Robust calibration estimators in surveys with outliers

Marcin Szymkowiak

Statistical Office in Poznań
University of Economics in Poznań

Outline

Theoretical background of calibration

Theoretical background of calibration

1. This technique was proposed by Devill and Särndal (1992) and is a method of searching for so called calibrated weights by minimizing distance measure between the sampling weights and the new weights, which satisfy certain calibration constraints.
2. As a consequence when the new weights are applied to the auxiliary variables in the sample, they reproduce the known population totals of the auxiliary variables exactly.
3. It is also important that the new weights should be as close as possible to sampling weights in sense of chosen 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 estimation of the total value of the variable y:

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

where y_i denotes the value of the variable y for 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 the auxiliary variable x_j, $j = 1, \ldots, k$, e.i.

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

  where x_{ij} odenotes 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 the vector of calibration weights.
- Our main goal is to look for new weights w_i which are as close as possible to the design weights d_i and which allow us to get known population totals from administrative registers exactly.
- The process of construction calibration weights depends on the properly chosen distance function.
- Let G denote 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 construction 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 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 total 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:

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

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

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 the 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

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

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

Statistical software

Statistical software

- **Bascula 4.0** - the statistical tool developed in the Delphi language by Statistics Netherlands for the calculation of estimates of population totals, means and ratios.
- **Calmar/Calmar 2** - the statistical software developed by INSEE.
- **Caljack** - this is a SAS macro written and developed by Statistics Canada and is an extension of the Calmar macro.
- **CALWGT** - this is a freely distributed program for calibration written by Li-Chun Zhang in S-plus for Unix.
- **CLAN 97** - the statistical software designed to handle surveys in Statistics Sweden.
- **G-Calib 2** - the statistical software developed in the SPSS language by Statistics Belgium.
- **GES** - this is a SAS-based application with a Windows-like interface which was developed in SAS/AF by Statistics Canada.
- **R** - this is a free statistical software. The calibrate function, which can be found in the survey package, reweights the survey design weights and also adds additional information about estimated standard errors.

CALMAR

CALMAR

- Although in many statistical packages the problem of finding calibration weights was implemented using different G functions in Poland CALMAR is preferred.
- In CALMAR, which is a macro written in 4GL in SAS four distance functions were implemented: the linear method, the raking ratio metod, the logit method, the truncated linear method.
- In CALMAR 2 which is a later version of CALMAR, the distance function based on hyperbolic sinus function was also implemented.

Example 1

Example 1

- We consider an artificial population of enterprises of size $N = 1000$ from which a simple random sample of size $n = 20$ is drawn. Hence design (initial) weights are equal $N/n = 1000/20 = 50$.
- We also consider a numerical variable x_1 (for instance monthly revenue of enterprise) and one categorical variable x_2 (for instance enterprise size i.e. large - L and medium - M).
- In this example it will be only shown how to compute calibration weights. We do not take into account the variable of interest y which is not necessary to compute calibration weights and would be necessary to calculate the variance of the estimator.

Example 1 – artificial data set

Example 1 – artificial data set

Number of enterprise	Monthly revenue x_1	Enterprise size x_2	d_i
1	18	M	50
2	14	M	50
3	16	M	50
4	35	L	50
5	30	L	50
6	10	L	50
7	15	M	50
8	23	M	50
9	23	L	50
10	12	M	50
11	18	M	50
12	16	M	50
13	22	L	50
14	15	M	50
15	15	M	50
16	10	M	50
17	18	M	50
18	18	M	50
19	35	L	50
20	16	M	50

Example 1

Example 1

- The weighted sum of variable x_1 is equal to 18950.
- Number of medium and large enterprises according to this survey is equal to 700 (14 medium enterprises x 50) and 300 (6 large enterprises x 50) respectively.
- **Assumption:** The exact population total of monthly revenue is known and equals 19000 and the real number of medium and large enterprises is equal to 720 and 280 respectively.
- **Problem** We would like to change the design weights in such a way that known auxiliary totals will be reproduced. In other words, we would like to slightly modify the initial weights so that the sum of x_1 based on the new weights is equal to 19000 and weighted sum of medium and large enterprises is equal to 720 and 280 respectively.
- **Solution:** Use calibration
- The SAS code which solves the problem for creating the preliminary datasets and rucalling the macro CALMAR2 command is given on the next slide.

