Necessary Condition Analysis (NCA) with Stata:

A Quick Start Guide

Version 0.0.2

December 11, 2023

Necessary Condition Analysis (NCA) is an approach and data analysis technique for identifying necessary conditions in datasets. It can complement traditional regression-based data analysis as well as other methods (see then NCA website www.erim.nl/nca for more information on NCA). This guide helps a Stata user to install and use the user written Stata package **nca**. It allows performing an NCA analysis with the example dataset within 15 minutes. Details of the method can be found in:

Dul, J. (2016) Necessary Condition Analysis (NCA). Logic and methodology of 'Necessary but not Sufficient' causality. *Organizational Research Methods* 19(1), 10- 52. (https://journals.sagepub.com/doi/pdf/10.1177/1094428115584005).

Suggested citation to the software:

Spinelli, D., Dul J. & Buijs, G. (2023). *nca: Stata module to perform Necessary Condition Analysis (NCA)*, Statistical Software Components S459269, Boston College Department of Economics.

1. Installation

Open the Stata package and enter **ssc install nca** in the command window for installing the package. Enter **net get nca** to download the data and the examples do-file in the working directory (enter **adopath** to identify the working directory).

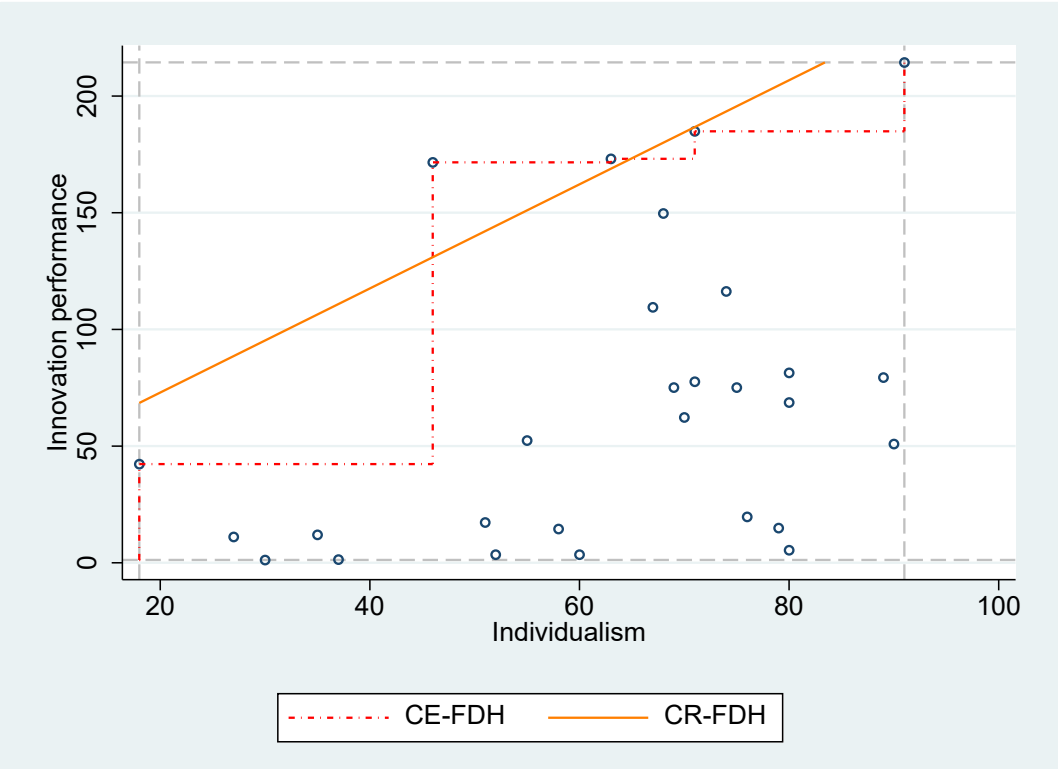
Enter **net get nca** to load the example data into a Stata session.

Enter **help nca** to get help about the **nca** functions.

2. NCA with one condition

In the example dataset, the dependent variable ('outcome' in NCA terminology) is innovationperformance (label: Innovation performance) and the two dependent variables (called 'conditions') are individualism (labeled Individualism) and risktaking (labeled Risk taking).

