New results on automatic editing using hard and soft edit rules

Sander Scholtus

Statistics Netherlands

Introduction

- Error localisation problem:
  - Find erroneous and missing values in data
- Edit rules:
  - Constraints that should be satisfied by the data
  - Hard edit: *e.g.* Profit = Turnover – Costs
  - Soft edit: *e.g.* Profit / Turnover ≤ 50%
- Automatic error localisation

Automatic error localisation

- Fellegi and Holt (1976):
  - Find the smallest (weighted) possible subset of variables that can be imputed so that all edits are satisfied
  - Mathematically,

$$\min \sum_j w_j y_j \quad , \quad y_j \in \{0,1\}$$

so that the variables with $y_j = 1$ can be imputed consistently

  - Confidence weights w_j
- Many algorithms/tools available to solve this problem
- No room for soft edit rules

Automatic error localisation

Example: why soft edits matter

Hard edit rules:

- Profit = Turnover – Costs
- Turnover / Employees ≤ 550
- Turnover ≥ 0
- Costs ≥ 0
- Employees ≥ 0

Soft edit rules:

- Profit / Turnover ≤ 0.5
- Profit / Turnover ≥ –0.1

	Employees (3)	Turnover (2)	Costs (1)	Profit (1)
raw data	5	100	60 000	40 000
hard edits	5	100	60 000	–59 900
all edits (e.g.)	5	100	60	40

Automatic error localisation

- Incorporating soft edits:
  - Assign failure weights s_k to soft edits
  - Alternative error localisation problem

  $$\min\{\lambda \sum_j w_j y_j + (1-\lambda) \sum_k s_k z_k\}, \; y_j, z_k \in \{0,1\}$$

  so that the variables with $y_j = 1$ can be imputed consistently with all edits <u>except</u> for the soft edits with $z_k = 1$
- Algorithm proposed by Scholtus (2011, 2013)
- Small-scale simulation study: Scholtus and Göksen (2012)

Automatic error localisation

$$\min\{\lambda\sum_j w_j y_j + (1-\lambda)\sum_k s_k z_k\} \;,\; y_j, z_k \in \{0,1\}$$

- This can be rewritten as a Fellegi-Holt-type problem involving only hard edits
  - Add z_k as a variable with confidence weight $(1-\lambda)s_k$
  - Initial value: $z_k = 0$
  - Rewrite each soft edit as a conditional hard edit
- Examples:

  Profit ≤ 0.5×Turnover → IF (z_1 = 0) THEN (Profit ≤ 0.5×Turnover)

  IF (X > 0) THEN (Y > 0) → IF (X > 0 AND z_2 = 0) THEN (Y > 0)

Automatic error localisation

- Result: any software that can solve the error localisation problem of Fellegi and Holt (with conditional edits) can also solve the problem with hard and soft edits
- In particular, the problem can be solved using the R package **`editrules`** (available on CRAN)

Simulation study

– Real data from Dutch SBS 2007 on wholesale
  - Manually edited during regular production
  - One half used as test data, other half as reference data
  - Within test data: selected records with at most 10 errors
  - Two types of questionnaire: long form and short form

test data	short form	long form
number of records	126	800
number of variables	69	89
number of hard edit rules	93	111
number of soft edit rules	24	37

Simulation study

- Evaluation measures:
  - Fraction of false negatives: $\alpha = FN/(TP+FN)$
  - Fraction of false positives: $\beta = FP/(FP+TN)$
  - Fraction of wrong decisions: $\delta = (FN+FP)/tot$
  - Records with exactly the right solution: ρ
  - Missing values excluded in α, β, δ (detection trivial)
  - Here: focus on 22 'core variables'

		detected: error	detected: no error	total
true:	error	*TP*	*FN*	*TP+FN*
	no error	*FP*	*TN*	*FP+TN*
	total	*P*	*N*	*tot*

Simulation study: results

– Long form:

editing approach	α	β	δ	$1-\rho$	#errors
only hard edits	0.813	0.007	0.058	0.493	497
all edits as hard edits	0.650	0.060	0.097	0.694	2168
soft edits A	0.750	0.012	0.058	0.479	786
soft edits B	0.778	0.010	0.058	0.479	638
soft edits C	0.778	0.011	0.059	0.483	660
soft edits C, $\lambda=0.3$	0.731	0.011	0.056	0.459	889
soft edits, ideal subset	0.703	0.011	0.054	0.440	865

2404

– Short form:

editing approach	α	β	δ	$1-\rho$	#errors
only hard edits	0.779	0.012	0.085	0.500	125
all edits as hard edits	0.618	0.093	0.143	0.849	448
soft edits, ideal subset	0.673	0.018	0.080	0.452	187

455

Discussion

- Results of simulation study:
  - Quality of error localisation rather poor (missed errors)
  - Marginal improvement by incorporating soft edits
- Possible explanations:
  - Study considered relatively ‘difficult’ records
  - Large fraction of missing data
  - Suboptimal soft edit rules for automatic editing?
  - Minimisation criterion incompatible with manual editing?
- Other suggestions for improvement?

Thank you for your attention.

Normal form of edits

- Normal form of edits:

  IF (condition on categorical variables)

  THEN (linear condition on numerical variables)

- In practice, most edits can be (re-)written in normal form
  - Ratio edit:

    $X / Y \leq c$ → $X \leq c \times Y$
  - Conditional edit on numerical variables:

    IF $(X > 0)$ THEN $(Y > 0)$ → $(X \leq 0)$ OR $(Y > 0)$

    → IF (aux = “a”) THEN $(X \leq 0)$

    IF (aux = “b”) THEN $(Y > 0)$