Use of Calibration in Official Statistics: Design versus Model-Based Calibration – Pros and Cons

Marcin Szymkowiak

Statistical Office in Poznan
Poznan University of Economics and Business

Outline

1. Theoretical background of calibration - the design-based approach
2. Theoretical background of calibration - the model-based approach
3. Simulation study
4. Design versus Model-Based Calibration – Pros and Cons

Design-based approach in calibration

Design-based approach in calibration

1. This technique was proposed by Devill and Särndal (1992) and is a method of searching for so called calibrated weights by minimizing a distance measure between sampling weights and new weights, which satisfies certain calibration constraints.
2. As a consequence, when new weights are applied to auxiliary variables in the sample, they corectly reproduce known population totals of the auxiliary variables.
3. It is also important that the new weights should be as close as possible to sampling weights in the sense of the selected distance measure (Särndal C-E., Lundström S. 2005, Särndal C-E. 2007).

Theoretical background of calibration

Theoretical background of calibration

- Let us assume that the whole population $U = \{1, 2, \ldots, N\}$ consists of N elements.
- From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements.
- Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = 1/\pi_i$ the design weight.
- Let us assume that our main goal is to estimate the total value of variable y:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

where y_i denotes the value of variable y for the i-th unit, $i = 1, \ldots, N$.

Theoretical background of calibration

Theoretical background of calibration

- Let $x_1, \ldots, x_k$ denote auxiliary variables which will be used in the process of finding calibration weights and let $\mathbf{X}_j$ denote the total value for auxiliary variable x_j, $j = 1, \ldots, k$, e.i.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{2}$$

where x_{ij} denotes the value of j-th auxiliary variable for the i-th unit.

- In practice it occurs that:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j \tag{3}$$

so calibration is required.

Theoretical background of calibration

Theoretical background of calibration

- Let $\mathbf{w} = (w_1, \ldots, w_n)^T$ denote a vector of calibration weights.
- Our main goal is to look for new weights w_i, which are as close as possible to design weights d_i and which allow us to get correct known population totals from administrative registers.
- The process of constructing calibration weights consists in properly selecting a distance function.
- Let G denote a function for which the second derivative exists and:
  - $G(\cdot) \geq 0$,
  - $G(1) = 0$,
  - $G'(1) = 0$,
  - $G''(1) = 1$.

Examples of G function

Examples of G function

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{4}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{5}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{6}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{7}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt. \tag{8}$$

The choice of G function

The choice of G function

- The most common G function which can be used in the process of constructing a distance function is $G_1(x) = \frac{1}{2}(x-1)^2$. In this case we have:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \quad (9)$$

The problem of finding calibration weights

The problem of finding calibration weights

(C1) Find the minimum of the distance function:

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{10}$$

(C2) Calibration equations:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{11}$$

(C3) Calibration constraints:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{where: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, n. \tag{12}$$

The calibration estimator for total

The calibration estimator for total

The calibration estimator for totals takes the form:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i \,, \tag{13}$$

where the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ is obtained as the following minimization problem:

$$\mathbf{w} = \operatorname{argmin}_v D(\mathbf{v}, \mathbf{d}) \,, \tag{14}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{15}$$

where

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i}, \tag{16}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{17}$$

Theorem

Theorem

The solution of the minimization problem is a vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, for which

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{18}$$

where

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \tag{19}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{20}$$

Model-based approach in calibration

Model-based approach in calibration

1. This technique was proposed by Wu and Sitter (2001) and is a method of searching for so called calibrated weights by minimizing a distance measure between sampling weights and new weights, which satisfies certain calibration constraints. In this approach we assume that the relationship between y and $\mathbf{x}$ can be described by a proper statistical model.
2. As a consequence, when the new weights are applied to predicted values of y variable in the sample, they correctly reproduce the known population totals of the predicted values of y variable.
3. It is also important that the new weights should be as close as possible to the sampling weights in the sense of the selected distance measure.

