Joint calibration estimators for totals and quantiles for probability and nonprobability samples

Maciej Beręsewicz[1,2], Marcin Szymkowiak[1,2]

[1]Poznań University of Economics and Business
[2]Statistical Office in Poznań

uRos 2024

29.11.2024

1. Introduction
2. Aim of the presentation
3. Calibration for total and quantile
4. `jointCalib` package
5. Examples
6. Literature

Introduction

- The authors' work has been financed by the National Science Centre in Poland, OPUS 22, grant no. 2020/39/B/HS4/00941.
- Detailed description can be found in two our working papers:
  - **A note on joint calibration estimators for totals and quantiles** (https://arxiv.org/abs/2308.13281)
  - **Quantile balancing inverse probability weighting for non-probability samples** (https://arxiv.org/abs/2403.09726; Minor Review at the Survey Methodology journal).
- Codes to reproduce the results are freely available from the github repository: https://github.com/ncn-foreigners/.
- R packages: **jointCalib** for joint calibration for totals and quantiles and **nonprobsvy** for non-probability samples. Both available through **CRAN**.
- The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Poland.

Introduction

- In this presentation we consider a method of joint calibration for totals (**Deville and Särndal 1992**) and quantiles (**Harms and Duchesne, 2006**).
- The proposed method is based on the classic approach to calibration and simultaneously takes into account calibration equations for totals and quantiles of all auxiliary variables.
- Final calibration weights w_k reproduce known population totals and quantiles for all auxiliary variables.
- At the same time, they help to reduce the bias and improve the precision of estimates.

Contribution

- We **extend** the calibration/IPW paradigm to **jointly account for totals/means and quantiles in probability and non-probability samples**.
- We propose a new package `jointCalib` which allows co create calibration weights to reproduce population totals and population quantiles for a set of auxiliary variables,
- The proposed approach allows the same vector of weights (calibration weights) to be used in the estimation of totals and quantiles for variables under study.
- The package implements calibration through `sampling`, `laeken` and `survey` packages as well as entropy balancing (via the `ebal` package) and empirical likelihood (using base R).

Setup (1)

- Let $U = \{1, \ldots, N\}$ denote the target population consisting of N labelled units.
- Each unit k has an associated vector of auxiliary variables $\boldsymbol{x}$ and the target variable y, with their corresponding values $\boldsymbol{x}_k$ and y_k, respectively.
- s denotes a probability sample of size n.
- $d_k = 1/\pi_k$ is a design weight and π_k is the first-order inclusion probability of the i-th element of the population U.

Calibration approach

- In most applications the goal is to estimate a finite population total

$$\tau_y = \sum_{k \in U} y_k \tag{1}$$

or the mean

$$\bar{\tau}_y = \tau_y / N \tag{2}$$

of the variable of interest y, where U is the population of size N.

- The well-known estimator of a finite population total is the Horvitz-Thompson estimator

$$\hat{\tau}_{y\pi} = \sum_{k \in s} d_k y_k. \tag{3}$$

- In most cases original weights d_k do not reproduce know population totals for auxiliary variables. They have to be calibrated.

Calibration approach for total

- Let $\boldsymbol{x}_k^\circ$ be a J_1-dimensional vector of auxiliary variables for which

$$\tau_{\boldsymbol{x}} = \sum_{k \in U} \boldsymbol{x}_k^\circ = \left(\sum_{k \in U} x_{k1}, \ldots, \sum_{k \in U} x_{kJ_1} \right)^T \qquad (4)$$

  is assumed to be known.
- In most cases in practice the d_k weights do not reproduce known population totals for auxiliary variables $\boldsymbol{x}_k^\circ$.
- It means that the resulting estimate $\hat{\tau}_{\boldsymbol{x}\pi} = \sum_{k \in s} d_k \boldsymbol{x}_k^\circ$ is not equal to $\tau_{\boldsymbol{x}}$.

Calibration approach for total

- The main idea of calibration is to look for new calibration weights w_k which are as close as possible to original design weights d_k and reproduce known population totals $\tau_{\boldsymbol{x}}$ exactly.
- In other words, in order to find new calibration weights w_k we have to minimise a distance function

$$D\left(\boldsymbol{d},\boldsymbol{v}\right)=\sum_{k\in s} d_k\; G\left(\frac{v_k}{d_k}\right)\rightarrow \min \tag{5}$$

to fulfil calibration equations

$$\sum_{k\in s} v_k \boldsymbol{x}_k^{\circ}=\sum_{k\in U} \boldsymbol{x}_k^{\circ}, \tag{6}$$

where $\boldsymbol{d}=\left(d_1,\ldots,d_n\right)^T$, $\boldsymbol{v}=\left(v_1,\ldots,v_n\right)^T$ and $G\left(\cdot\right)$ is a function which must satisfy some regularity conditions.