For evaluating the necessity of Individualism for Innovation performance you can run NCA by typing the command **nca individualism innovationperformance**. Note that in **nca** first the conditions are specified and last the outcome. After running this command two outputs appear.



```
NCA Parameters : individualism - innovationperformance

   Number of observations         28.0
                   Scope      15563.6
                    Xmin         18.0
                    Xmax         91.0
                    Ymin          1.2
                    Ymax        214.4

                            individ~m
                               ce_fdh      cr_fdh

             Ceiling zone      6466.8    4772.541
              Effect size     .415508    .3066477
                  # above           0           2
          c-accuracy (%)         100    92.85714
                  Fit(%)         100    73.80067

                            individ~m
                               ce_fdh      cr_fdh

                    Slope           .     2.23023
                Intercept           .    28.35271
             Abs. ineff.      3000.3    6018.517
        Rel. ineff. (%)    19.27767    38.67047
   Condition ineff. (%)           0    10.38263
     Outcome ineff. (%)    19.27767    31.56513
```

The first output (top) is a scatter plot showing the condition (X) on the horizontal axis and the outcome (Y) on the vertical axis. The two default ceiling lines CE-FDH (step function) and CR-FDH (straight line) are shown as well. A ceiling line is a line on top of the data (rather than a regression line through the middle of the data). It is a border line between the area with cases and the area without cases ('empty space'). In the example the upper left corner is expected to be empty given the necessary condition hypothesis that the presence or high level of X is necessary for the presence or high level of Y. No high Y is possible if X is low.

The second output is displayed in Stata's Results window. The output starts with basic information on the data (Number of observations, Scope, Xmin, Xmax, Ymin, and Ymax). Scope refers to the area of the possible combinations of and X and Y, given the minimum and maximum X and Y values. By default, the observed minimum and maximum values are selected, but it is also possible to select a theoretical scope using the **scope** option. The remaining lines present the NCA parameters for each of the selected ceiling techniques.

The 11 displayed NCA parameters are:

- Ceiling zone, which is the size of the 'empty' area above the ceiling line and within the scope in the selected expected empty corner (by default upper-left corner),

- Effect size, which is the ceiling zone divided by the scope,
- # above, which is the number of observations that are above the ceiling line, and hence in the ceiling zone,
- c-accuracy, which is the number of observations on or below the ceiling line divided by the total number of observations and multiplied by 100%,
- Fit, which relates to the closeness of the selected ceiling line to the CE-FDH ceiling line,
- Slope and Intercept, which are the slope and the intercept of the straight ceiling line (no values are displayed if the ceiling line is not a straight line),
- Abs. ineff., which is the absolute inefficiency: the area where X does not constrain Y, and Y is not constrained by X,
- Rel. ineff., which is the relative inefficiency: the area where X does not constrain Y, and Y is not constrained by X as percentage of the scope,
- Condition ineff., which is the condition inefficiency that indicates for which range of X (as a percentage of the total range) X does not constrain Y (i.e., there is no ceiling line in that X-range),
- Outcome ineff., which is the outcome efficiency that indicates for which range of Y (as a percentage of the total range of Y) Y is not constrained by X (i.e., there is no ceiling line in that Y-range).

3. NCA's statistical test

You can perform NCA's statistical test (a permutation test) to estimate a p value, by using the **testrep(#)** option where # is the number of resamples (e.g., 10000). The command is: **nca individualism innovationperformance, testrep(10000)**. The computations may take some time. With this option the estimated p value and its accuracy (p accuracy) are also displayed in the output. You may want to use a random number seed before using the statistical test to ensure replicability.

```
. set seed 1234567

. nca individualism innovationperformance, testrep(10000)
executing permutations


NCA Parameters : individualism - innovationperformance
```

Number of observations	28.0
Scope	15563.6
Xmin	18.0
Xmax	91.0
Ymin	1.2
Ymax	214.4

