β_0 is a constant and the other β_i's are the regression coefficients of the terms, indicating how strong the term is related to the outcome (when all other terms are constant). The regression model is:

$$\begin{aligned}
\text{Attention} = {} & \beta_0 + \beta_1 \mathit{subsidiary\ size} + \beta_2 \mathit{subsidiary\ age} + \beta_3\ \mathit{subsidiary\ age}^2 \\
& + \beta_4 \mathit{subsidiary\ autonomy} + \beta_5 \mathit{subsidiary\ performance} \\
& + \beta_6 \mathit{subsidiary\ performance}^2 + \beta_7\ \mathit{subsidiary\ functional\ scope} \\
& +\ \beta_8 \mathit{subsidiary\ market\ scope} + \beta_9\ \mathit{geographic\ are\ structure} \\
& + \beta_{10} \mathit{matrixstructure} + \beta_{11} \mathit{geographic\ scope} \\
& + \beta_{12}\ \mathit{headquarter\ AsiaPacific\ parentage} \\
& + \beta_{13} \mathit{headquarter\ NorthAmerican\ parentage} \\
& + \beta_{14} \mathit{headquarter\text{-}subsiduary\ cultural\ distance} \\
& + \beta_{15} \mathit{presence\ of\ MNEs\ in\ local\ market} + \beta_{16} \mathit{local\ market\ size} \\
& + \beta_{17} \mathit{subsiduary\ strength\ within\ MNE\ Network} \\
& + \beta_{18} \mathit{subsidiary\ initiative\ taking} + \beta_{19} \mathit{subsidiary\ profile\ building} \\
& + \beta_{20} \mathit{headquarter\text{-}subsiduary\ geographic\ distance} \\
& + \beta_{21} \mathit{initiative\ taking} * \mathit{geographic\ distance} + \beta_{22} \mathit{profile\ building} \\
& * \mathit{geographic\ distance} + \beta_{23} \mathit{subsidiary\ downstream\ competence} \\
& + \beta_{24} \mathit{initiative\ taking} * \mathit{downstream\ competence} \\
& + \beta_{25} \mathit{profile\ building} * \mathit{downstream\ competence}\ + \varepsilon
\end{aligned}$$

Source:
Bouquet, C., & Birkinshaw, J. (2008). Weight versus voice: How foreign subsidiaries gain attention from corporate headquarters. *Academy of Management journal*, 51(3), 577-601.

The example shows that adding more factors makes the model more complex and less understandable and therefore less useful in practice. The contrast with NCA is large. NCA can have a model with only one factor that perfectly explains the absence of a certain level of an outcome when the factor is not present at the right level for that outcome.

Whereas regression models must include factors that correlate with other factors and with the outcome to avoid biased estimation of the regression coefficient, NCA's effect size for a necessary factor is not influenced by the absence or presence of other factors in the model. This is illustrated with and example about the effect of a sales persons personality on sales performance using data of 108 cases (sales representatives from a large USA food manufacturer) obtained with The Hogan Personality Inventory (HPI) personality assessment tool for predicting organizational performance (Hogan & Hogan, 2007). Details of the example are in Dul et al. (in press). The statistical descriptives of the data (mean, standard deviation, correlation) are shown in Figure 3A. Ambition and Sociability are correlated with Y as well as each other. Hence, if one of them is omitted from the model the regression results may be biased.

