Calibration Techniques

Olivier Sautory

Calibration techniques, as developed by Deville and Särndal [3] and [4], can be used to adjust a sample, through individual re-weighting using available auxiliary information for a certain number of variables, known as calibration variables. Such calibration weightings are used to calibrate the sample of known population totals of quantitative variables, and of known population distributions for levels of categorical variables. They also improve the accuracy of estimates of variables of interest that are highly correlated with calibration variables.

INSEE has applied these techniques since 1990 using the SAS Calmar macro[1] (see Sautory [10]). Calmar is an abbreviation for Calibration on Margins: this refers to the technique used to adjust the margins (from sample estimates) of a contingency table, cross-tabulating two (or more) category variables, to the known population margins. However, the application is more general than "calibration on margins" strictly speaking, as it can be used to calibrate quantitative variable totals.

The R Icarus package, developped by A. Rebecq [8], can also be used to apply these methods. It is available on CRAN.

1. Calibration: Theoretical Aspects

1.1 The Problem and the Solution

We consider a population U of individuals, from which probability sample s is selected. For each individual k in U, π_k denotes the probability of its inclusion in s. ***Y*** is a variable of interest for which we would like to estimate the total in population $Y=\sum_{k\in U} y_k$.

The estimator for Y using survey data takes the form $\hat{Y}=\sum_{k\in s} d_k y_k$ in almost all cases, where values for d_k are estimation weights associated with sample observations. These weights are often "survey weights", equal to the inverse of the probabilities of inclusion π_k: from this, we obtain the Horvitz-Thompson estimator: $\hat{Y}_{\text{HT}}=\sum_{k\in s}\frac{1}{\pi_k} y_k$.

We assume that the totals in population J of auxiliary variables[2] $\boldsymbol{X_1} \dots \boldsymbol{X_j} \dots \boldsymbol{X_J}$, available for all survey observations, are known: $X_j=\sum_{k\in U} x_{\text{jk}}$

We will try to obtain new weightings, "calibration weights" w_k that are as close as possible, according to a "distance function" G, to the initial weightings d_k, and which are used to calibrate the totals of variables X_j, i.e. which verify the **calibration equations**:

$$\forall\ j=1\cdots J \sum_{k\in s} w_k x_{\text{jk}}=X_j \quad (1)$$

The distance function G, with argument r = w_k / d_k, used to measure the distances between values for w_k and d_k is positive and convex, and verifies G(1) = 0. The unknown weights w_k minimise the value $D=\sum_{k\in s} d_k G(w_k/d_k)$ under calibration constraints (1).

1 available to download from the INSEE website, www.insee.fr

2 These are quantitative or indicative variables associated with levels of categorical variables.

The solution to this problem is given by $w_k = d_k F(x_k' \lambda)$, where $x_k' = (x_{1k} \cdots x_{Jk})$, λ is a vector of J Lagrange multipliers linked to the constraints (1). F, known as the calibration function, is the inverse function of the derivative of function G.

Vector λ is obtained by solving the non-linear system of J equations with unknown values for J resulting from calibration equations: $\sum_{k \in s} d_k F(x'_k \lambda) x_k = X$ where X is the vector of totals for X_j.

We can solve this system numerically using Newton's iterative method; we calculate a series of vectors $\lambda^{(i)}$ defined by a recurrence relation, initialising the algorithm with vector $\lambda^{(0)} = 0$. Convergence is obtained where the ratio of weights w_k / d_k obtained in two successive iterations "almost stops moving":[3]

$$\underset{k \in s}{\mathrm{Max}} \left| \frac{w_k^{(i-1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \varepsilon$$

When calibration weights w_k have been calculated, the estimator of all variable of interest totals ***Y*** will be the "calibrated" estimator, in the form $\hat{Y}_w = \sum_{k \in s} w_k y_k$.

1.2 Calibration Functions

Four calibration methods, corresponding to four distance functions, are provided in the SAS Calmar macro and the R Icarus package. They are defined by the form of function F. Below we indicate for each method the function G(r) (where $r = w_k / d_k$ denotes the "weight ratio"), and function F(u) (where $u = x_k' \lambda$).

a) <u>"Linear" Method</u>

$$G(r) = \frac{1}{2}(r-1)^2,\ r \in \mathbb{R} \qquad F(u) = 1 + u\ (\in \mathbb{R})$$

