New Strategies in Calibration Estimation at Statistics Belgium

Camille Vanderhoeft
Statistics Belgium, Department for Methodology and Co-ordination
Rue de Louvain, 44, B-1000 Brussels, Belgium
camille.vanderhoeft@statbel.mineco.fgov.be

1. Introduction

Traditionally, Statistics Belgium uses post-stratification techniques for calculation of extrapolation coefficients for sample surveys. Since the last decade, more sophisticated techniques are made available within the framework of generalised calibration. Deville and Särndal (1992) and Deville *et al* (1993) are setting out the general statistical theory, and many others after them have investigated its applicability in daily practice. This has resulted in several advanced software modules, among which we mention Calmar, Clan, GES, Bascula. Others are still showing up. One of the results of our study is a set of modules, called g-CALIB, which provides support for calibration estimation under SPSS. Statistics Belgium is likely to benefit from it, although general implementation in daily practice is still far from realised.

2. Generalised calibration as a convex programming problem

In Vanderhoeft (2001), a thorough understanding of the calibration problem could be achieved, by formulating it, for a given sample, as a convex mathematical programming problem: $\{\min \mathbf{d}^T G(\mathbf{g}); \tilde{\mathbf{X}}^T\mathbf{g} = \mathbf{t}, \mathbf{g} \in \Omega_B\}$, where $\mathbf{d}$ is the n-vector of initial (sampling) weights, $\mathbf{g}$ is the n-vector of g-weights, $\mathbf{X}$ is the $n \times m$ calibration design matrix, $\tilde{\mathbf{X}} = \mathbf{X} \cdot diag(\mathbf{d})$ is the expanded design matrix, G is the "distance" function, and Ω_B is a rectangular (bounded or unbounded) subset of n-dimensional Euclidean space, determining additional boundaries for the g-weights. For some G the user may define a set $\Omega_B = [L,U]^n$. The problem is solved by Lagrange-methods. The final vector of calibrated weights is $\mathbf{w}(\lambda) = diag(\mathbf{d})\mathbf{g}(\lambda) = diag(\mathbf{d})F(\tilde{\mathbf{X}}\lambda)$, where F is the calibration function (inverse of the derivative of G) and λ is the vector of Lagrange multipliers. The Lagrange-method provides an updating formula $\lambda^{(l)} = \lambda^{(l-1)} - \left(\tilde{\mathbf{X}}^T diag\left(\mathbf{w}(\lambda^{(l-1)})\right)\tilde{\mathbf{X}}\right)^- \left(\tilde{\mathbf{X}}^T \mathbf{w}(\lambda^{(l-1)}) - \mathbf{t}\right)$, from which the Lagrange multipliers can be found iteratively. We have shown that the choice of the g-inverse $(...)^-$, as well as the choice of the representation of the calibration design matrix $\mathbf{X}$, does influence neither the solution, nor the iterative path to the solution of the calibration problem.

Another interesting result involves the most extreme values L and U for which the problem still has at least one solution. Provided that the calibration constraints are consistent, L and U can be found by solving the related convex programming problem $\{\min \|\mathbf{g} - \mathbf{1}_n\|_\infty; \tilde{\mathbf{X}}^T\mathbf{g} = \mathbf{t}, \mathbf{g} \geq 0\}$, where $\mathbf{1}_n$

is a target point (could be any other point), and where the constraints $\mathbf{g} \geq 0$ are included to avoid a negative solution for L. This new optimisation problem does not involve the distance function G.

3. Calibration on element-level and/or cluster-level auxiliary information, and g-CALIB

Having formulated the basic calibration problem as in the previous section, it became quickly apparent how (1°) to structure the software g-CALIB, and (2°) to extend its applicability for handling more complex situations. The main input for the basic model is (**X**, **d**, **t**). We may have several such input sets if auxiliary information is available at several levels, e.g. for elements (individuals, local units, ..), for clusters (households, enterprises, ..), …We outline how, if available, element-level and cluster-level auxiliary information can be used simultaneously. It is shown that the calibration problem can be transformed into the basic form (section 2). Next we show the advantage of designing calibration software in order to handle the required transformations of elementary input data sets automatically. This feature is implemented in our modules g-CALIB. Hopefully, this will speed up the process of introducing new calibration techniques at Statistics Belgium. Of course, others have introduced such techniques for integrated multi-level auxiliary information in the calibration process, but it may become fully operational only if calibration software provides features for automated data manipulation for that purpose.

More strengths and weaknesses of g-CALIB will be discussed.

4. Applications

The first application is to the Structural Business Survey. We deal with the problem of over-coverage, and we shall point out future extensions of the current calibration model for the SBS.

Our second application is to the Labour Cost Survey. Here, aggregated data are available in contingency tables and adjustments need to be made such that known marginal totals are fitted. It is shown how contingency tables can be easily dealt with in g-CALIB, although the software was initially intended to deal with non-aggregated data only. This is closely related to a clustering technique applied to calibrate simultaneously on element- and cluster-level auxiliary data.

REFERENCE

Vanderhoeft, C. (2001) Generalised Calibration at Statistics Belgium – SPSS® Modules g-CALIB-S and Current Practices. Working Paper, ix + 192 pp. http://statbel.fgov.be/news/studies/home_en.htm

RESUME

L'Institut National de Statistique (INS) Belge a commencé d'adopter des méthodes avancées de calage. Nous présenterons quelques résultats théoriques d'une étude approfondie (Vanderhoeft, 2001), et nous discuterons des applications (potentielles). En plus, cette étude a mené à des modules SPSS, nommé g-CALIB, qui nous permettront d'améliorer les techniques d'extrapolation d'une manière uniforme à l'INS. Les exemples choisis concernent l'Enquête structurelle des Entreprises, et la correction des statistiques agrégées sur le Coût du Travail.