Statistical disclosure control and output checking

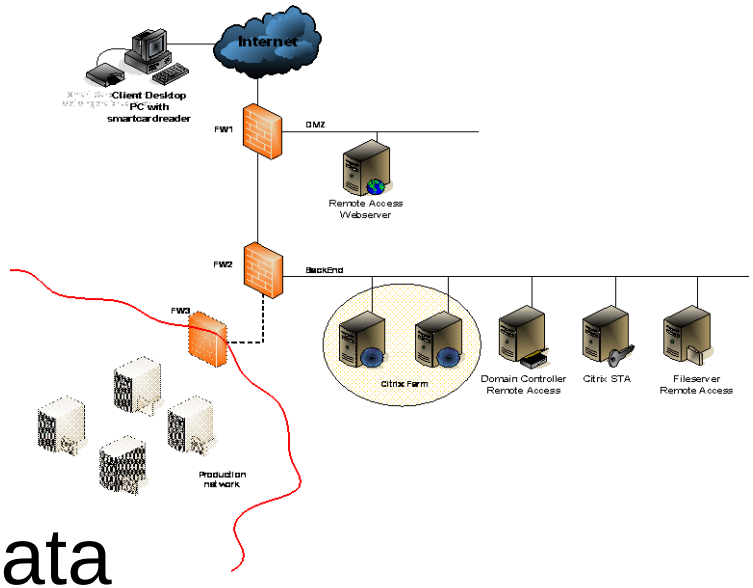
Remote access and on-site

Contents

- Settings
- Problems
- Guidelines
- Practical issues

Settings

Remote acces and on-site



- Data remain at NSI
- Researcher analyses data
- Results
  - Intermediate
  - Final

Settings

Check results that leave the NSI

- On-site: intermediate and final
- Remote access: (intermediate and) final

Settings

Output of OS/RA can be everything…

Transforming data, combining data, …

Set of rules that cover every possible output?

mission impossible?

Problems

What to check?

- Traditional output (tables)
- Descriptive statistics
- Output of many analyses
  - Regression
- Graphics

Problems (traditional)

Tabular output

- Frequency tables
- Magnitude tables

You now know all about that

Problems (descriptives)

Extreme values (max, min)

- Often extremes belong to identifiable respondents
- Extremes are very visible

Problems (descriptives)

Mean

- Single respondent?
- Two respondents?
- Dominating respondent?

- Group disclosure?

Problems (linear regression)

Observations $(x_{1i}, x_{2i}, x_{pi}, y_i)$, $i = 1, \ldots, n$

Linear regression: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$

with $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} \cdots & x_{p1} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} \cdots & x_{pn} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$

and $\boldsymbol{u}$ is n-vector with error-terms.

Problems (linear regression)

Estimator

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Simple case: $p = 1$, dummy variable

$$x_{1i} = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{else} \end{cases} \qquad \text{hence } \boldsymbol{X} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$$

Problems (linear regression)

Then, calculation of estimator yields

$$\hat{y}_1 = \cdots = y_1$$

Problems (linear regression)

Differencing two regressions, with deletion/addition of single observation

Subtract two regression coefficients

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0$$

Problems (linear regression)

May lead to information about individual values

- Known means
- Binary explanatory variables
- Binary dependent variable

Problems (graphics)

- Possible to identify individual respondent?
- Outliers present?
- Is underlying data safe?
- Does picture have “data embedded”?

Problems (graphics)

Guidelines

Started as deliverable of

ESSnet on common tools and harmonised methodology for SDC in the ESS

included in Wiley book

Statistical Disclosure Control

currently improved upon and extended in FP7

Data without Boundaries

Guidelines

Classify output as either

generally safe

or

generally unsafe

Guidelines

Generally safe:

researcher can expect to have output cleared with no or minimal changes

Generally unsafe:

not to be cleared unless researcher can demonstrate non-disclosiveness

Guidelines

Types of error with output checking:

- Releasing disclosive output
- Not releasing safe output

Guidelines

Two model approach:

- Rule-of-thumb 

  (**simple** rules that make a first distinction between safe output and potentially unsafe output)

- Principle Based 

  (further check potentially unsafe output)

Guidelines

Overall rules-of-thumb 

- 10 units
- 10 degrees of freedom
- No group disclosure
- No dominance

Guidelines (example)

Descriptive statistics

Max, min

 : Not released (refer to single unit)

 : Rules for magnitude tables apply (can be released if not associated with single unit)

Guidelines (example)

Descriptive statistics

Means, indices, ratios, indicators

should come from ≥ 10 units (unweighted)
largest contributor not more than 50%

indices considered more closely

Guidelines (example)

Complex indices like Fisher Price Index:

$$\sqrt{\frac{\sum_{j=1}^{m} p_{1,j} q_{0,j}}{\sum_{j=1}^{m} p_{0,j} q_{0,j}} \frac{\sum_{j=1}^{m} p_{1,j} q_{1,j}}{\sum_{j=1}^{m} p_{0,j} q_{1,j}}}$$

Simple indices like:

$$\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{N} Y_i}$$

Guidelines (example)

Complexity yields some protection

Simplicity:

- magnitude tables (X_i quantitative)
- frequency tables (X_i dichotomous)

Guidelines (example)

Linear regression coefficients

 : safe provided at least one estimated coefficient is withheld

 : check on degrees of freedom, not only categorical variables, not on a single unit

Guidelines (example)

Graphs

 : graphs are not allowed, unless based on undisclosive datapoints

 : data point not identified with units, no significant outliers, no “embedded” data (e.g. .jpg, .bmp)

Guidelines (overview)

- Frequency tables
- Magnitude tables
- Max, min, percentiles (incl. median)
- Mode
- Means, indices, ratios, indicators
- Concentration ratios
- Higher moments like (co)variance, kurtosis, skewness, …
- Graphics

Guidelines (overview)

- Linear regression coefficients
- Non-linear regression coefficients
- Estimation residuals
- Summary and test-statistics (R^2, χ^2, …)
- Correlation coefficients
- Factor analysis
- Correspondence analysis

Missing something? Contact us or Expert Group SDC.

Practical issues

- SDC awareness of researcher
- Joint responsibility of researcher and NSI
- Output checking is context specific
  - Check by SDC-expert and context-expert
- Consistency between checkers
- Costs and time-limits

Practical issues

Output checking is labour intensive!

Possibilities:

- Researcher facilitates NSI for checking
- Check sample of outputs
- Put price on checking of output
- Facilitate writing of reports at safe setting