Effect of multivariate outlier detection and imputation on poverty and inequality indicators

Beat Hulliger, Tobias Schoch
University of Applied Sciences Northwestern Switzerland
FHNW

SST2011, 25.10.2011, Fribourg
Motivation

Definition
An **income concept** defines the set of income component/sources (e.g., employee cash income, capital income, unemployment benefits, etc.) and defines the aggregation/composition rule to generate the income.
Motivation

- How do multivariate outliers in the income components affect the estimates of poverty and income-inequality measures?
- How can we successfully invoke Multivariate Outlier Detection and Imputation (MODI) methods?
- What is the impact of the methods on the poverty and income-inequality measures?
EU-SILC Data and Indicators of Social Cohesion

- EU-SILC: European Statistics on Income and Living Conditions
- 18 income components at the household level, 14 at the individual
- (Distributed) Aggregate: equivalized disposable personal income
- Key poverty and income-inequality measures in the EU
Multivariate Outliers in the Income Components

- Aggregation of income components may mask or propagate outliers
- Marginal distributions of income components are skewed and zero inflated
- Joint distribution is not elliptically contoured
- Structural zeros in multiple dimensions (an overwhelming majority of the observations lies in subspaces)
Parametric Detection: BACON-EEM

- BACON algorithm for complete, iid-data (Billor et al., 2000)
- BACON-EEM extended to work with missing values (EM-algorithm) and adapted for finite population sampling (Béguin and Hulliger, 2008)
- Model: elliptically contoured distribution, $EC(\mu, \Sigma)$
- Algorithm:
  - Small subset $G_1$ of outlier-free data: center $\mu_{G_1}$ and scatter $\Sigma_{G_1}$
  - Mahalanobis distance (MD) of all points
  - Choose a new (larger) subset, $G_2$, of observations with MD below a threshold
  - Iterate until convergence
- Output: $\hat{\mu}^R$ and $\hat{\Sigma}^R$ and outlier indicator $o_i, i = 1 \ldots , n$
Parametric Imputation

Treat Outliers as if they were missing values (TOaM)

- set outliers to missing
- impute for missing observations from $EC(\mu^R, \Sigma^R)$

Winsorized Gaussian Imputation (WGI)

- impute $\hat{y}_i = \mu^R + (y_i - \mu^R) \cdot (c/d_i)$, with $c$ a tuning constant and $d_i$ the MD of observation $i$ under model $EC(\mu^R, \Sigma^R)$ (Hulliger, 2007)
- preserves direction
Non-parametric Detection: Epidemic Algorithm (EA)

- Non-parametric data depth (Béguin and Hulliger 2004)
- Algorithm:
  - Start an epidemic at a central observation (e.g., spatial median) and let it spread through the point cloud. Late infected observations are outliers.
  - Transmission/ infection probability decreases with distance
  - Infection probabilities are estimated in the population (w.r.t. sampling design)
- Output: outlier indicator $o_i = 1$ if infection time large.
Non-parametric Imputation Reverse EA

- Requirement: outlier indicator $o_i$, $i = 1, \ldots, n$ (declared outliers)
- For each declared outlier and obs. with missing values:
  - start an epidemic at the outlier/obs with missings
  - stop when potential donor(s) infected
  - impute values of donor
- tuning of REA requires some experience
- high computational workload
- preserves neighbourhood
Data Processing

Simulation 1
- sampling
  - A-AT-SILC population of Alfons et al. (2010)
  - ST SIC

Prepare 1
- aggregate
  - work inc
  - capital inc
  - transfers pers
  - transfers hh

Prepare 2
- treatment of zeros
  - remember which obs is zero
  - exclude/ set to NA

Prepare 3
- transform (symmetry)
  - component-wise transform by e.g., \( \log_{10} \)

Simulation 2
- contaminate (outliers)
  - Replacement of a scaled point cloud
  - shift outliers

Simulation 3
- missing values
  - MCAR
  - MAR

©23. Oktober 2011 FHNW/Beat Hulliger
Data Processing (ctd.)

