Automatic error localisation

Automatic error localisation

Two possible approaches:

- the deterministic approach;
- the probabilistic approach.

Automatic error localisation: the deterministic approach

The deterministic approach is based on the definition of the deterministic edit rules, of the kind:

IF *incoherence* THEN *correction action*

The “incoherence” is expressed as a logical contradiction among two or more categorical variables, or as a mathematical constraint among two, or more, continuous variables

The “correction action” is an imputation action that is carried out on a pre-defined subset of variables appearing in the IF section.

Automatic error localisation: the deterministic approach

Example of deterministic error localisation in categorical data

if sex of spouse (not head) is equal sex of head, change the first one

```
IF Hsex=Ssex THEN  do;
     if Ssex=1 then Ssex=2;
     if Ssex=2 then Ssex=1;
End;
```

Automatic error localisation: the deterministic approach

Example of variables in a dataset

Record:

condition	(emploee/unemployed)
age	(classes of age)
position	(manager)
Educational attainment	(elementary/degree…)
unemployment subsidy	(yes/no)

Automatic error localisation: the deterministic approach

Example of deterministic error localisation in categorical data:

Problem… Let us define the set of deterministic rules:

Children can only start full-time work once they've reached at least 15 years of age.

Two rules are possible:

Rule1A: IF (age=0-14) AND (condition=employee)
THEN (age = 15-64)

Rule1B: IF (age=0-14) AND (condition=employee)
THEN (condition = not applicable)
Other variables can help in the decision (example: civil status)

Automatic error localisation: the deterministic approach

Example of deterministic error localisation in categorical data:

One can be manager only if he/she has a degree

Rule 2: IF (position=manager) AND NOT(condition=employee)
OR NOT(educational attainment=degree)
THEN (condition=employee) AND (educational attainment=degree)

Rule 3: IF (unemployment subsidy=yes) AND (condition=employee)
THEN (condition=unemployed)

Automatic error localisation: the deterministic approach

Example of deterministic error localisation in categorical data

Record:

condition	= unemployed
age	= 34
position	= manager
educational attainment	= elementary
unemployment subsidy	= yes

Automatic error localisation: the deterministic approach

Processing

STEP 1: Rule 1 is not activated

Rule 2 is activated and value “employed” is imputed to condition and “degree” to instruction level

Rule 3 is activated because of these new values and “unemployed” is imputed to condition

Record after rule2

condition	= employed
age	= 34
position	= manager
educational attainment	= degree
unemployment subsidy	= yes

Record after rule3

condition	= unemployed
age	= 34
position	= manager
educational attainment	= degree
unemployment subsidy	= yes

Automatic error localisation: the deterministic approach

Processing

STEP 2:

Rule 1 is not activated

Rule 2 is activated and value “employed” is imputed to condition

Rule 3 is activated and “unemployed” is imputed to condition

Record:

condition	= unemployed
age	= 34
position	= manager
educational attainment	= degree
unemployment subsidy	= yes

⇨ endless loop with the impossibility to eliminate the incoherence

Automatic error localisation: the probabilistic approach

Example of probabilistic error localisation in categorical data

Edit 1: IF (age=0,14) AND (condition=employed)

Edit 2: IF (age=0,14) AND (educational attainment=degree)

Edit 3: IF (position=manager) AND NOT(condition=employed)

Edit 4: IF (position=manager) AND NOT(educational attainment=degree)

Edit 5: IF (unemployment subsidy=yes) AND (condition=employed)

Automatic error localisation: the probabilistic approach

Example of probabilistic error localisation in categorical data

Record:

condition	= unemployed
age	= 34
position	= **manager**
educational attainment	= elementary
unemployment subsidy	= yes

Edit 1: IF (age=0,14) AND (condition=employed)	NO
Edit 2: IF (age=0,14) AND (educational attainment=degree)	NO
Edit 3: IF (position=manager) AND NOT(condition=employed)	YES
Edit 4: IF (position=manager) AND NOT(educational attainment=degree)	YES
Edit 5: IF (unemployment subsidy=yes) AND (condition=employed)	NO

Automatic error localisation: the Fellegi-Holt probabilistic approach

The Fellegi-Holt algorithm has been originally proposed for the localisation of random errors in categorical variables subject to logical constraints (*edits*).

