Data Cleaning Tutorial: adjusting numerical data

Mark van der Loo

https://github.com/markvanderloo/UFPEL2019

Try the code

03valid/adjusting.R

Adjusting numerical values

Minimally adjust values so that they conform to rules after imputation.

Imputation

- Most imputation methods do not take the data restrictions/rules into account.
- This means that valid data can be become invalid after missing values have been imputed.

Successive projection algorithm

Idea

Alter (imputed) values in a record $\boldsymbol{x}$ *as little as possible* to satisfy all restrictions.

As little as possible?

The minimal Eucledian distance between the original $\boldsymbol{x}$ and the adjusted record $\boldsymbol{x}^*$.

$$\boldsymbol{x}^* = \min_{\boldsymbol{x}}(\boldsymbol{x}^* - \boldsymbol{x})'(\boldsymbol{x}^* - \boldsymbol{x})$$

Successive Projection Algorithm (sketch)

Project $\boldsymbol{x}$ on each (in)equality restriction sequentially and iteratively until convergence.

Hildredth (1957) *Naval Research Logistics* **4** 79–85

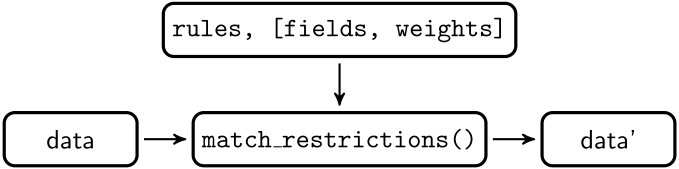
Extension: weighted distance

$$\boldsymbol{x}^* = \min_{\boldsymbol{x}}(\boldsymbol{x}^* - \boldsymbol{x})'\boldsymbol{W}(\boldsymbol{x}^* - \boldsymbol{x})$$

Property

If $W_{ij} = \delta_{ij}x_j^{-1}$, then the ratios between altered variables are preserved to $\mathcal{O}(1)$.

Pannekoek & Zhang (2015) *Survey Methodology* **41** 127–144; SDCR §10.11



rules, [fields, weights]

data

match_restrictions()

data'

Assignments

- load "03valid/errors_located.csv" into `errors_located`
- load "03valid/imputed.csv" into `imputed`.
- use `confront` to find out how many values are invalid in `imputed` and make a plot of the object
- Use `is.na` to store all `NA` values of `errors_located` into `adjust`
- apply `rspa::match_restrictions` to the data and use the `adjust` argument: we are restricting adjustments to the data that are imputed.
- use `confront` to find out how many values are invalid and make a plot of the object