D is therefore a chi-squared distance between weights d_k and w_k. The linear form of F gives its name to this method, and the calibrated estimator is therefore the generalised regression estimator:

$$\hat{Y}_{\text{reg}} = \hat{Y}_{\text{HT}} + \left(X - \hat{X}_{\text{HT}}\right) \hat{B}_s \qquad \text{where } \hat{B}_s = \left(\sum_{k \in s} d_k x_k x_k'\right)^{-1} \left(\sum_{k \in s} d_k x_k y_k\right)$$

This is the quickest method because Newton's algorithm always converges after two iterations. It can generate negative w_k weights, and the weights have no upper limit.

b) <u>"Exponential" or "Raking Ratio" Method</u>

$$G(r) = r\,\mathrm{Log}\,r - r + 1,\ r > 0 \qquad F(u) = \exp u\ (> 0)$$

D is then an "entropic" distance between weights d_k and w_k. Where the auxiliary variables are categorical variables for which the size of population levels is known, selecting function G results in a conventional adjustment method, proposed by Deming and Stephan [2], known as a raking ratio; it is also referred to (particularly in SAS) as IPF ("Iterative Proportional Fitting").

This method results in positive weights that have no upper limit, which are generally higher (for the highest weights) than those obtained for the linear method.

3 ε is a threshold defined by the user (e.g. 10^{-4})

c) "Logit" Method

We select two real numbers for L and U such as L < 1 < U.

$$G(r) = \left[(r - L)\ \mathrm{Log}\frac{r - L}{1 - L} + (U - r)\ \mathrm{Log}\frac{U - r}{U - 1} \right] \frac{1}{A},$$ if L < r < U (and + ∞ otherwise) with

$$A = \frac{U - L}{(1 - L)\ (U - 1)}$$

$$F(u) = \frac{L(U - 1) + U(1 - L)\exp(Au)}{U - 1 + (1 - L)\exp(Au)} \in]L, U[$$

This method is named after the logistic form for function F and ensures that the ratios of weights w_k / d_k are included in interval]L, U[. However, we cannot select a priori any values for L and U: in general there is a maximum value L_{max} for L (less than 1), and a minimum value U_{min} for U (more than 1). These values depend on the data and the calibration margins: the more the sample structure varies from that of the real population one as regards calibration variables, the further these values are from 1.

d) "Truncated Linear" Method

We select two real numbers for L and U such as L < 1 < U.

$$G(r) = \frac{1}{2}(r - 1)^2$$ if L ≤ r ≤ U (+ ∞ otherwise) $$F(u) = 1 + u \in [L, U]$$

This method ensures that ratios w_k / d_k fall within the interval [L, U], and as for the "logit" method, L_{max} and U_{min} generally exist.

The logit and truncated linear methods are used most often, as they help to avoid over-weighting, which may reduce the robustness of estimations, and under-weighting, weights less than 1 or even negative, which may be obtained using the linear methods.

1.3 Precision

Calibrated estimators $\widehat{Y}_w$ all have the same precision (asymptotic), regardless of the method used: the variance approximate to $\widehat{Y}_w$ is therefore equal to the regression estimator $\widehat{Y}_{reg}$: the variance falls, the higher the correlation between the variable of interest ***Y*** and calibration variables $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$.

If we use a formula - or software - to estimate the variance of the Horvitz-Thompson estimator for variable of interest ***Y***, the variance of $\widehat{Y}_w$ is obtained by replacing y_k values in the formula by the regression residuals (weighted by the d_k or calibration weights w_k) for ***Y*** on $\boldsymbol{X_j}$ values in sample s.

2. Calibration in the Presence of Total Non-response

Total non-response is usually adjusted using re-weighting methods for respondent units. The two main calibration strategies that may be implemented in the presence of total non-response are as follows (see methodological note *Adjusting for non-response by re-weighting, in particular §V. Margin calibration):*

- calibration after adjusting for non-response: in the first instance, the total non-response is adjusted for by re-weighting, followed by a standard calibration using the non-response-adjusted responses $d_k^a = d_k / \widehat{p}_k$, where $\widehat{p}_k$ are estimates of the probability responses (e.g. using for example the homogeneous response groups method);

- “direct” calibration using respondent survey weights. This is justified where the calibration variables contain explanatory variables of non-response, and where a particular form of non-response model is assumed (generalised linear model linked to the chosen calibration function) (see Dupont [5]).

3. Penalised Calibration

(this paragraph borrows heavily from the presentation by Rebecq [8]; see also presentation by Rebecq [9] at the INSEE statistical methodology seminar from15 March 2016).

With penalised calibration, it is accepted that calibration is not completed perfectly on some margins in a way that facilitates convergence, thereby enabling an increase in the number of variables for which the estimate after calibration is “controlled”, while maintaining a narrow distribution of weight ratios. The method (see Beaumont and Bocci [1], and Guggemos and Tillé [7]) involves releasing calibration constraints and incorporating them within the optimisation programme.

