Towards generic analyses of data validation functions

Mark van der Loo
Jeroen Pannekoek
Data validation

Informally
Determine whether a (subset) of data satisfies or violates a presumption that is based on domain knowledge.

Examples

— Do variables revenue and cost add up to profit?
— Is the average revenue within 10% of last year's?
— Is the most common educational level in this branch 'high'?

Some characteristics

— univariate vs multivariate
— in-dataset vs cross-dataset
— in-record vs cross-record
Data validation

Observation: validation is a two-step process

1. Compute a `score`
2. Compare score with a valid region of scores

Examples

— Does revenue minus cost, minus profit equal 0?
— Is the average revenue within 10% of last year's?
— Is the most common educational level in this branch `high`?

Possible outcomes

— 1: Score is in valid region
— 0: Score not in valid region
— NA: Score can't be computed (because of missing values)
Data validation

Formally...

A validation function $v$ is the composition of two functions:

$$v = i_V \circ s,$$

where

$s$ : A score function
$V$ : Set of valid score functions
$i_V$ : Indicator function over $V$.

Explicitly

$$v(x) = i_V(s(x)) = \begin{cases} 
1 & \text{if } s(x) \in V \\
0 & \text{if } s(x) \notin V \\
\text{NA} & \text{if } s(x) = \text{NA}
\end{cases}$$

⇒ A validation function is fixed by determining $s$ and $V$. 
When \( x \) violates a presumption...

**Severity**

How much does the score differ from a valid score?

\[
R(x) = \inf\{d(s(x), s') : s' \in V\}
\]

\( \Rightarrow d \) is a distance function on the codomain of \( s \).

**Impact**

How much do I need to change \( x \) to obtain a valid score?

\[
I(x) = \inf\{\tilde{d}(x, x') : s(x') \in V\}
\]

\( \Rightarrow \tilde{d} \) is a distance function on the set of possible data.
When $x$ violates a presumption... (VALS)

**Severity**

*Indicates the significance of invalid records [...]*
(user-specified record-wise indicator)

**Discrepancy**

*The discrepancy between the validated data and reference data [...]*
(user-specified, rule-wise indicator)

*A. Simón(2013), Validation syntax: VALS version 0.1309*
Example

\[ f_{hb}(x) = \max \left( \frac{y}{y^*}, \frac{y^*}{y} \right) - 1 \]

\((y > 0)\)

rule \quad f_{hb}(y) < h

score \quad s(y) = f_{hb}(y)

valid \quad V = [0, h]

distance \quad d, \tilde{d} = |x - x'|
Example

- 840 × 80 Linear (in)equality checks
- Euclidean distance.
Example

- 840 × 80 Linear (in)equality checks
- Euclidean distance.
Example: impact function for all rules simultaneously

- 80 Linear (in)equality checks
- Euclidean distance.
Conclusions & outlook

Conclusions

— General approach to analysing validation results
— Applied to several types of data and rules

Outlook

— Derive $R$ and $I$ for other rule- and distance types
— General implementation