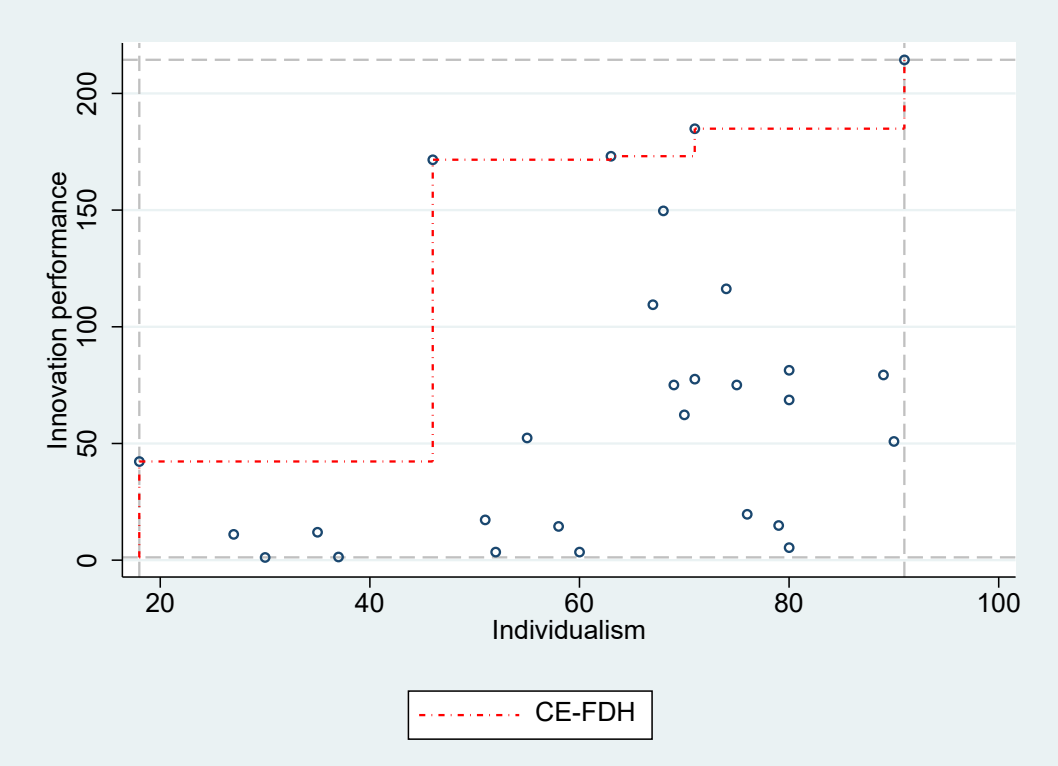
	individ~m ce_fdh	cr_fdh
Ceiling zone	6466.8	4772.541
Effect size	.415508	.3066477
# above	0	2
c-accuracy (%)	100	92.85714
Fit(%)	100	73.80067
p-value	.0828	.1749
p-accuracy	.0054014	.0074457

	individ~m ce_fdh	cr_fdh
Slope	.	2.23023
Intercept	.	28.35271
Abs. ineff.	3000.3	6018.517
Rel. ineff. (%)	19.27767	38.67047
Condition ineff. (%)	0	10.38263
Outcome ineff. (%)	19.27767	31.56513

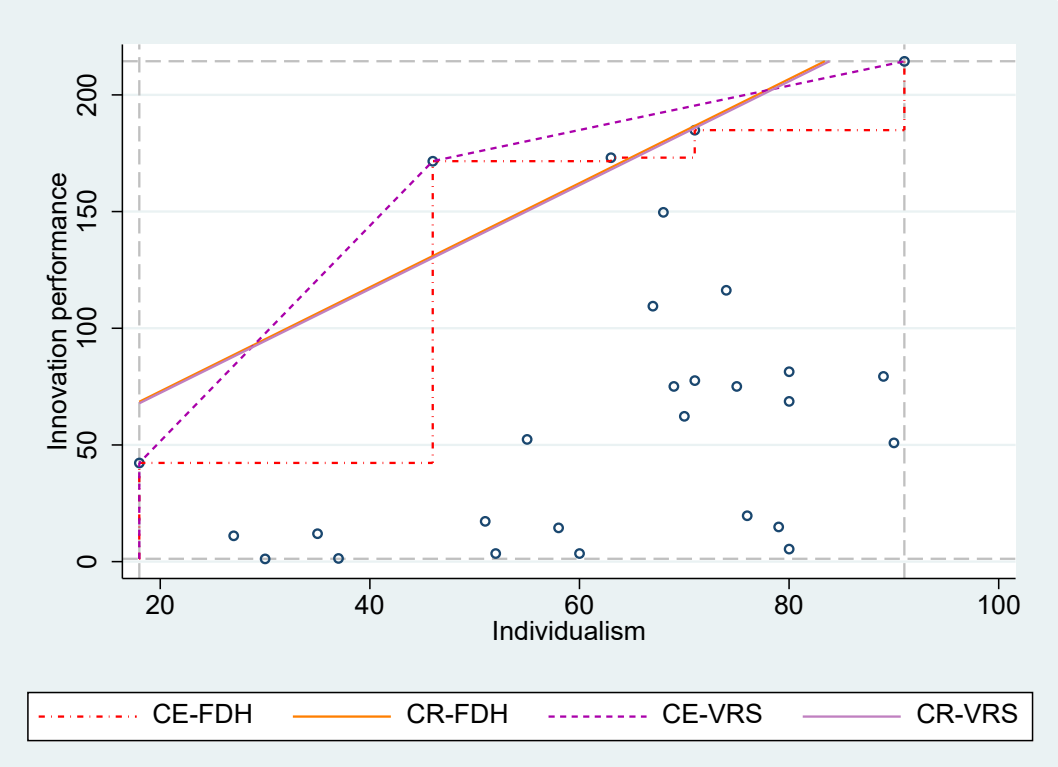
```
graphs individualism have been saved, check  graph dir
```