Simulation 1
- sampling
- Prepare 1: aggregate
- Prepare 2: treatment of zeros
- transform (symmetry)
- Prepare 3: contaminate (outliers)

Simulation 2
- missing values

Simulation 3
- detect
- MODI 1: detect
- MODI 2: impute
- Prepare 4: back-transform
- Prepare 5: equivalized disposable income
- finally: estimation

- BACON-EEM
- Epidemic Algorithm
- TOaM
- WGI
- REA
- work inc
- capital inc
- transfers pers
- transfers hh
- mean
- at-risk-of-poverty rate
- relative-median poverty gap
- quintile share ratio
- Gini coefficient

©23. Oktober 2011 FHNW/Beat Hulliger
Income Components: AT-SILC 2004, Public Use Sample
Contaminated Income Components: AT-SILC 2004 (PUS), log-scale
How to tune BACON-EEM?

Abbildung: Upper tail of the empirical CDF of the Mahalanobis distances
Relative bias (in %): Parametric MODI

$n = 6000$, $r = 1000$, contamination OCAR 1%, component-wise 0 or 2% MCAR, Imputation with TOaM

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0% NA</th>
<th>2% NA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>ARPR</td>
</tr>
<tr>
<td>8E-08</td>
<td>2.40</td>
<td>0.78</td>
</tr>
<tr>
<td>8E-09</td>
<td>2.41</td>
<td>0.79</td>
</tr>
<tr>
<td>8E-10</td>
<td>3.22</td>
<td>0.83</td>
</tr>
<tr>
<td>8E-11</td>
<td>4.63</td>
<td>0.86</td>
</tr>
<tr>
<td>8E-12</td>
<td>6.55</td>
<td>0.88</td>
</tr>
<tr>
<td>0</td>
<td>92.79</td>
<td>1.89</td>
</tr>
</tbody>
</table>
How to tune the Epidemic Algorithm

Average prop. false negatives

Average prop. false positives

©23. Oktober 2011  FHNW/Beat Hulliger
Relative Bias (in %): Non-parametric MODI

\( n = 6000, \ r = 1000, \) contamination 1\% OCAR, EA with DecOut outliers, Imputation with REA

<table>
<thead>
<tr>
<th>DecOut</th>
<th>Mean</th>
<th>QSR</th>
<th>GINI</th>
<th>ARPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>52.2</td>
<td>141.9</td>
<td>88.6</td>
<td>1.5</td>
</tr>
<tr>
<td>130</td>
<td>30.9</td>
<td>84.2</td>
<td>60.8</td>
<td>1.3</td>
</tr>
<tr>
<td>140</td>
<td>27.3</td>
<td>72.4</td>
<td>55.0</td>
<td>1.3</td>
</tr>
<tr>
<td>160</td>
<td>22.6</td>
<td>60.8</td>
<td>47.2</td>
<td>1.3</td>
</tr>
<tr>
<td>200</td>
<td>19.1</td>
<td>50.8</td>
<td>40.5</td>
<td>1.3</td>
</tr>
<tr>
<td>0</td>
<td>93.8</td>
<td>251.4</td>
<td>122.0</td>
<td>1.89</td>
</tr>
</tbody>
</table>
Conclusion

- Multivariate outliers have strong influence on inequality indicators
- BACON-EEM for detection good; Gaussian imputation unstable
- Epidemic Algorithm for detection moderate; Reverse EA imputation rarely better than BACON-EEM
- Important:
  - Income components must be aggregated/segmented into 4–6 variables (maybe up to 8 or 10)
  - Split data into subpopulations (using socio-economic status and other household characteristics)
  - (Robust) cluster analysis before MODI (data-based)
- \texttt{R}-package modi
- TOaM better than WGI
Literature


http://ameli.surveystatistics.net