Calibration approach for total

- The final calibration estimator of a population total τ_y can be expressed as

$$\hat{\tau}_{y\boldsymbol{x}} = \sum_{k \in s} w_k y_k, \tag{7}$$

where w_k are calibration weights obtained for instance for $G(x) = \frac{(x-1)^2}{2}$ as follows:

$$w_k = d_k + d_k \left(\tau_{\boldsymbol{x}} - \hat{\tau}_{\boldsymbol{x}\pi}\right)^T \left(\sum_{j \in s} d_j \boldsymbol{x}_j^{\circ} \boldsymbol{x}_j^{\circ T}\right)^{-1} \boldsymbol{x}_k^{\circ}.$$

Calibration approach for quantile

- We assume that
$$\boldsymbol{Q}_{\boldsymbol{x},\alpha} = \left(Q_{x_1,\alpha}, \ldots, Q_{x_{J_2},\alpha}\right)^T \tag{8}$$
is a vector of known population quantiles of order α for a vector of auxiliary variables $\boldsymbol{x}_k^*$, where $\alpha \in (0,1)$ and $\boldsymbol{x}_k^*$ is a J_2-dimensional vector of auxiliary variables.
- It is worth noting that, in general, the numbers J_1 and J_2 of the auxiliary variables are different.
- It may happen that for a specific auxiliary variable its population total and the corresponding quantile of order α will be known. However, in most cases quantiles will be known for continuous auxiliary variables, unlike totals, which will be generally known for categorical variables.

Calibration approach for quantile

- A calibration estimator of quantile $Q_{y,\alpha}$ of order α for variable y is defined as

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{9}$$

where a vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ is a solution of optimization problem

$$D(\boldsymbol{d}, \boldsymbol{v}) = \sum_{k \in s} d_k \, G\left(\frac{v_k}{d_k}\right) \to \min \tag{10}$$

subject to the calibration constraints

$$\sum_{k \in s} v_k = N \tag{11}$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{x},cal,\alpha} = \left(\hat{Q}_{x_1,cal,\alpha}, \ldots, \hat{Q}_{x_{J_2},cal,\alpha}\right)^T = \boldsymbol{Q}_{\boldsymbol{x},\alpha}, \tag{12}$$

where $j = 1, \ldots, J_2$.

Calibration approach for quantile

- If $G(x) = \frac{(x-1)^2}{2}$ then using the method of Lagrange multipliers the final calibration weights w_k can be expressed as

$$w_k = d_k + d_k \left(\boldsymbol{T_a} - \sum_{k \in s} d_k \boldsymbol{a}_k \right)^T \left(\sum_{j \in s} d_j \boldsymbol{a}_j \boldsymbol{a}_j^T \right)^{-1} \boldsymbol{a}_k, \qquad (13)$$

where $\boldsymbol{T_a} = (N, \alpha, \ldots, \alpha)^T$ and the elements of the vector $\boldsymbol{a}_k = (1, a_{k1}, \ldots, a_{kJ_2})^T$ are given by

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \leq L_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ N^{-1}\beta_{x_j,s}\left(Q_{x_j,\alpha}\right), & x_{kj} = U_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ 0, & x_{kj} > U_{x_j,s}\left(Q_{x_j,\alpha}\right), \end{cases} \qquad (14)$$

with $j = 1, \ldots, J_2$.

A joint approach to calibration for total and quantile

- Let us assume that we are interested in estimating a population total τ_y and/or quantile $Q_{y,\alpha}$ of order α, where $\alpha \in (0,1)$ for variable of interest y.
- Let $\boldsymbol{x}_k = \begin{pmatrix} \boldsymbol{x}_k^{\circ} \\ 1 \\ \boldsymbol{x}_k^{*} \end{pmatrix}$ be a $J+1$-dimensional vector of auxiliary variables, where $J = J_1 + J_2$.
- We assume that for J_1 variables a vector of population totals $\tau_{\boldsymbol{x}}$ is known and for J_2 variables a vector $\boldsymbol{Q}_{\boldsymbol{x},\alpha}$ of population quantiles is known.
- In practice it may happen that for the same auxiliary variable we know its population total and quantile.
- We do not require that the complete auxiliary information described by the vector $\boldsymbol{x}_k$ is known for all $k \in U$.

