Mechanisms for multivariate outliers and missing values

Beat Hulliger, Tobias Schoch
School of Business FHNW

NTTS2013, Brussels
The goal to turn the EU into the most competitive and dynamic economy by 2010 demands a full benchmarking system to monitor policy performance and their impact on progress. For this reason, the European Commission has engaged in selecting, collecting and analysing a set of indicators that are published each year. The Stockholm European Council has further emphasised the need for effective, timely and reliable statistics and indicators. A main challenge is to develop indicators for the main characteristics and key drivers. An utmost important and challenging area to be measured is social cohesion. Based on a clear definition of social cohesion, a universally-accepted high-quality and robust statistics to adequately measure social cohesion is required. Further, tools for measuring temporal developments and regional breakdowns to sub-populations of relevance will be of great importance. In order to measure social cohesion with Laeken indicators adequately while regarding national characteristics and practical peculiarities from the newly created EU-SILC, an improved methodology will be elaborated within AMELI. This will ensure that future political decision in the area of quality of life can be based on more adequate and high-quality data and a proper understanding of the Laeken indicators by the users. The study will include research on data quality including its measurement, treatment of outliers and nonresponse, small area estimation and the measurement of development over time. A large simulation study based on EU-SILC data will allow a simultaneous elaboration of the methodology focusing on practical issues aiming at support for policy. Due to the fact that the Laeken indicators are based on a highly sophisticated methodology the project’s outcome may also serve as a methodological complement for other FP7 projects in the area of indicators.

**PROJECT INFO**

**Project period:** 1 April 2008 - 31 March 2011

**Project No.:** 217322

**Workshops:**
- Vienna, Spring 2010
- Trier, Spring/Summer 2011

**COORDINATION**

Ralf Münnich
University of Trier
Faculty IV
Dept. of Economic and Social Statistics
Universitätsring 15, D-54286 Trier, Germany
muennich@uni-trier.de

**PARTNERS**

University of Trier
Federal Statistical Office of Germany
University of Applied Sciences NW Switzerland
Swiss Federal Statistical Office
Statistics Austria
Statistics Finland
University of Helsinki
Vienna University of Technology
Statistical Office of the Republic of Slovenia
Statistics Estonia

http://ameli.surveystatistics.net/
Contents

Introduction

Mechanisms

Inference

Final remarks
AMELI Project and EU-SILC data

- Advanced Methodology for European Laeken Indicators
- EU FP 7 research project
- Methods (Robustness)
- Evaluation with simulation study
- Data EU-SILC: disposable income constructed from some 18 household and 14 personal income components: highly multivariate with structural zeros and missing values
Multivariate outliers and missing values

- Outliers may occur in any direction of the space
- Missing values may occur for each dimension (in patterns)
- Outlier dimensions may be missing (all or some of them) because of outlyingness
- Further variables are available to help to nominate outliers in particular dimensions
- What is our target population?
Introduction

Robustness

- How to model outlyingness and missingness in real survey data?
- Basis for simulation?
- Under what conditions can we make inference on the intended population?
Classical contamination model

- Classical contamination model:

\[ F(y) = (1 - \varepsilon) G(y) + \varepsilon H(y), \quad \varepsilon \in [0, 0.5] \]

- Mixture distribution: Independent mixing with proportion \( \varepsilon \). No missing values. Often no bias \((E_G[Y] = E_H[Y])\).

- Two-phase procedure:
  - Outlyingness: determination of outlier, i.e. indicator per obs.
  - Contamination: determination of value of outlier, i.e. values per obs.
Representative outliers

- Chambers 1986: Outlier from the point of view of the sample: extrapolate or not?
- “not” means: observation values are incorrect or unique.
- practical, but not directly usable for simulation.
- how to distinguish between rep. and non-rep. outliers?
Mechanisms with dependence on observed values

- Missing at random (Rubin 1976)
- Outlying at random (Beguin and Hulliger 2008)
- Contamination at random (Hulliger and Schoch 2013)
A mechanism
Notation

- $Y^*$ is the population of inference
- $O^n$ indicator vector for non-rep. outliers
- $Y^n$ matrix of non-rep. outlier values
- $R$ response indicator matrix
Outlier mechanisms (narrow sense)

- Outlying at random (OAR): The outlier indicator $O^n$ may depend on (known) covariates $X$ but not on other variables.
- Contamination at random (CAR): The contamination $Y^n$ may depend on the (known) covariates $X$ but not on other variables.
- Outlyingness and Contamination completely at random (OCAR, CCAR) and not at random (ONAR, CNAR)
- idem for $O^r$
- In addition MCAR, MAR, MNAR.
Simulation set up (example)

- AAT-SILC universe from AMELI project: 8 Mio persons with 27 variables.
- ST SI sample, 6000 households (approx. 14000 persons); 1000 Monte Carlo replications;
- 1% outliers \(\approx 127.9\) outlier-persons.
- Outlyingness completely at random (OCAR)
- Contamination \textbf{not} at random (CNAR) (income \(\times 12\)).
- 0 and 2 % missing values completely at random (MCAR).
- R-packages \texttt{simFrame} and \texttt{DBsim}.

©2012 FHNW/Beat Hulliger
Inference

- Conditions for inference on true population $Y^*$?
- No valid inference if unobserved variables $Z$ influence the mechanisms!
- Assume sample design $S$ ignorable.
- Nice to know which observed data is good (non-outlying): $Y^{og}$
OMAR

- Assume $O^n$
- Outlyingness and Missingness jointly at random (OMAR):
  \[
  f(O^n_i, R_i| Y^*, Y^n, X; \xi^{nR}) = f(O^n_i, R_i| Y^{*og}, X; \xi^{nR})
  \]
- Observed likelihood can be integrated over observed bad (outlying) data and missing data.
- Proper inference based on $f(Y^{og})$ becomes possible (if $O^n$ were known).
Separability

- How to discover safely the latent outlyingness with $\hat{O}^n$?
- Conditions:
  1. There is a safe region, where only good data from $Y^*$ occurs but no (non-rep.) outliers.
  2. Not all data in the safe region is missing.
- We call these conditions “separable contamination”.
- “Clustering” of good and bad observed data.
- As long as there is a region free of non-rep. outliers the confounding with rep. outliers is not problematic.
- “Censoring” to safe region must be accounted for.
The interplay between outliers and missing values must be taken into account for simulation and for inference.

Mechanisms for simulation which are realistic may be formulated.

OMAR (joint outlyingness and missingness at random) and separable contamination is necessary to enable inference on the true population.

BACON-EEM algorithm (Beguin and Hulliger 2008) implicitly assumed these conditions.

These conditions rarely hold in practice!

But they help to understand the challenge.