Handling of outliers at SFSO

Beat Hulliger and Daniel Kilchmann

Statistical Methods Unit
Swiss Federal Statistical Office

25. - 27. September 2006

Introduction

- Until 1990, robust estimation methods had rarely been applied in public statistics due to technical limits and the complexity of some methods.
- Outliers were normally treated manually.
- Theoretical and technical development allowed the implementation of new estimators for detection and treatment of outliers.

Robust methods applied at SFSO

SFSO developed and/or implemented procedures for several estimators adapted to sampling weights.

- Trimming one-step estimator for univariate outlier detection and treatment.
- Woodruff method for estimating the confidence interval of the Median.
- One-step ratio estimator for outlier detection and treatment (re-weighting).
- L_1-regression for outlier detection and imputation.
- Transformed rank correlations estimates used for the definition of a robust Mahalanobis distance. This distance function is used for multivariate outlier detection and for nearest neighbor imputation.

Surveys where robust methods were applied

- Survey on housing rents.
- Earning structure survey.
- Enterprise census.
- Environment protection expenditures.
- Survey on production and value added.
- Survey on energy consumption.
- Retail trade statistics.
- Hospital statistics.
- Survey on income and expenditure (Household budget survey).

Experience with robust methods in surveys

- Robust methods help to limit the influence of outliers.
- Robust methods must be adapted to sampling.
- One-step estimators are good approximations of M-estimators.
- Choice of tuning constants is often difficult (possible bias).
- The degree of robustification normally must be discussed with subject matter specialists.
- Outliers must often be checked individually to decide how they should be treated.
- Robustification should be limited to the most extreme observations; be aware of 'over-robustification'.

- A few large weights may dominate the estimation. $\longrightarrow$ Analyse the sampling weights (median-dominance).
- The total weight (robustness weight $\times$ sampling weight) may become much smaller than 1. $\longrightarrow$ Lower limit for the robustness weights.
- The definition of an outlier depends on models which must be checked carefully. The model should be adequate for the bulk of the data. $\longrightarrow$ Different models in different sub-populations.

Annexe: One-step ratio estimator adapted to sampling

- Adaptation of a ratio estimator.
- Initial robust estimate of the slope $\longrightarrow$ residuals.
- Downweight the observations with extreme residuals with a robustness weight (decision based on tuning constant).
- Robust re-estimation of the slope with sampling weights $\longrightarrow$ one-step estimator of the slope.
- The one-step estimator can be used as initial estimation of the slope for the next iteration step. Convergence $\longrightarrow$ weighted M-estimator.
- Mean of the robustness weights should not be much below 1.

Cf. (Hulliger 1995), (Hulliger 1999), (Peters, Renfer, and Hulliger), (Salamin 2005) and (Bendel, Scherer, Salamin, and Gülden 2006)

Annexe: L_1-regression

L_1-regression: least absolute deviation regression (LAD-regression).

Minimize the absolut values of the residuals r of the linear regression model.

Outlier detection:

- Lower and upper limits for outlier detection (boxplot): $r_{0.25} \pm 1.5 \times (r_{0.75} - r_{0.25})$, with r_p the p-th quantile of r
- $y_i > c\hat{\beta}_{L_1}$ flagged as outlier

Imputation: $\hat{y}_i = x_i^T \hat{\beta}_{L_1}$

Cf. (Oetliker 2002), (Renfer 2006)

Annexe: Transformed rank correlations (TRC)

- Robust estimate of center m and covariance matrix S based on bivariate Spearman rank correlations.
- Detection of outliers with Mahalanobis distance.
- Adaptation to sampling and missingness

Cf. (EUREDIT Project 2004a), (EUREDIT Project 2004b), (Béguin and Hulliger 2004), (Kilchmann 2006).

Annexe: Woodruff method

Back-transforming a confidence interval on the probability with the inverse of the empirical distribution function.

Cf. (Peters and Hulliger 1996), (Graf 2002).

Annexe: Median-dominance

- Minimal part of the largest weights accounting for more than 50% of the total weights.
- The lower $\mathrm{dom}_{0.5}$ is the more unbalanced are the weights.
- E.g. if $\mathrm{dom}_{0.5} = 30\%$ then the empirical breakdown point of the weighted median is 30% instead of the 50% of an unweighted median.
- Winsorization of the weights may resolve the problem.
- The use of the unweighted median in the one-step ratio estimator my be an alternative.

Biblography

Béguin, C. and B. Hulliger (2004).
Multivariate Oulier Detection in Incomplete Survey Data: The Epidemic Algorithm and Transformed Rank Correlations.
J.R.Statist.Soc.A 167(Part 2.), 275–294.

Bendel, R., R. Scherer, P.-A. Salamin, and J. Gülden (2006).
Energieverbrauch in der Industrie und im Dienstleistungssektor. Resultate 2002 bis 2004.
Report, Swiss Federal Institute of Energie, Bern.

EUREDIT Project (2004a).
Methods and Experimental Results from the Euredit Project, Volume 2.
http://www.cs.york.ac.uk/euredit/results/results.html.

EUREDIT Project (2004b).
Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project, Volume 1.
http://www.cs.york.ac.uk/euredit/results/results.html.

Graf, M. (2002).
Enquête suiss sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé.
Methodological Report 338-0010, Swiss Federal Statistical Office.

Hulliger, B. (1995).
Outlier Robust Horvitz-Thompson Estimators.
Survey Methodology 21(1), 79–87.
Statistics Canada.

Hulliger, B. (1999).
Simple and Robust Estimators for Sampling.
In *Proceedings of the Section on Survey Research Methods*, pp. 54–63. American Statistical Association.

Kilchmann, D. (appears 2006).
Krankenhausstatistik und Statistik der sozialmedizinischen Institutionen 1999-2004. Einsetzungsverfahren.
Methodological Report 338-00XX, Swiss Federal Statistical Office, Neuchâtel.

Oetliker, U. (2002, Août).
Traitement des données manquantes et aberrantes dans le domaine des revenus de l'enquête sur les revenus et la consommation (ERC98).
Mémoire du diplôme postgrade en statistique, Université de Neuchâtel.

Peters, R. and B. Hulliger (1996, June).
Schätzverfahren für die Lohnstruktur-Erhebung.
Methodological report, Swiss Federal Statistical Office.

Peters, R., J.-P. Renfer, and B. Hulliger.
Technical report.

Renfer, J.-P. (appears 2006).
Enquête sur les chiffres d'affaire du commerce de détail. Elaboration du plan d'échantillonnage et méthodes d'estimation.
Methodological Report 338-00XX, Swiss Federal Statistical Office, Neuchâtel.

Salamin, P.-A. (2005).
Extrapolation pour la statistique de la consommation d'énergie.
In *Swiss Statistics Meeting*.