Example 1 – solution using CALMAR2

```
/***********Creation of input dataset with drawn units*************/
data sample;
input enterprise $ size $ revenue weight;
cards;
ent01   M       18  50
ent02   M       14  50
ent03   M       16  50
ent04   L       35  50
ent05   L       30  50
ent06   L       10  50
ent07   M       15  50
ent08   M       23  50
ent09   L       23  50
ent10   M       12  50
ent11   M       18  50
ent12   M       16  50
ent13   L       22  50
ent14   M       15  50
ent15   M       15  50
ent16   M       10  50
ent17   M       18  50
ent18   M       18  50
ent19   L       35  50
ent20   M       16  50
;
run;
```

Example 1 – solution using CALMAR2

```
/**********Creation dataset with known population totals************/
data totals;
input var $ n mar1 mar2;
cards;
size 2 280 720
revenue 0 19000 .
;
run;

/*****************Library containing CALMAR**********************/
libname calm 'D:\Lamborghini\Calibration';
options mstored sasmstore=calm;

/**********************Call to CALMAR****************************/
%CALMAR2(DATAMEN=sample, POIDS=weight, IDENT=enterprise, MARMEN=totals,
M=1,DATAPOI=wcal, POIDSFIN=cal_weights)
```

Example 1 – calibration weights

Example 1 – calibration weights

Number of enterprise	Monthly revenue x_1	Enterprise size x_2	d_i	w_i
1	18	M	50	52,275
2	14	M	50	50,5821
3	16	M	50	51,4286
4	35	L	50	50,5462
5	30	L	50	48,4301
6	10	L	50	39,9657
7	15	M	50	51,0054
8	23	M	50	54,3911
9	23	L	50	45,4675
10	12	M	50	49,7357
11	18	M	50	52,275
12	16	M	50	51,4286
13	22	L	50	45,0443
14	15	M	50	51,0054
15	15	M	50	51,0054
16	10	M	50	48,8893
17	18	M	50	52,275
18	18	M	50	52,275
19	35	L	50	50,5462
20	16	M	50	51,4286

Example 2 – register based statistics (artificial data set)

Example 2 – register based statistics (artificial data set)

No.	Enterprise size	Section	Revenue	Legal status
1	Small	Section 1	NA	A
2	Large	Section 2	Small	B
3	Large	Section 2	High	NA
4	Small	Section 2	Small	C
5	Small	Section 1	NA	C
6	Small	Section 1	High	C
7	Large	Section 1	High	C
8	Large	Section 2	Small	C
9	Large	Section 1	Small	B
10	Small	Section 1	High	B
11	Large	Section 2	Small	B
12	Small	Section 2	Small	C
13	Large	Section 1	Small	A
14	Small	Section 2	Small	NA
15	Small	Section 1	High	B
16	Large	Section 2	Small	B
17	Large	Section 1	High	C
18	Small	Section 2	High	A
19	Small	Section 1	High	NA
20	Large	Section 1	Small	B

Example 2 – register based statistics

Two-way contingency table – the problem of nonresponse

- The main goal is to create two-way contingency table which shows the structure of revenue and legal status. Because of the fact that variables revenue and legal status are affected by nonresponse final table will not be correct.
- Description of variables: **Enterprise size** (Small, Large), **Legal status** (A, B, C), **Section** (Section 1, Section 2, Section 3), **Revenue** (Small, High), NA – not available.

	Revenue		
Legal status	Small	High	Total
A	1	1	2
B	5	2	7
C	3	3	6
Total	9	6	15

- The number of enterprises in two-way contingency tables does not add up to 20.
- **Solution:** Use calibration approach to adjust numbers in particular cells.

How to find calibration weights?

How to find calibration weights?

1. Create artificial "design weights". If for any enterprise the legal status or revenue is not known than initial weight $d_i = 0$. Otherwise $d_i = 1$.
2. Choose auxiliary variables. Because for all enterprises in register information about section and enterprise size is known use theme as covariates to find calibration weights w_i. In this example three variables were taken into account: $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$.

$$\mathbf{x}_{i1} = \begin{cases} 1 & \text{if } i\text{-th enterprise is large,} \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

$$\mathbf{x}_{i2} = \begin{cases} 1 & \text{if } i\text{-th enterprise is small,} \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

$$\mathbf{x}_{i3} = \begin{cases} 1 & \text{if } i\text{-th enterprise is from section 1} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

3. Use statistical software and find calibration weights w_i.

Example 2 – register based statistics (artificial data set)

Example 2 – register based statistics (artificial data set)