Model-based approach in calibration

Model-based approach in calibration

- In model-based calibration Wu and Sitter (2001) assume that the relationship between y and $\mathbf{x}$ can be described by a superpopulation model through the first and second moments:

$$\begin{cases} E_U(y_i) = f(\mathbf{x}_i, \beta) \\ D_U^2(y_i) = v_i^2 \sigma^2 \end{cases}, \tag{21}$$

  where $f(\mathbf{x}_i, \beta)$ is a known function of x and β, $\beta = (\beta_0, \ldots, \beta_k)^T$ and σ^2 are unknown superpopulation parameters, which have to be estimated and v_i is a known function of $\mathbf{x}_i$. E_U and D_U^2 denote the expectation and variance with respect to the superpopulation model. In this approach we also assume that $(y_1, \mathbf{x}_1), \ldots, (y_k, \mathbf{x}_k)$ are mutually independent.
- The model described by Wu and Sitter is very general and includes both a linear and nonlinear regression model.

Model-based approach in calibration

Model-based approach in calibration

- We assume that the main goal is to estimate the total value of variable y given by (1). Moreover, we assume that auxiliary variables $x_1, \ldots, x_k$ exist and $f(\mathbf{x}_i, \beta)$ is the linking model between y and auxiliary variables. Using data from sample s and all auxiliary variables $x_1, \ldots, x_k$ we find predicted values

$$\hat{y}_i = f\left(\mathbf{x}_i, \hat{\beta}\right), \tag{22}$$

where the estimator of parameter β is given by:

$$\hat{\beta} = \left(\mathbf{X}_s^T \boldsymbol{\Pi}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}^{-1} \mathbf{y}_s, \tag{23}$$

$\boldsymbol{\Pi}$ is a diagonal matrix consisting of first order inclusion probabilities π_i:

$$\boldsymbol{\Pi} = diag(\pi_1, \ldots, \pi_n) = \begin{bmatrix} \pi_1 & 0 & \ldots & 0 \\ 0 & \pi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \pi_n \end{bmatrix}. \tag{24}$$

The problem of finding model-based calibration weights

The problem of finding model-based calibration weights

Taking into account distance function D given by (9) the problem of finding calibration weights in the model-based approach can be formulated as follows:

$$\begin{cases} D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i} \rightarrow min \\ \sum_{i \in s} w_i \hat{y}_i = \sum_{i \in U} \hat{y}_i = \sum_{i \in U} f\left(\mathbf{x}_i, \hat{\beta}\right) \\ \sum_{i \in s} w_i = N \end{cases} \tag{25}$$

The problem of finding model-based calibration weights

The problem of finding model-based calibration weights

Taking into account distance function D given by (9) the problem of finding calibration weights in the model-based approach can be formulated equivalently as follows:

$$\begin{cases} D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i} \rightarrow min \\ \sum_{i \in s} w_i \mathbf{z}_i = \sum_{i \in U} \mathbf{z}_i = \mathbf{z}_U \end{cases}, \tag{26}$$

where:

$$\mathbf{z}_i = (\hat{y}_i, 1)^T \tag{27}$$

and

$$\mathbf{z}_U = \left(\sum_{i \in U} \hat{y}_i, N \right)^T. \tag{28}$$

The model-based calibration estimator for total

The model-based calibration estimator for total

The model-based calibration estimator for totals takes the form:

$$\hat{Y}_{mcal} = \sum_{i \in s} w_i y_i, \tag{29}$$

where the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ is obtained as the following minimization problem:

$$\mathbf{w} = \text{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{30}$$

under the constraint:

$$\sum_{i \in s} w_i \mathbf{z}_i = \mathbf{z}_U. \tag{31}$$

Theorem

Theorem

The solution of the minimization problem is a vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, for which

$$w_i = d_i + d_i \left(\mathbf{z}_U - \hat{\mathbf{z}}\right)^T \left(\sum_{i \in s} d_i \mathbf{z}_i \mathbf{z}_i^T\right)^{-1} \mathbf{z}_i, \tag{32}$$

