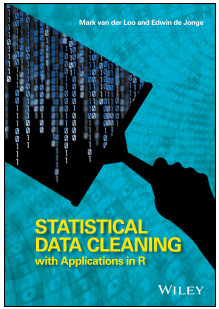

Error Localization

Mark van der Loo and Edwin de Jonge

Statistics Netherlands Research & Development
@markvdloo @edwindjonge

useR!2019

Try the code

03valid/errorlocalization.R

Error localization

Error localization is a procedure that points out fields in a data set that can be altered or imputed in such a way that all validation rules can be satisfied.

useR! 2019

Example

Ruleset

```
if (married == TRUE ) age >= 16
if (attends == "kindergarten") age <= 6
```

Data

age	married	attends
3	TRUE	kindergarten

Question

Which field or fields would you change?

Principle of Fellegi and Holt

Find the minimal (weighted) number of fields to adjust such that all rules, including implied rules, can be satisfied.

IP Fellegi and D Holt, JASA **71** 353 17–35 (1976).

Note

This should be used as a last resort, when no further information on the location of errors is available.

Implied rules?

```
turnover - total.cost == profit
          profit <= 0.6 * turnover
```

This implies (substituting profit):

```
          total.cost >= 0.4 * turnover
```

We need to take into account such *essentially new* rules: a rule set forms a system of rules and its implied rules. `errorlocate` takes this into account

Choosing weights

All weights equal (usually to one)

Least nr of variables adapted. In case of multiple solutions: choose randomly (e.g. by adding a small random perturbation to the weights).

Weights represent reliability

Heigher weight $\rightarrow$ variable is less likely chosen.

- Can be made to depend on 'outlierness', or expert judgement.
- Possible problem: minimal weights vs minimal nr of variables?

Choosing weights

Question

Is it possible to choose a set of weights, such that

a. The smallest number of variables is chosen
b. The weights are minimized

Intuition

If the weights do not differ too much, no extra variables will be introduced on top of the variables in a feasible solution of of minimal size.

errorlocate

`errorlocate` formulates a Mixed Integer Problem with:

- `validate` rules set R as a hard constraints
- values of record x_0 are soft constraints R_0
- objective function: minimize

$$f(x_0, \delta) = \sum_i w_i \delta_i$$

with $\delta_i \in \{0, 1\}$ and $\delta_i = 1$ if field i is an invalid value.

useR! 2019

Assignment

```
# we have much confidence in turnover, since these are
# also collected via the tax register
weight <- sapply(data_with_errors, function(x) 1)
# Set the weight of turnover to 10 and supply the weight to
# locate_errors
```

- Are less errors found in `turnover`?
- Replace errors with `NA` using the `replace_errors` with the weights used above
- Store the results in "my_errors_located.csv".