No.	Enterprise size	Section	Revenue	Legal status	d_i	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$	$\mathbf{x}_{i3}$	w_i
1	Small	Section 1	NA	A	0	0	1	1	0
2	Large	Section 2	Small	B	1	1	0	0	1,0447761
3	Large	Section 2	High	NA	0	1	0	0	0
4	Small	Section 2	Small	C	1	0	1	0	1,6069652
5	Small	Section 1	NA	C	0	0	1	1	0
6	Small	Section 1	High	C	1	0	1	1	1,7263682
7	Large	Section 1	High	C	1	1	0	1	1,1641791
8	Large	Section 2	Small	C	1	1	0	0	1,0447761
9	Large	Section 1	Small	B	1	1	0	1	1,1641791
10	Small	Section 1	High	B	1	0	1	1	1,7263682
11	Large	Section 2	Small	B	1	1	0	0	1,0447761
12	Small	Section 2	Small	C	1	0	1	0	1,6069652
13	Large	Section 1	Small	A	1	1	0	1	1,1641791
14	Small	Section 2	Small	NA	0	0	1	0	0
15	Small	Section 1	High	B	1	0	1	1	1,7263682
16	Large	Section 2	Small	B	1	1	0	0	1,0447761
17	Large	Section 1	High	C	1	1	0	1	1,1641791
18	Small	Section 2	High	A	1	0	1	0	1,6069652
19	Small	Section 1	High	NA	0	0	1	1	0
20	Large	Section 1	Small	B	1	1	0	1	1,1641791

Example 2 – register based statistics

Two-way contingency table – before calibration

	Revenue		
Legal status	Small	High	Total
A	1	1	2
B	5	2	7
C	3	3	6
Total	9	6	15

Two-way contingency table – after calibration

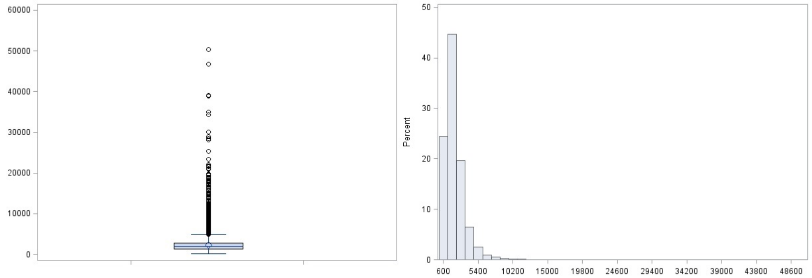
	Revenue		
Legal status	Small	High	Total
A	1,16	1,61	2,77
B	6,14	2,77	8,91
C	4,27	4,05	8,32
Total	11,57	8,43	20

The problem of outliers

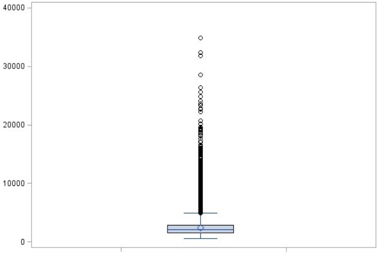
The problem of outliers

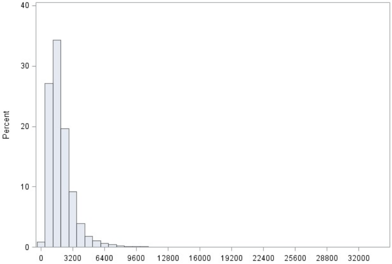
- The problem of outliers is very crucial in many statistical surveys.
- Outliers may influence the final results i.e., they produce significantly biased results and can considerably affect the survey quality.
- As a rule, this problem is evident in all kinds of surveys conducted by statistical offices of many countries, including those focused on business statistics, where asymmetric distribution of many variables is quite normal, although definitely undesirable from the point of view of estimation.

Boxplot and histogram of Household's available income

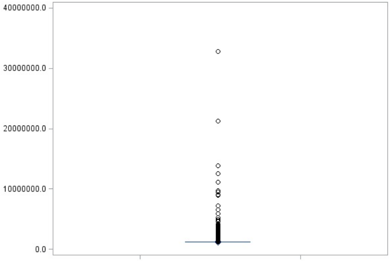


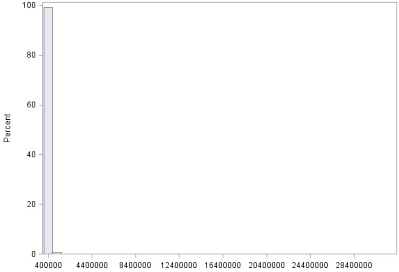
Boxplot and histogram of Household's total expenditures





Boxplot and histogram of Annual revenue for enterprises





The problem of outliers

The problem of outliers

- It has to be mentioned that the classic calibration estimator is not robust

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

- The final formula on weight w_k does not take into account the values of the variable of interest y_k.