	M	S.D.	Y	X_1	X_2	X_3
Y = Sales Performance	49.0	18.1				
X_1 = Ambition	57.7	27.0	0.24			
X_2 = Sociability	53.1	28.5	0.28	0.35		
X_3= Interpersonal sensitivity	54.0	33.5	0.06	0.19	0.03	
X_4= Learning approach	49.5	29.8	-0.08	0.23	0.23	-0.02

A. Descriptive statistics.

MODEL 1 with 4 factors	**Regression coefficient (B)** **$R^2 = 0.13$**	**NCA effect size (*d*)** **CR-FDH**
X_1 = Ambition	0.13 (p=0.059)	0.18 (p=0.047)
X_2 = Sociability	0.16 (p=0.012)	0.19 (p=0.003)
X_3 = Interpersonal sensitivity	0.01 (p=0.883)	0.11 (p=0.371)
X_4 = Learning approach	-0.11 (p=0.055)	0.14 (p=0.167)
MODEL 2 with 3 factors (Sociability omitted)	**Regression coefficient (B)** **$R^2 = 0.08$**	**NCA effect size (*d*)** **CR-FDH**
X_1 = Ambition	0.18 (p=0.006)	0.18 (p=0.047)
~~X_2 = Sociability~~	-	-
X_3 = Interpersonal sensitivity	0.00 (p=0.952)	0.11 (p=0.371)
X_4 = Learning approach	-0.09 (p=0.135)	0.14 (p=0.167)

(N=108)

B. Results of regression analysis (middle column) and results of NCA (right column). Model 1 is the full model and Model 2 has an omitted variable (Sociability).

Figure 3 Example of the results of a regression analysis and NCA. Effect of four personality traits of sales persons on sales performance. A. Descriptive statistics. B. Results of regression analysis (middle column) and results of NCA (right column) for two different models. Model 1 is the full model and Model 2 has an omitted variable (Sociability). Data from Hogan & Hogan (2007).

The omission of one variable is shown in Figure 3B, middle column. The full model (Model 1) includes all four personality factors. The regression results show that Ambition and Sociability have a positive average effects on Sales performance (regression coefficients 0.13 and 0.16 respectively, and Learning approach has a negative average effect on Sales performance (regression coefficient -0.11). Interpersonal sensitivity has virtually no average effect on Sales performance (regression coefficient 0.01). The p-values for Ambition and Sociability are relatively low. In the model with 3 factors Sociability is omitted (Model 2). The regression results show that all three remaining regression coefficients have changed. The regression coefficient for Ambition has increased to 0.18, and the regression coefficients of the other two factors have minor differences (because these factors are less correlated with the omitted variable). Hence, in a regression model that is not correctly specified because a factor that correlates with factors that are included in the model and with the outcome are not included, the correlation coefficients of the included factors may be biased (omitted variable bias).

The results of the NCA analysis does not change when a variable is omitted (Figure 3B, right column). All factors are necessary with an effect size of greater than 0.10. The results for the remaining three factors do not change when Sociability is excluded from the model. The example also shows that a factor that is important according to a regression analysis may also be necessary (Ambition, Sociability). But the example also show that a factor that is not important (Interpersonal sensitivity according to the average contribution to the outcome may still be necessary for the outcome. Even a factor this has a negative average effect on the outcome (Learning approach) may be necessary for the outcome.

Combining NCA and regression

This example illustrates that regression and NCA are fundamentally different and complementary. A regression analysis can be added to a NCA study to evaluate the average effect of the identified necessary condition on the outcome. However, the researcher must then include all relevant factors, also those that are not expected to be necessary, to avoid omitted variable bias, and must obtain measurement scores for these factors.

When NCA is added to a regression study not much extra effort is required. If a theoretical argument is available for a factor being necessary, any factor that is included in a regression model (independent variables, moderators, mediators) can also be treated as a potential necessary condition that can be tested with NCA. This could be systematically done:

- For all potential necessary conditions;
- For those factors that provide a surprising result in the regression analysis (e.g. in terms of direction of the regression coefficient) to better understand the result;
- For those factors that show no or a limited effect in the regression analysis (small regression coefficient) to check whether such ‘unimportant’ factors *on average* still may be necessary for a certain outcome;
- For those factors that have a large effect in the regression analysis (large regression coefficient) to check whether an ‘important’ factor *on average* may also be necessary or not.

When adding NCA to a regression analysis more insight about the effect of X on Y can be obtained.

What is the same in NCA and regression?

