Robust Multivariate Methods for Income Data

Hulliger, Beat
University of Applied Sciences Northwestern Switzerland, School of Business
Riggenbachstrasse 16
CH-4600 Olten, Switzerland
E-mail: beat.hulliger@fhnw.ch

Schoch, Tobias
University of Applied Sciences Northwestern Switzerland, School of Business
Riggenbachstrasse 16
CH-4600 Olten, Switzerland
E-mail: tobias.schoch@fhnw.ch

With the EU Statistics on Income and Living Conditions (EU-SILC), the European Union established a coordinated survey and adopted a set of indicators (Laeken indicators) to tackle poverty and social exclusion. In particular, the monetary Laeken indicators are based on the equivalized disposable income, an aggregation of person- and household-specific income components (e.g., income from employment and capital; unemployment-, old-age-, survivors'-, and disability benefits, etc.).

Technically speaking, the income components at the individual level (expressing those figures that are exclusively measured at household-level as per capita) span a multidimensional space with the following characteristics: (1) the marginal distribution of each component is very skewed and features a remarkable point mass at zero, (2) thus, the joint distribution of the components is far from being elliptically contoured (even after appropriate transformation), (3) an overwhelming majority of observations lies on subspaces i.e., exhibits intrinsic zeros on certain dimension (e.g., individuals on working age with a positive employee-cash income do neither receive old-age nor unemployment benefits, and vice versa), (4) within subspaces, the observations are clustered with respect to non-monetary, socio-economic characteristics. In addition, and due to the complex nature of the complete data, the problem of missing values is accentuated. Further, the skewed marginal distribution of each component in connection with the complex overall data structure and the missing values renders outlier detection very difficult.

We are interested in the effect that any treatment or slight modification of the components has on both the equivalized disposable income and the measures of social cohesion. In particular, we show that outliers may have a considerable effect on the estimates of the Laeken indicators. We require the methods (1) to be capable to treat the intrinsic zeros appropriately, (2) to work with missing values, (3) to cope with the complex nature of the data, (4) to take the sampling design into account, and (4) to be computationally practicable.

The proposed methods have been studied by extensive simulation in the EU FP 7 project AMELI (Advanced methods for European Laeken Indicators). For further information, visit our webpage http://www.ameli.surveystatistics.net.

REFERENCES (RÉFERENCES)

Béguin, C. and B. Hulliger (2008): The BACON-EEM Algorithm for Multivariate Outlier Detection in Incomplete Survey Data, Survey Methodology 34, 91-103.

Béguin, C. and B. Hulliger (2004): Multivariate Outlier Detection in Incomplete Survey Data: the Epidemic Algorithm and Transformed Rank Correlations, Journal of the Royal Statistical Society, Series A: Statistics in Society 167,275-294.

Charlton, J. (ed.) (2003) Towards Effective Statistical Editing and Imputation Strategies Findings of the EUREDIT Project, Vol. 1 and 2, EUREDIT consortium, www.cs.york.ac.uk/euredit/results/results.html

Hulliger, B. and T. Schoch (2011): Robust Multivariate Methods for Income Data, in Proceedings of the New Technologies and Techniques (NTTS) Conference. Brussels, Eurostat, Februrary 2011.