$\widehat{X}_w = \sum_{k \in s} w_k x_k$ denotes the vector of estimators of calibration variable totals using “calibration weights”. We use a cost vector C, equal in size to the number of calibration variables J, and diag(C) denotes the diagonal matrix of dimensions J×J where the diagonal coefficients are values of vector C.

The penalised calibration program is written as follows:

$$\min_{w_k} \sum_{k \in s} d_k G(w_k / d_k) + \lambda (\widehat{X}_w - X)' \operatorname{diag}(C)(\widehat{X}_w - X)$$

Parameter λ is a value between 0 and +∝ and represents the relative importance assigned to the distance between final weights and initial weights (the first part of the function should be minimised), by comparison with the deviation at margins X of adjusted estimates $\widehat{X}_w$ (the second part of the function should be minimised, the “cost” part). Where λ→ +∝, the cost term is preponderant: margin constraints are satisfied first of all, which removes weight ratios of 1. Where λ→ 0, the distance term is preponderant: weight ratios approach 1, but margin constraints are released to a large extent.

The cost linked to a calibration variable $\boldsymbol{X_j}$ increases, the more we wish to see “close” proximity between $\widehat{X}_{jw}$and the total X_j. By setting an infinite cost, we can require an exact calibration.

The penalised calibration may be implemented using the R Icarus package. The user selects the cost vector and the value of a *gap* parameter that gives a maximum value for the distribution range for weight ratios: the programme then determines the highest value for λ. It is not (at present) possible to impose *a priori* an estimation error for calibration variables: statisticians adjust both cost and gap parameters to empirically obtain a satisfactory solution.

We can use a number of distance functions G. But using a “bounded” distance is not necessary on account of the *gap* parameter. The two methods proposed by Icarus are therefore the linear method, for which an analytical solution exists, and the exponential method (Icarus uses the ICRS algorithm described by Bocci and Beaumont).

4. Practical Aspects of Calibration

4.1 Calibration Variables

A variable may be used in a margin calibration on the dual condition that it is available for all sample observations used in calibration, and that its population total is known. It could therefore be variables on the survey frame, or variables measured when collecting survey data and for which the total is known through other sources. In the latter case, it is essential that the variable available on the survey corresponds exactly to the variable for which the total is known: variables must be measured at the same time, according to the same concepts, the total that is settled on must correspond to the population that the sample aims to describe.

Ideally, the exact total should be known. It may also be estimated using another survey where such a source can be used to obtain more accurate estimators than the survey to which calibration has been applied. In practice, we may use a survey to calculate margins for another survey provided that the sample for the first survey is ten times larger than that for the second. Therefore, for INSEE household surveys, many calibration margins are calculated using the Labour Force Survey (LFS).

4.2 Out-of-scope Units and Calibration

Due to imperfections in the sample frame from which samples are taken, units surveyed may not actually belong within the survey's target population (i.e. the survey "scope"): businesses that have shut down all operations, accommodation units that are vacant or used as a secondary residence are typical examples of out-of-scope units in INSEE surveys. Out-of-scope units are most often detected at the collection stage.

Where the margins used in calibration are taken from the sample frame, out-of-scope units detected in collection must be used in calibration: margins calculated in the sample frame are in fact relative to a population that contains both out-of-scope units and units within the survey scope. If, on the other hand, margins only relate to the survey scope, out-of-scope units must not be used in calibration. Where margin calibration uses margins relating to the survey scope as well as margins calculated using the sample frame, the out-of-scope sample units detected in collection must be used in calibration, but, for these observations, the values of calibration variables for margins relating to the scope are set to 0.

5. Examples

5.1 Annual Sectoral Surveys

Annual Sectoral Surveys (ESAs) are used to provide a breakdown of annual turnover figures for French companies by business area. The survey includes approximately 160,000 companies, of which half - the largest - are systematically collected (probability of inclusion set to 1), while the other half is selected at random from French small and medium-sized enterprises.

After applying methods for non-response adjustment methods and handling prominent values (see methodological note *Non-response adjustment through re-weighting* and *Handling prominent values in surveys*), calibration is completed: insofar as possible, it consists of calibrating non-exhaustive strata in the survey for the total annual turnover per NAF (2008) group and the number of units per division of the NAF 2008 in the population without exhaustive strata. For this, we use the Calmar macro with the truncated linear methods as the default parameter, forcing weight ratios to stay within the range [0.5; 2]. Where the calibration does not converge, we widen the weight ratio bounds, and if this is still not sufficient, sectors are grouped together in calibration.