4. Selection of ceiling lines

Ceiling lines can be selected by using the **ceilings** options. The CE-FDH line is often used for discrete data or for a non-linear border, and the CR-FDH for continuous data and a linear border. . For example for selecting the CE-FDH ceiling line the command is: **nca individualism innovationperformance , ceilings(ce_fdh)**

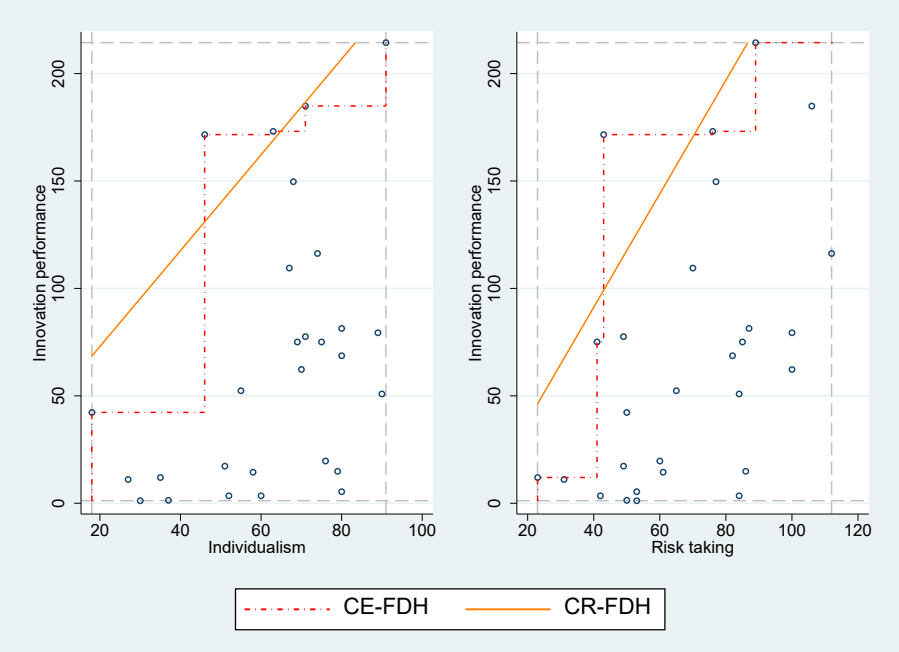


Other ceiling lines can be selected as well. All currently available lines are shown using **nca individualism innovationperformance , ceilings(ce_fdh cr_fdh ce_vrs cr_vrs)**



5. NCA with multiple conditions

Several conditions can be analyzed at once using the command **nca individualism risktaking innovationperformance.** Now the output consists of two graph and two displayed tables. See **graph dir** command to list the saved plots.



```
. nca individualism risktaking innovationperformance

NCA Parameters : individualism - innovationperformance
```

