# Multivariate Outlier Detection and Treatment in Business Surveys

Beat Hulliger
University of Applied Sciences Northwestern Switzerland FHNW

Montréal, 21 June 2007
ICES-III

# Content

# Introduction

# Multivariate outliers with missing values

- Outlier with missing values: If the outlier direction is not observed, the outlier cannot be detected!
- If values are missing because they are outlying we may not detect the outlier.
- We need a missing at random assumption (MAR) to impute missing values.
- MAR includes that, conditionally on observed data, unobserved outliers do not influence missingness.

# Mahalanobis distance with missing values

- Assume $m$ an estimate of the mean and $C$ an estimate of the covariance matrix
- For an observation $x_i$ let $C_{ioo}$ denote the sub-matrix of the covariance matrix with entries corresponding to $x_{io}$
- Marginal MD (Little and Smith 1987):

$$d_{io} = MD^2_{marg}(x_i) = \frac{p}{q_i}(x_{io} - m_{io})^\top (C_{ioo})^{-1}(x_{io} - m_{io})$$

  ($q_i$ the number of observed values)
- An observation $x_i$ is an outlier if $d_{io} > k$ for a constant $k$ to be chosen.

# Multivariate Outlier Detection Methods

# BACON for complete non-sampling data

Lit: Billor, N., Hadi, A.S. , and Vellemann, P.F. (2000)

Multivariate normal distribution:
outlier=large Mahalanobis distance for robust center and scatter.

| |
|---|
| Add non-outlying points to a small subset of good data as long as possible. |

- Robust: High breakdown point
- Tolerates a few outliers in the good subset
- Computationally fast
- Needs roughly elliptical distribution

# BACON-EEM algorithm

- Adapt BACON-algorithm to sampling: weighted mean and weighted covariance estimator
- Adapt EM-algorithm to sampling: estimate the quasi-likelihood from the sample (EEM)
- Combine BACON and EEM efficiently

Béguin and Hulliger (Submitted 2007 to Survey Methodology)

# ER-algorithm

- M-step of EM-algorithm: Do one robustification step (weights) (Little and Smith 1987)
- Non-robust start for robustification step!
- Original proposal without weights
- Here: Implementation in R with weights (EER).

# Transformed Rank Correlations

1. Calculate pairwise covariances with MAD and Spearman Rank Corelation (Gnanadesikan and Kettenring 1972).
2. Transform data to space of eigenvectors of $S$.
3. Calculate componentwise median and MAD and transform back into original space.

Maronna and Zamar 2002: iterate to convergence.

Béguin and Hulliger 2004: sampling and missing values.

# GIMCD

Robustify after non-robust EM-algorithm

1. Non-robust EM algorithm (unweighted): $m$ and $C$
2. Gaussian imputation under multivariate normal distribution with $m$ and $C$.
3. MCD algorithm on imputed data.

# MU281

- Data set MU284 (Särndal, Swensson, Wretman 1992) without the three largest municipalities $\longrightarrow$ MU281
- RMT85, ME84 and REV84 are divided by P85.
- Log of REV84/P85 and of P75.
- MAR with decreasing missingness for increasing P75.
- Hypothetical weighting: $w_i = 10$ if P75 $\leq 20$, otherwise $w_i = 1$.
- There are outliers in the original data: representative outliers.
- Additional artificial outliers: non-representative outliers.

# Detection of outliers in MU281

| miss. rate | outliers | ER | BEM | TRC | GIMCD |
|---|---|---|---|---|---|
| 10.7 | 34 | 18 | 24 | 27 | 20 |
| 10.7 | 85 | 43 | 66 | 69 | 71 |
| 30.1 | 85 | 42 | 61 | 44 | 64 |
| 30.1 | 108 | 56 | 85 | 65 | 43 |

- ER worst and slowest.
- GIMCD better than expected
- TRC good for low missingness rate
- BACON-EEM best when high missingness and outlyingness.

# Influential observations

# Influence

- Theory: Influence function (Hampel 1974).
- Sensitivity curve for sampling: Reaction of a statistic $T$ to a value $x$ replacing the value $y_i$ observed for observation $i$ in sample $S$.
- Sensitivity curve at $x = y_i$: **Impact**

$$SC(y_i; T, y_S, i) = n\left(T(y_S) - T(y_{S\setminus i})\right)$$

- $T(y_{S\setminus i})$ is the estimator $T$ evaluated at the sample without observation $i$, i.e. we treat $i$ as a complete non-response.
- $T$ can be a statistic on a sub-population.
- $T$ can be simple (Horvitz-Thompson) or complex (Quintile Share Ratio, Spearman Rank Correlation).

# Impact on Horvitz-Thompson type estimator

$$SC(y_i; T_{HT}, y_S, i) = nw_i(y_i - \hat{y}_i),$$

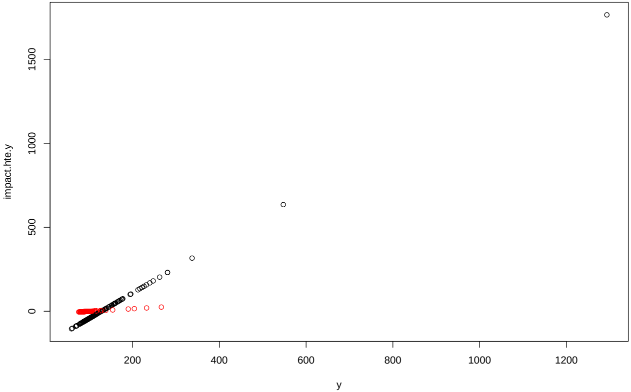
where $\hat{y}_i = \frac{\sum_{k \in S \setminus i} w_k y_k}{\sum_{k \in S \setminus i} w_k}$ is the Hajek-estimator based on the rest of the observations.