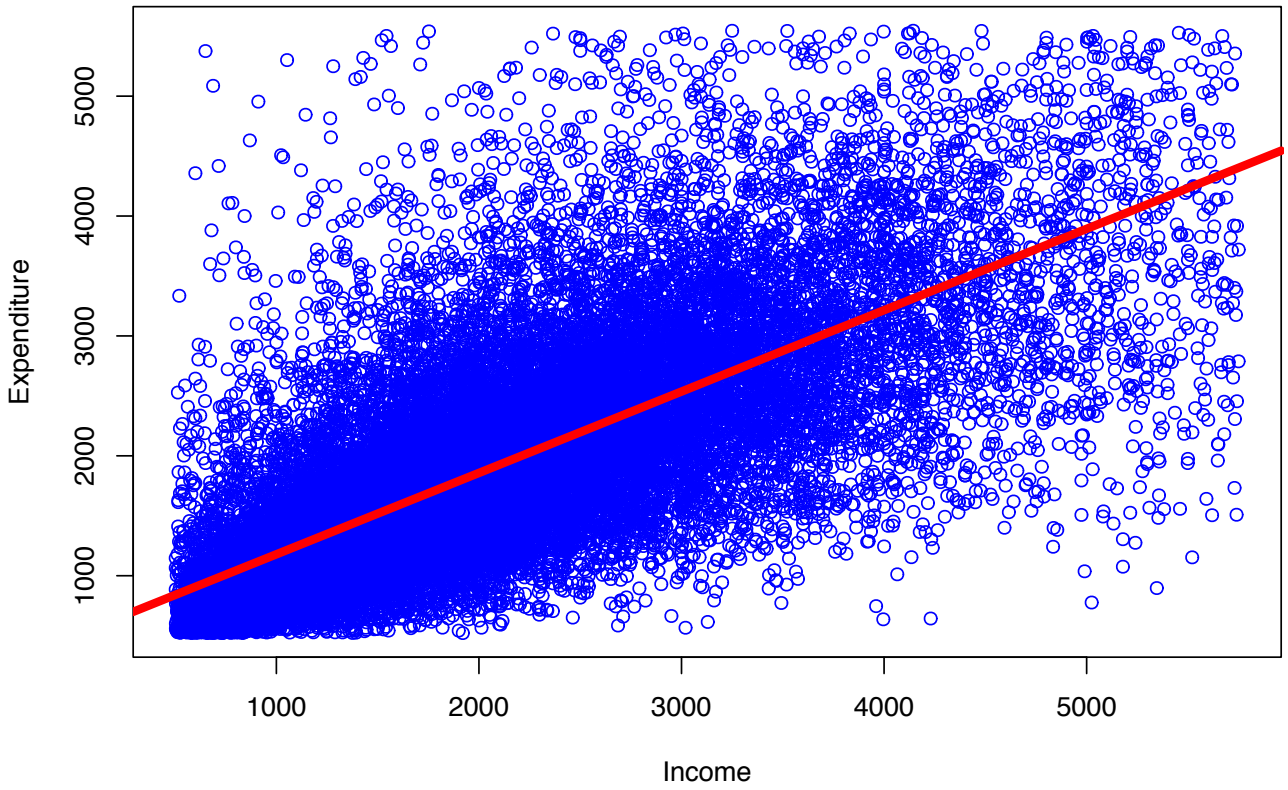
where:

$$\hat{\mathbf{z}} = \left(\sum_{i \in s} d_i \hat{y}_i, \sum_{i \in s} d_i\right)^T. \tag{33}$$

Population under study

- The simulation study investigated two variables: total expenditures of households or output variable (y) and the household's income (x), which was the only auxiliary variable.
- Data about variables came from the Polish Household Budget Survey (HBS).
- To conduct the simulation study, a pseudo-population was created, consisting of all households included in the HBS survey for which information about the variables was available.
- The resulting dataset consisted of about 30,000 records containing complete information about the variables of interest.
- Coefficient of correlation between y and x variable was 0.74.

Population under study



Estimators selected for the study

Selected estimators

Six different estimators were taken into account:

- Horvitz-Thompson estimator (HT),
- Design-based calibration estimator with $G_3\,(x)$ function (Logit),
- Design-based calibration estimator with $G_1\,(x)$ function (GREG),
- Design-based calibration estimator with $G_5\,(x)$ function (Sinh),
- Model-based calibration estimator (MC_1) (linear model in the sample $y = a_0 + a_1 x$),
- Model-based calibration estimator (MC_2) (logarithmic model in the sample $y = a_0 + a_1 \ln x$),
- Model-based calibration estimator (MC_3) (power model in the sample $y = a_0 x^{a1}$),
- $k = 500$ replications were done using simple random sampling without replacement. The sample size: $n_1 = 500$ and $n_2 = 1000$.