I showed that regression has several characteristics that are fundamentally different from the characteristics of NCA. Regression is about average trends, uses additive logic, assumes unbounded Y values, is prone to omitted variable bias, needs control variables, and is used for testing sufficiency-type of hypotheses, whereas NCA is about necessity logic, assumes limited X and Y, is immune for omitted variable bias, does not need control variables, and is used for testing necessity hypotheses. However, NCA and regression also share several characteristics. Both NCA and regression are variance-based approaches and use linear algebra (although NCA can also be applied with the set theory approach with Boolean algebra; see the supplement on NCA and QCA). Both methods need good (reliable and valid) data without measurement error, although NCA may be more prone to measurement error. For statistical generalization from sample to population both methods need to have a probability sample that is representative for the population, and having larger samples usually give more reliable estimations of the population parameters, although NCA can handle small sample sizes. Additionally, for generalization of the findings of a study both methods need replications with different samples; a one-shot study is not conclusive. Both methods cannot make strong causal interpretations when observational data are used; then at least also theoretical support is needed. When null hypothesis testing is used in both methods, such tests and the corresponding p-values have strong limitations and are prone to misinterpretations; a low p value only indicates a potential randomness of the data and is not a prove of the specific alternative hypothesis of interest (average effect, or necessity effect).

When a researcher uses NCA or OLS, these common fundamental limitations should be acknowledged. When NCA and OLS are used in combination the fundamental differences between the methods should be acknowledged. It is important to stress that one method is not better than the other. NCA and OLS are different and address different research questions. To ensure theory-method fit, OLS is the preferred method when the researcher is interested in an average effect of X on Y, and NCA is the preferred method when the researcher is interested in the necessity effect of X on Y.

References

Berry, W. D. (1993). Understanding regression assumptions (Vol. 92). Sage Publications

Dul, J., Hauff, S., Tóth, Z. (in press). *Necessary condition analysis in marketing research*. In: Handbook of Research Methods for Marketing Management. Eds. Robin Nunkoo, Viraiyan Teeroovengadum, and Christian Ringle. Edgard Elgar.

Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory manual* (3rd ed.). Tulsa, OK: Hogan Assessment.

Appendix. Script for obtaining and analysing the Galton dataset.

```
library(HistData) # R package that contains the Galton dataset
library(NCA) # R package for conducting NCA
data("GaltonFamilies") # get the Galton data from the HistData
package
head(GaltonFamilies) # print the head of the data file
Parent.Height<-GaltonFamilies[,4]*2.54 # parent height in cm
Child.Height<-GaltonFamilies[,8]*2.54 # child height in cm
data<-as.data.frame (cbind(Parent.Height,Child.Height)) # make a
data frame
#pdf("Galton OLS.pdf") # delete '#' for storing a pdf file of the
scatter plot
plot(data, xlab = "Parent height", ylab = "Child height") # make
scatter plot
modelOLS<- lm(Child.Height~Parent.Height, data) # perform OLS
regression analysis
slopeRegression<-modelOLS$coefficients[2] # slope of the OLS
regression line
interceptRegression<-modelOLS$coefficients[1] # intercept of the
OLS regression line
abline(modelOLS) # draw OLS regression line
#dev.off() #delete '#' for storing a pdf file of the scatter plot
#pdf("Galton NCA.pdf") #delete '#' for storing a pdf file of the
scatter plot
plot(data, xlab = "Parent height", ylab = "Child height")
modelNCA<- nca_analysis(data, 'Parent.Height', 'Child.Height',
ceilings = "c_lp") # perform NCA
modelNCA
slopeCeiling<-modelNCA$summaries$Parent.Height$params[9] # slope
of the C-LP ceiling line
interceptCeiling<-modelNCA$summaries$Parent.Height$params[10] #
intercept of the C-LP ceiling line
abline(interceptCeiling,slopeCeiling) # draw the C-LP ceiling line
#dev.off() #delete '#' for storing a pdf file of the scatterplot
```