Number of observations	28.0
Scope	15563.6
Xmin	18.0
Xmax	91.0
Ymin	1.2
Ymax	214.4

	individ~m ce_fdh	cr_fdh
Ceiling zone	6466.8	4772.541
Effect size	.415508	.3066477
# above	0	2
c-accuracy (%)	100	92.85714
Fit(%)	100	73.80067

	individ~m ce_fdh	cr_fdh
Slope	.	2.23023
Intercept	.	28.35271
Abs. ineff.	3000.3	6018.517
Rel. ineff. (%)	19.27767	38.67047
Condition ineff. (%)	0	10.38263
Outcome ineff. (%)	19.27767	31.56513

```
NCA Parameters : risktaking - innovationperformance

  Number of observations |      28.0
                   Scope |   18974.8
                    Xmin |      23.0
                    Xmax |     112.0
                    Ymin |       1.2
                    Ymax |     214.4

                         | risktak~g
                         |   ce_fdh     cr_fdh
-------------------------+----------------------
            Ceiling zone |    5871.1   5348.968
             Effect size |   .3094156   .2818985
                 # above |          0          1
         c-accuracy (%)  |        100   96.42857
                 Fit(%)  |        100   91.10675

                         | risktak~g
                         |   ce_fdh     cr_fdh
-------------------------+----------------------
                   Slope |          .   2.641869
               Intercept |          .   -14.4777
            Abs. ineff.  |     5616.4   8276.863
       Rel. ineff. (%)   |   29.59926   43.62029
    Condition ineff. (%) |    25.8427   28.50028
     Outcome ineff. (%)  |   5.065666   21.14695

graphs individualism risktaking have been saved, check  graph dir
```

6. Displaying the bottleneck table with default values

NCA's bottleneck table is a tabular representation of the ceiling line. It is particularly useful when there are multiple conditions. The default way to show a bottleneck table is to add the **bottlenecks** option in the **nca** command: **nca individualism risktaking innovationperformance, nograph nosummaries bottlenecks.**

The output is suppressed (options **nograph** and **nosummaries**) as the plots and the summaries are the same as above. The default setting of the bottleneck table is to display the X and the Y as percentage of their observed range in steps of 10%. There is one bottleneck table for each ceiling line.

Bottlenecks: - innovationperformance (ce-fdh) (Y=perc_range, X=perc_range, cutoff=0)

innovationperformance	individualism	risktaking
0	NN	NN
10	NN	20.22472
20	38.35616	20.22472
30	38.35616	20.22472
40	38.35616	22.47191
50	38.35616	22.47191
60	38.35616	22.47191
70	38.35616	22.47191
80	61.64384	59.55056
90	100	74.1573
100	100	74.1573

Bottlenecks: - innovationperformance (cr-fdh) (Y=perc_range, X=perc_range, cutoff=0)

innovationperformance	individualism	risktaking
0	NN	NN
10	NN	NN
20	NN	NN
30	NN	8.027471
40	11.0457	17.09493
50	24.14098	26.1624
60	37.23626	35.22986
70	50.33153	44.29733
80	63.42681	53.36479
90	76.52209	62.43225
100	89.61736	71.49972

NN: not necessary, NA: not available

7. Displaying the bottleneck tables with custom values

The **bottlenecks** options also allow a *numlist* to control the values of Y to be displayed in the bottleneck table. In this example, the bottleneck table is to display the X and the Y as percentage of their observed range in steps of 5%. The command is **nca individualism risktaking innovationperformance, nograph nosummaries bottlenecks(0(5)100)**. There are other options to customize the bottleneck table; please check options **steps**, **stepsize** and **bottlenecks** in **help nca**.

```
. nca individualism risktaking innovationperformance, nograph nosummaries bottlenecks
 Bottlenecks:  - innovationperformance (ce-fdh)   (Y=perc_range, X=perc_range, cutoff=0)
```

innovationperformance	individualism	risktaking
0	NN	NN
10	NN	20.22472
20	38.35616	20.22472
30	38.35616	20.22472
40	38.35616	22.47191
50	38.35616	22.47191
60	38.35616	22.47191
70	38.35616	22.47191
80	61.64384	59.55056
90	100	74.1573
100	100	74.1573

```
 Bottlenecks:  - innovationperformance (cr-fdh)   (Y=perc_range, X=perc_range, cutoff=0)
```

innovationperformance	individualism	risktaking
0	NN	NN
10	NN	NN
20	NN	NN
30	NN	8.027471
40	11.0457	17.09493
50	24.14098	26.1624
60	37.23626	35.22986
70	50.33153	44.29733
80	63.42681	53.36479
90	76.52209	62.43225
100	89.61736	71.49972

```
NN: not necessary, NA: not available
```