5.2 Household Surveys on Information and Communication Technology (ICT)

This annual survey combines a number of collection methods: a sample of 3,500 households is surveyed by telephone, a sample of 22,500 households is surveyed online/by hard-copy questionnaire. Samples are calibrated after adjusting for non-response, at individual level, for metropolitan France and separately for overseas France. The data used in calibration are those from the Labour Force Survey (LFS) from year N-1.

For metropolitan France, the calibration variables selected are the following: cross-tabulations for sex-age (14 levels), sex-qualification (nine levels), age-qualification (seven levels), Social categories (11 levels), size of urban area crossed with the category of municipality in urban zones (eight levels), size of urban area (eight levels), new region (12 levels), number of persons aged 15 or over in the household crossed with the three main age groups (nine levels), household type (based on Eurostat definition - four levels) and nationality (two levels).

For overseas France, the calibration variables selected are the crossing of sex and age for all overseas departments and territories (nine levels), the qualification for all overseas departments and territories (four levels), the Social categories (nine levels), population distribution by sub-departmental geographic areas (15 levels in total) and the number of households.

5.3 National Survey on Youth Resources (ENRJ)

(This paragraph is an extract from an internal INSEE note by Gros [6]).

The ENRJ survey was conducted by DREES and INSEE between 1 October and 31 December 2014. It describes the resources and living standards of young adults aged between 18 and 24 in France, living in ordinary housing or community accommodation.

The only reliable and accurate source of data available to INSEE in this specific topic is the age pyramid by sex dated 1 January 2015 based on the 2014 demographics report (Census). Figures collected using administrative sources consist more of calibration data than calibration margins per se: possibility of double-counting for some variables, concepts that differ slightly between administrative sources and variables taken from survey for others. It was therefore decided to use penalised calibration, as follows:

- exact calibration for the age pyramid by sex as of 1 January 2015;
- approximate calibration for a certain number of administrative datasets: number of people with a Bachelor’s degree by sex, number of people with a Bachelor’s degree by field of study, number of people enrolled in university by sex, number of people enrolled for the higher technical certificate (BTS) by sex, number of people enrolled and business schools or engineering schools, number or bursaries based on social criteria by sex, number of single people receiving housing benefit (APL) and the number of young women receiving a basic allowance under the Early Childhood Benefit Programme (PAJE).

The penalised calibration makes it possible to contain the weight ratios within the [0.2; 1.8] interval and to ensure exact calibration on the age pyramid while adjusting for the main imbalances that were observed in the calibration data when the age pyramid by sex is adjusted solely by post-stratification.

REFERENCES

[1] Beaumont, J.-F and Bocci, C.(2008), "Another look at ridge calibration", *Metron*, vol 66.1, pp. 5-20.

[2] Deming, W.E. and Stephan, F.F. (1940), "On a least squares adjustment of a sampled frequency table when the exact totals are known", *Annals of Mathematical Statistics*, 11, pp. 427-444.

[3] Deville, J.-C. and Särndal, C.-E (1992), "Calibration estimation in survey sampling", *Journal of the American Statistical Association,* 87, n°418, pp. 375-382.

[4] Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993), "Generalized raking procedures in survey sampling", *Journal of the American Statistical Association,* 88, n°423, pp. 1013-1020.

[5] Dupont, F. (1996), « Calage et redressement de la non-réponse totale », *Actes des journées de méthodologie statistique, 15 et 16 décembre 1993, INSEE-Méthodes n°56-57-58.*

[6] Gros, E. (2015), « Note de bilan relative aux calculs de pondérations dans l'enquête nationale sur les ressources des jeunes », note interne Insee.

[7] Guggemos, F. and Tillé (2010), Y., "Penalized calibration in survey sampling : Design based estimation assisted by mixed models", *Journal of statistical planning and inference*, 140.11 pp. 3199-3212.

[8] Rebecq, A. (2016), « Icarus : un package R pour le calage sur marges et ses variantes », *9e colloque francophone sur les sondages, Gatineau (Canada)* - http://sondages2016.sfds.asso.fr/.

[9] Rebecq, A. (2016), « Le calage pénalisé - Théorie et application », *Séminaire de méthodologie statistique de l'Insee « Miscellanées sur la calage »,* https://www.insee.fr/fr/information/2387498

[10] Sautory, O. (1993), « La macro Calmar. Redressement d'un échantillon par calage sur marges », *Document de travail F9310 de la DSDS*, Insee.

Department of statistical methods
Version 1, published 5 March 2018