Validatetools

Validatetools: Check and resolve contradictory rule sets

Edwin de Jonge

Statistics Netherlands / eRum 2018


CAUTION: BAD DATA
BAD DATA QUALITY
MAY RESULT IN
FRUSTRATION AND
LEAD TO DROP
KICKING YOUR
COMPUTER

Data cleaning...

A large part of your job is spent in data-cleaning:

- getting your data in the right shape (e.g. `tidyverse`)
- assessing missing data (e.g. `VIM`)
- checking validity (e.g. `validate`)
- locating and removing errors: `errorlocate`!
- impute values for missing or erroneous data (e.g. `simputation`)


KEEP
CALM
AND
VALIDATE

Completeness

Validation rules?

Package `validate` allows to:

- formulate explicit data rule that data must conform to:

```
library(validate)
check_that( data.frame(age=160, driver_license=TRUE),
  age >= 0,
  age < 150,
  if (driver_license == TRUE) age >= 16
)
```


RULES

Rules

A lot of datacleaning packages are based on explicit rules that data must conform to.

- `validate` to check **validity** of data
- `errorlocate` to find **errors**.
- `rspa`, `deductive`, `dcmodify` for **correction** and **imputation** using data rules.

Rules (2)

- Data rules are solidified domain knowledge.
- Real world knowledge e.g. :
  - age is not negative.
  - human age is less then 150 years.
- Expert knowledge, e.g:
  - if `profit > 0` then `turnover > 0`
  - if `married` then 'age $>$ 16'

Data checking

- A large part of data quality assurance in Official Statistics is checking data validity:
- n, number of records is high, typically > 0.5M
- p, number of columns is high, typically > 20
- population is diverse, different rules for different subpopulations.
- often many processing steps from input to statistic each checking/using (implicit) domain knowledge.

Result:

- Often many rules, great and small
- Rules often defined multiple times at different processing steps.
- Rules may partially contradict each other.

Validate tools

- Use `validate` to collect

Formally...

Rule set S

$$S = r_1 \wedge \cdots \wedge r_n$$

Rule r_i

$$r_i = \bigvee_j C_i^j$$

Atomic clause C_i^j

$$C_i^j(\boldsymbol{x}) = \begin{cases} \boldsymbol{a}^T \boldsymbol{x} \leq b \\ \boldsymbol{a}^T \boldsymbol{x} = b \\ x_j \in F_{ij} \text{with } F_{ij} \subseteq D_j \\ x_j \notin F_{ij} \text{with } F_{ij} \subseteq D_j \end{cases}$$