Evaluation of estimators

Evaluation of estimators

Evaluation of estimators taken into account in the Monte Carlo simulation study was based on three measures:

- The relative bias of estimators (RB):

$$RB = \frac{E\left(\hat{T}\right) - T_y}{T_y} \tag{34}$$

- The coefficient of variation (CV):

$$CV = \frac{\sqrt{V}}{E\left(\hat{T}\right)} \tag{35}$$

where $E\left(\hat{T}\right) = \frac{1}{k}\sum_{i=1}^{k}\hat{T}_i$ and $V = \frac{1}{k}\sum_{i=1}^{k}\left(\hat{T}_i - E\left(\hat{T}\right)\right)^2$.

- Mean square error (MSE).

Results

	$n = 500$			$n = 1000$		
Estimator	RB	CV	MSE	RB	CV	MSE
HT	-0.12	2.19	122260917	-0.02	1.58	87987455
Logit	-0.01	1.64	91670118	0.04	1.15	64443178
GREG	-0.01	1.64	91670531	0.04	1.15	64449652
Sinh	-0.01	1.64	91674483	0.04	1.15	64444523
MC_1	-0.01	1.59	88816622	0.03	1.09	61058534
MC_2	-0.04	1.61	89710573	0.04	1.11	61968056
MC_3	-0.04	1.58	88078254	0.03	1.09	60682219

Results

- The Relative Bias (RB) of all estimators is within a reasonable range, with the HT having the largest at -0.12%.
- The Coefficient of Variation (CV) is higher for the HT estimator than for design and model-based calibration estimators; it can be seen especially when the sample size is smaller.
- The gain in precision from using the model-based calibration estimators compared to the design-based calibration estimators is visible but rather small. This is the consequence of the strong relationship between y and x variables.
- The results obtained in the simulation study show that methods that are conceptually different, can lead to similar final results.

Design versus Model-Based Calibration – Pros and Cons

	Calibration approach	
	Design-based calibration	**Model-based calibration**
Purpose	Reproducion of known population totals of all auxiliary variables	Reproduction of population totals of the predictions using a properly selected model
Popularity	Common in official statistics	The application of model-based estimation procedures in official statistics is limited
Aggregate data	Aggregate data as auxiliary variables are sufficient at the population level	Unit data as auxiliary variables are necessary at the population level
Model	Model-free - no model specification is required	A properly selected model of the relationship between y and all auxiliary variables is required
Consistency	The weight system is consistent with the known population total for each auxiliary variable	The weight system may not be consistent with the known population total for each auxiliary variable
Multipurpose weighting	Calibration weights do not depend on y values - coherence of estimates with published statistics is possible (multipurpose weighting)	Calibration weights depend on the y values, implying a loss of the multipurpose property

Design versus Model-Based Calibration – Pros and Cons

	Calibration approach	
	Design-based calibration	**Model-based calibration**
Robustness	Robust against model-misspecification	Not robust against model-misspecification
Precision and accuracy	It requires strong auxiliary variables	It improves the precision and accuracy when the model is well specified
Timeliness	This form of calibration is an attractive option for producing timely official statistics (one set of weights for the estimation of all target parameters)	The need to build different models for different target parameters would negatively affect the timely production of official statistics (more than one set of weights for the estimation of all target parameters)

Literature

Literature


Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.


Brakel van den J., Bethlehem J. (2008), „*Model-Based Estimation for Official Statistics*, Statistics Netherlands, Voorburg/Heerlen.


Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.


Lehtonen R., Särndal C-E., Veijanen A .(2008), „*Generalized regression and model-calibration estimation for domains: Accuracy comparison*.


Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.


Wu C., Sitter R.R. (2001), „*A model-calibration approach to using complete auxiliary information from survey data*, Journal of the American Statistical Association, 96, 185—193.

Thank you very much for your attention!