8. Displaying the bottleneck tables with actual values

The X and Y values in the bottleneck table can be displayed with different types of values using the **xbottlenecks** and **ybottlenecks** options. For selecting the original ('actual') values the command is **nca individualism risktaking innovationperformance, nograph nosummaries xbottlenecks(actual) ybottlenecks(actual) bottlenecks**

```
. nca individualism risktaking innovationperformance, nograph nosummaries xbottlenecks(actual) ybottlenecks(actual) bottlenecks
 Bottlenecks:  - innovationperformance (ce-fdh)   (Y=actual, X=actual, cutoff=0)
```

innovationperformance	individualism	risktaking
1.2	NN	NN
22.52	NN	41
43.84	46	41
65.16	46	41
86.48	46	43
107.8	46	43
129.12	46	43
150.44	46	43
171.76	63	76
193.08	91	89
214.4	91	89

```
Bottlenecks:  - innovationperformance (cr-fdh)   (Y=actual, X=actual, cutoff=0)
```

innovationperformance	individualism	risktaking
1.2	NN	NN
22.52	NN	NN
43.84	NN	NN
65.16	NN	30.14445
86.48	26.06336	38.21449
107.8	35.62291	46.28453
129.12	45.18247	54.35458
150.44	54.74202	62.42462
171.76	64.30157	70.49467
193.08	73.86112	78.5647
214.4	83.42068	86.63475

```
NN: not necessary, NA: not available
```

The Y values in the bottleneck table can be customized also when using actual values. The following outputs reports the bottleneck table for Y={50, 75, 100,180} by using the command line **nca individualism risktaking innovationperformance, nograph nosummaries xbottlenecks(actual) ybottlenecks(actual) bottlenecks(50(25)100 180)**

```
. nca individualism risktaking innovationperformance, nograph nosummaries xbottlenecks(actual) ybott
> lenecks(actual) bottlenecks(50(25)100 180)
 Bottlenecks:  - innovationperformance (ce-fdh)   (Y=actual, X=actual, cutoff=0)
```

innovationperformance	individualism	risktaking
50	46	41
75	46	41
100	46	43
180	71	89

```
Bottlenecks:  - innovationperformance (cr-fdh)   (Y=actual, X=actual, cutoff=0)
```

innovationperformance	individualism	risktaking
50	NN	24.40609
75	20.91591	33.86908
100	32.12552	43.33208
180	67.99626	73.61367

9. Displaying the bottleneck tables with percentile and actual values

When the X values in the bottleneck table are displayed as percentiles and the Y value as actual value, the interpretation of the bottleneck table is that for a given row, the X value corresponds to the percentage of cases that was unable to achieve the required X level for the desired Y level of that row. For selecting the actual values for Y and the percentiles for X the command is **nca individualism risktaking innovationperformance, nograph nosummaries xbottlenecks(percentile) ybottlenecks(actual) bottlenecks**

```
. nca individualism risktaking innovationperformance, nograph nosummaries ///
>         xbottlenecks(percentile) ybottlenecks(actual) bottlenecks
 Bottlenecks:  - innovationperformance (ce-fdh)   (Y=actual, X=percentile, cutoff=0)
```

innovationperformance	individualism	risktaking
1.2	0 (0)	0 (0)
22.52	0 (0)	7.142857 (2)
43.84	17.85714 (5)	7.142857 (2)
65.16	17.85714 (5)	7.142857 (2)
86.48	17.85714 (5)	14.28571 (4)
107.8	17.85714 (5)	14.28571 (4)
129.12	17.85714 (5)	14.28571 (4)
150.44	17.85714 (5)	14.28571 (4)
171.76	39.28571 (11)	53.57143 (15)
193.08	96.42857 (27)	82.14286 (23)
214.4	96.42857 (27)	82.14286 (23)

```
Bottlenecks:  - innovationperformance (cr-fdh)   (Y=actual, X=percentile, cutoff=0)
```

innovationperformance	individualism	risktaking
1.2	0 (0)	0 (0)
22.52	0 (0)	0 (0)
43.84	0 (0)	0 (0)
65.16	0 (0)	3.571429 (1)
86.48	3.571429 (1)	7.142857 (2)
107.8	14.28571 (4)	17.85714 (5)
129.12	17.85714 (5)	39.28571 (11)
150.44	28.57143 (8)	46.42857 (13)
171.76	42.85714 (12)	53.57143 (15)
193.08	64.28571 (18)	60.71429 (17)
214.4	89.28571 (25)	78.57143 (22)

```
NN: not necessary, NA: not available
Number of cases that are unable the reach the required x level for the desired y level in parentheses
```