The algorithm requires that the constraints to be satisfied by data are expressed through *explicit* rules (edit rules) having the general form:

IF [*error condition*]

Example: *IF Age<14 and Marital status=married*

Automatic error localisation: the Fellegi-Holt probabilistic approach

The set of explicit rules form an initial set of constraints used for identifying errors.

For each unit failing one or more of the explicit rules, the identification of the subset of variables having the higher probability of being in error is demanded to a probabilistic algorithm.

For each unit in error, on the basis of the failed rules, this algorithm determines the minimum number of items to be changed in order to let the unit pass all the edits without introducing new inconsistencies among data.

Automatic error localisation: the Fellegi-Holt probabilistic approach

The most important aspects are:

- **➔the minimum number of changes, thus the risk of introducing errors among observed data is minimised**
- **➔the localisation and correction of errors is performed taking into account simultaneously all edits**
- **➔errors are dealt within a similar manner on the basis of a theoretically valid approach that does not dependent on any personal decision**

Automatic error localisation: deterministic vs probabilistic approach

The probabilistic approach is based on the **minimum change criterion**.

In doing that, under the condition that **errors** in data are of the **random** type, the probabilistic approach is the best one in finding errors.

On the contrary, in case of **systematic errors**, the algorithm does not provide good results: in this case, determinstic rules are to be used.

Automatic error localisation: deterministic vs probabilistic approach

Example of systematic errors

In a questionnaire is given the following flow of questions:

1. POSITION IN THE PROFESSION

DEPENDENT WORKER	1
INDEPENDENT WORKER	2

(answer to the following question only if dependent worker)

2. DO YOU WORK IN PUBLIC OR IN PRIVATE SECTOR?

PUBLIC	1
PRIVATE	2

Automatic error localisation: deterministic vs probabilistic approach

The consequence is that, in correspondence to values of the variable "position in Let us suppose that respondents ignore the rule for the compilation of the second question, and answer independently from the response given to the first question.the profession" different from 1, responses in the variable "do you work in public or private sector?" equal to" 1" and" 2" rather than [blank], are found in many cases.

Since almost all independent workers are in the private sector, the final effect of the error consists in a systematic error in the variable WORK, i.e. an increase of the frequency corresponding to the value "private", and a decrease of the frequency related to the value "public."

The presence of this situation is detectable by simply applying the rule:

IF (POSPRO ≠ 1) AND (WORK ≠ [blank]) THEN incoherence

Automatic error localisation: deterministic vs probabilistic approach

Under the probabilistic approach, the criterion of the minimum change is neutral, because it is possible to consider as erroneous both POSPRO than WORK.

So, the algorithm determines a change of the value of POSPRO from 2 to 1 in the 50% of the cases, and of WORK from 1 or 2 to [blank] in the remaining 50%.

But we know that a large percentage of the times this edit is failed, this is due to the fact that the respondent has ignored the rule for the compilation of the question related to the sector of work, and therefore it is WORK that actually contains the error.

If we ignore this consideration, systematic errors in WORK are certainly reduced, but new systematic errors in POSPRO are introduced, as POSPRO will be systematically imputed when its value is true.

On the contrary, the application of a deterministic rule such as

$$\text{IF (POSPRO} \neq 1\text{) AND (WORK} \neq \text{[blank]) THEN WORK} \leftarrow \text{[blank]}$$

is able to locate correctly the errors in WORK most of times

Automatic error localisation: mixed deterministic and probabilistic approach

Once the presence of systematic errors has been detected in data (e.g. by analysing the frequencies of the edit failures), the best solution is:

- first, locate and correct systematic errors by applying a deterministic procedure based on IF-THEN rules;
- then, locate random errors by applying a probabilistic procedure based on the minimum change algorithm

Imputation

The idea of imputation is to find subset of units (*imputation cells*) where the population of units with missing items (*recipient*) and population with no errors and no missing (*donor pool*) are similar.

Remember: we do not search for a unit such that the exact value is recovered but we look for similar populations so that the estimated mean is recovered (statistician), (*when we want estimate the mean of a quantity based on that).*