# Impact and selective editing

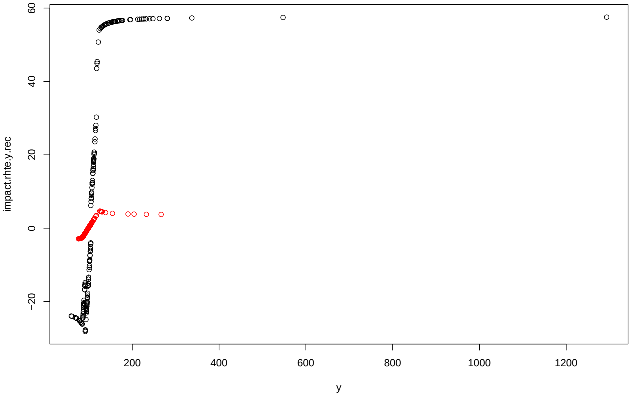
- Scores and impacts are closely related: Replace $\hat{y}_i$ in HT-impact by $\tilde{y}_i$ to obtain the local score $s_i = w_i(y_i - \tilde{y}_i)$:
- Some scores are very complex (e.g. Hidiroglou-Berthelot score) and relation to impact is unclear.
- Only particular impacts are covered by the scores: No guarantee for limitation of impact on other estimators!

EDIMBUS-RPM: Project of ISTAT, CBS, SFSO to develop a manual on Editing and Imputation for Cross-Sectional Business Surveys. Partially funded by Eurostat.

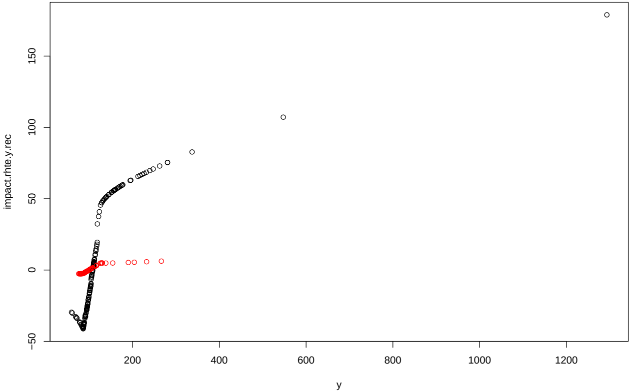
# Impact on HTE of rev84



# Impact on RHTE of rev84, robustified on rev84



# Impact on RHTE of rev84, robustification on lre84



# Winsorization and Imputation

# Winsorization

- Mahalanobis distance of observed part of outlier $d_{io}$ with $m$ and $C$ robust.
- Robustness weight $u_i$: $u_i = k/d_{io}$ if $d_{io} > k$ for a tuning constant $k$, otherwise $u_i = 1$.
- Winsorization for observations with $u_i < 1$:

$$\hat{x}_{io} = m_o + u_i\,(x_{io} - m_o). \tag{1}$$

  For $d_{io} \leq k$,i.e. $u_i = 1$ we have $\hat{x}_{io} = x_{io}$, i.e. no change.
- We may choose another tuning constant for imputation than for detection to allow for representative outliers.

# Gaussian imputation

- Imputation of missing values given the observed values under the multivariate normal model with or without error term.
- $\hat{x}_i = (\hat{x}_{io}, \hat{x}_{im})^\top$, with
  $\hat{x}_{im} = m_m + C_{mo} C_{oo}^{-1} (\hat{x}_{io} - m_o) + \epsilon_m$
- Implementation with package `norm` of R.
- To prevent imputation of outliers: Winsorize before imputation!

# MU281: Weighted means with TRC

| data | rmt85 | me84 | lre84 | lp75 |
|---|---|---|---|---|
| complete | 6.92 | 49.86 | 2.061 | 1.059 |
| complete winsorised | 6.90 | 49.53 | 2.044 | 1.061 |
| raw | 6.94 | 49.97 | 2.062 | 1.059 |
| raw winsorised | 6.93 | 49.80 | 2.049 | 1.060 |
| imputed | 6.91 | 49.70 | 2.047 | 1.060 |

# MU281: Weighted correlations with TRC

| data | rmt85,me84 | rmt85,lre84 | me84,lre84 |
|---|---|---|---|
| complete | 0.630 | 0.151 | 0.182 |
| complete winsorised | 0.624 | 0.159 | 0.005 |
| raw | 0.625 | 0.120 | 0.130 |
| raw winsorised | 0.627 | 0.098 | 0.022 |
| imputed | 0.671 | 0.083 | -0.036 |

# Some Conclusions

# Methods

- MOD: BACON-EEM, TRC
- GIMCD should be researched better
- Gaussian imputation after winsorization is relatively simple but more research is needed, e.g. comparison with Nearest Neighbour Imputation with robust metric (POEM).

# Influence and outliers

- The scores of selective editing often are particular instances of impacts: Selective editing cannot protect all possible statistics.
- Outliers and influential observations do not necessarily coincide, in particular not, when the model involves transformations.
- Check outliers and impacts on the result of your interest during macro-editing, even if selective editing was applied.