A joint approach to calibration for total and quantile

- In our joint approach we are looking for a vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ which is a solution of the optimization problem

$$D(\boldsymbol{d}, \boldsymbol{v}) = \sum_{k \in s} d_k \, G\left(\frac{v_k}{d_k}\right) \to \min \tag{15}$$

subject to the calibration constraints

$$\sum_{k \in s} v_k = N, \tag{16}$$

$$\sum_{k \in s} v_k \boldsymbol{x}_k^{\circ} = \tau_{\boldsymbol{x}}, \tag{17}$$

$$\hat{\boldsymbol{Q}}_{\boldsymbol{x},cal,\alpha} = \boldsymbol{Q}_{\boldsymbol{x},\alpha}. \tag{18}$$

A joint approach to calibration for total and quantile

- Alternatively, the last calibration constraint can be expressed as

$$\sum_{k \in s} v_k \boldsymbol{a}_k = \boldsymbol{T}_{\boldsymbol{a}}, \tag{19}$$

where as previously $\boldsymbol{T}_{\boldsymbol{a}} = (N, \alpha, \ldots, \alpha)^T$ and the elements of the vector $\boldsymbol{a}_k = (1, a_{k1}, \ldots, a_{kJ_2})^T$ are given by

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \leq L_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ N^{-1}\beta_{x_j,s}\left(Q_{x_j,\alpha}\right), & x_{kj} = U_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ 0, & x_{kj} > U_{x_j,s}\left(Q_{x_j,\alpha}\right), \end{cases} \tag{20}$$

with $j = 1, \ldots, J_2$.

A joint approach to calibration for total and quantile

- Assuming $G(x) = \frac{(x-1)^2}{2}$ function, an explicit solution of the above optimization problem can be derived.
- Let $\boldsymbol{h}_{\boldsymbol{x}} = \binom{T_{\boldsymbol{x}}}{\boldsymbol{T}_{\boldsymbol{a}}}$ and $\hat{\boldsymbol{h}}_{\boldsymbol{x}} = \binom{\sum_{k \in s} d_k \boldsymbol{x}_k^{\circ}}{\sum_{k \in s} d_k \boldsymbol{a}_k}$.
- Then the vector of calibration weights $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ which solves the above optimization problem satisfies the relation:

$$w_k = d_k + d_k \left(\boldsymbol{h}_{\boldsymbol{x}} - \hat{\boldsymbol{h}}_{\boldsymbol{x}} \right)^T \left(\sum_{j \in s} d_j \boldsymbol{x}_j \boldsymbol{x}_j^T \right)^{-1} \boldsymbol{x}_k. \quad (21)$$

A joint approach to calibration for total and quantile

- Under this function, the calibration estimator using (21) is equivalent to a generalised linear regression estimator (GREG) given by

$$\hat{\tau}_{y\boldsymbol{x}}^{GREG} = \sum_{k \in S_A} d_k^A y_k + \left(\boldsymbol{h}_x - \hat{\boldsymbol{h}}_x \right)^T \hat{\boldsymbol{\beta}},$$

- Therefore, we assume that the relationship between auxiliary variables $\boldsymbol{x}_k^\circ$ and $\boldsymbol{x}_k^*$ through $\boldsymbol{a}_k$ is linear as in

$$\hat{y}_k = (\boldsymbol{x}_k^\circ)^T \hat{\boldsymbol{\beta}}^\circ + \boldsymbol{a}_k^T \hat{\boldsymbol{\beta}}^*. \quad (22)$$

jointCalib – overview

jointCalib 0.1.2 Reference Articles Changelog Search fo

Overview

Details

A small package for joint calibration of totals and quantiles (see Beręsewicz and Szymkowiak (2023) working paper for details). The package combines the following approaches:

- Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.
- Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. Survey Methodology, 32(1), 37.
- Wu, C. (2005) Algorithms and R codes for the pseudo empirical likelihood method in survey sampling, Survey Methodology, 31(2), 239.
- Zhang, S., Han, P., and Wu, C. (2023) Calibration Techniques Encompassing Survey Sampling, Missing Data Analysis and Causal Inference, International Statistical Review 91, 165–192.

which allows to calibrate weights to known (or estimated) totals and quantiles jointly. As an backend for calibration sampling (`sampling::calib`), laeken (`laeken::calibWeights`), survey (`survey::grake`) or ebal (`ebal::eb`) package can be used. One can also apply empirical likelihood using codes from Wu (2005) with support of `stats::constrOptim` as used in Zhang, Han and Wu (2022).

Links

View on CRAN

Browse source code

Report a bug

License

GPL-3

Citation

Citing jointCalib

Developers

Maciej Beręsewicz

Author, maintainer

Dev status

R-CMD-check passing

CRAN 0.1.0

DOI 10.5281/zenodo.8355993

mentioned in awesome

jointCalib – the main function

Usage

```
joint_calib(
  formula_totals = NULL,
  formula_quantiles = NULL,
  data = NULL,
  dweights = NULL,
  N = NULL,
  pop_totals = NULL,
  pop_quantiles = NULL,
  subset = NULL,
  backend = c("sampling", "laeken", "survey", "ebal", "base"),
  method = c("raking", "linear", "logit", "sinh", "truncated", "el", "eb"),
  bounds = c(0, 10),
  maxit = 50,
  tol = 1e-08,
  eps = .Machine$double.eps,
  control = control_calib(),
  ...
)
```

jointCalib – the main function

formula_totals

a formula with variables to calibrate the totals,

formula_quantiles

a formula with variables for quantile calibration,

data

a data.frame with variables,

dweights

initial d-weights for calibration (e.g. design weights),

N

population size for calibration of quantiles,

pop_totals

a named vector of population totals for `formula_totals`. Should be provided exactly as in `survey` package (see `survey::calibrate`),

pop_quantiles

a named list of population quantiles for `formula_quantiles` or an `newsvyquantile` class object (from `survey::svyquantile` function),

An example

```
set.seed(123)
N <- 1000
x <- runif(N, 0, 80)
y <- exp(-0.1 + 0.1*x) + rnorm(N, 0, 300)
p <- rbinom(N, 1, prob = exp(-0.2 - 0.014*x))
df <- data.frame(x, y, p)
df_resp <- df[df$p == 1, ]
df_resp$d <- N/nrow(df_resp)
```

An example – known quantiles and totals

```
## information about population quantiles and totals
probs <- seq(0.1, 0.9, 0.1)
y_quant_true <- quantile(y, probs)
quants_known <- list(x=quantile(x, probs))
totals_known <- c(x=sum(x))

## standard calibration
result0 <- sampling::calib(Xs = cbind(1, df_resp$x),
                           d = df_resp$d,
                           total = c(N, totals_known),
                           method = "linear")
```

An example – calibration of totals and quantiles

```
result <- joint_calib(formula_totals = ~x,
                      formula_quantiles = ~x,
                      data = df_resp,
                      dweights = df_resp$d,
                      N = N,
                      pop_quantiles = quants_known,
                      pop_totals = totals_known,
                      method = "linear",
                      backend = "sampling")
```

An example – results

```
> result
Weights calibrated using: linear with sampling backend.
Summary statistics for g-weights:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.244   1.431   1.888   2.037   2.378   4.172
Totals and precision (abs diff: 8.409491e-08)
           totals     precision
N        1000.00 -2.141746e-09
x 0.10      0.10 -2.887413e-13
x 0.20      0.20 -6.920020e-13
x 0.30      0.30 -7.922552e-13
x 0.40      0.40 -1.168732e-12
x 0.50      0.50 -1.096512e-12
x 0.60      0.60 -1.642686e-12
x 0.70      0.70 -1.754152e-12
x 0.80      0.80 -1.792344e-12
x 0.90      0.90 -2.097322e-12
x       39782.22 -8.194183e-08
```

An example – comparison of estimates of τ quantiles of Y

```
> data.frame(total = y_quant_hat0,
+            totals_and_quant =y_quant_hat1,
+            true=y_quant_true)

         total totals_and_quant       true
10% -284.3574       -285.34675 -292.97255
20% -131.7079       -131.70792 -128.19010
30%  -25.2815        -21.94192  -10.07312
40%   80.5919         84.23786   84.64057
50%  175.5490        178.96015  184.87445
60%  274.0404        279.73343  294.76788
70%  412.2826        426.98679  453.35435
80%  592.0840        606.73082  669.36570
90% 1105.6883       1172.38891 1163.92646
```

```
> data.frame(total,
+            totals_and_quant)
     total     totals_and_quant
1  109.0958            71.85097
```

Summary

- The approach that extends calibration to simultaneously account for:
  - Known population totals for auxiliary variables
  - Known population quantiles for auxiliary variables
- Final calibration weights w_k reproduce both totals and quantiles
- Implementation available in `jointCalib` R package:
  - Supports multiple calibration methods (linear, raking, entropy)
  - Integrates with existing R packages (`sampling`, `survey`)
  - Allows flexible specification of totals and quantiles
- We encourage you observe our organization at Github (**github.com/ncn-foreigners**) and the repo for the package (**ncn-foreigners/jointCalib**).

Literature

- Chen Y., Li P., Wu C. (2020), „*Doubly robust inference with nonprobability survey samples*", Journal of the American Statistical Association, 115 (532), 2011–2021.
- Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.
- Harms T., Duchesne, P. (2006), „*On calibration estimation for quantiles*", Survey Methodology, 32(1), 37.
- Kott P.S., Chang T. (2010), „*Using calibration weighting to adjust for nonignorable unit nonresponse*", Journal of the American Statistical Association, 105 (491), 1265–1275.
- Wu C., Thompson M.E. (2020), „*Sampling theory and practice*", Springer.

Thank you for your attention!