Wright estimator

Wright estimator

- The process of construction robust calibration estimators was described by Duchesne
- The starting point constitutes so called class of Wright estimators. For $\{(q_i, r_i), q_i > 0, r_i \geq 0\}$ the Wright estimator of total value $Y = \sum_{i=1}^{N} y_i$ can be defined as follow:

$$\hat{Y}_{Wr} = \mathbf{X}^T \hat{B}_q + \sum_{i=1}^{n} r_i e_i, \qquad (24)$$

where $\hat{B}_q$ is definied as

$$\hat{B}_q = \left(\sum_{i=1}^{n} q_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T \right)^{-1} \sum_{i=1}^{n} q_i \underline{\mathbf{x}}_i y_i \qquad (25)$$

and

$$e_i = y_i - \underline{\mathbf{x}}_i^T \hat{B}_q. \qquad (26)$$

Wright estimator

Wright estimator

- Well known GREG estimator, which in fact belongs to the family of calibration estimators, is a particular example of Wright estimator obtained by assuming that $(q_i, r_i) = (d_i/c_i, d_i)$. For such chosen q_i and r_i we have:

$$\hat{B}_q = \left(\sum_{i=1}^{n} \frac{d_i}{c_i} \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T \right)^{-1} \sum_{i=1}^{n} \frac{d_i}{c_i} \underline{\mathbf{x}}_i y_i \tag{27}$$

and

$$\begin{aligned} \hat{Y}_{Wr} = \mathbf{X}^T \hat{B}_q + \sum_{i=1}^{n} r_i e_i &= \mathbf{X}^T \hat{B}_q + \sum_{i=1}^{n} d_i \left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_q \right) \\ &= \left(\mathbf{X} - \hat{\mathbf{X}} \right)^T \hat{B}_q + \sum_{i=1}^{n} d_i y_i \\ &= \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}} \right)^T \hat{B}_q = \hat{Y}_{GREG} \end{aligned} \tag{28}$$

Wright estimator

Wright estimator

Wright estimator can be considered as calibration estimator obtained by solving following optimalization problem:

(C1) Find the minimum of distance function:

$$D(\mathbf{w}, \mathbf{r}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(w_i - r_i)^2}{q_i} \longrightarrow \min, \tag{29}$$

(C2) Calibration equations:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{30}$$

(C3) Calibration constraints:

$$L \leq w_i \leq U. \tag{31}$$

The choice of q_i and r_i

The choice of q_i and r_i

- The problem of constructing robust calibration estimator depends on a proper choice of factors q_i and r_i. The first proposal is as follow:

$$(q_i, r_i) = \left(d_i h_i^{1-\alpha} u_i / c_i, r_i\right) \tag{32}$$

where

$$h_i = min\left(1, \frac{t}{x_i / med\,(x_i)}\right), \tag{33}$$

$$u_i = \frac{\psi\left(\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_g\right) / \left(\sigma h_i^{\alpha} \sqrt{c_i}\right)\right)}{\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_g\right) / \left(\sigma h_i^{\alpha} \sqrt{c_i}\right)}, \tag{34}$$

$$\psi = \psi_{Hub}\,(x; c) = \begin{cases} c & \text{if } x > c, \\ x & \text{if } |x| \le c, \\ -c & \text{if } x < -c, \end{cases} \tag{35}$$

and t, c, σ, α are properly chosen constants and $\hat{B}_g$ is a robust estimator of a regression coefficient.
- For r_i Lee (1991) suggested $r_i = \theta d_i$, where $\theta \in [0, 1]$.

The choice of q_i and r_i

The choice of q_i and r_i

- Another proposal is given by Duchesne (1999):

$$(q_i, r_i) = \left(d_i h_i^{1-\alpha} \hat{u}_i / c_i, d_i u_i^*\right) \tag{36}$$

where

$$\hat{u}_i = \frac{\psi\left(\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_0\right) / \left(\sigma h_i^\alpha \sqrt{c_i}\right)\right)}{\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_0\right) / \left(\sigma h_i^\alpha \sqrt{c_i}\right)}, \tag{37}$$

$$\hat{u}_i^* = \frac{\psi^*\left(\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_r\right) / \left(\sigma h_i^\alpha \sqrt{c_i}\right)\right)}{\left(y_i - \underline{\mathbf{x}}_i^T \hat{B}_r\right) / \left(\sigma h_i^\alpha \sqrt{c_i}\right)}, \tag{38}$$

$$\psi^* = \psi_{Hubmod}(x; c) = \begin{cases} x & \text{if } |x| \leq a, \\ asign(x) & \text{if } |x| > a \text{ and } |x| \leq a/b, \\ bx & \text{if } |x| > a/b, \end{cases} \tag{39}$$

The choice of q_i and r_i

The choice of q_i and r_i

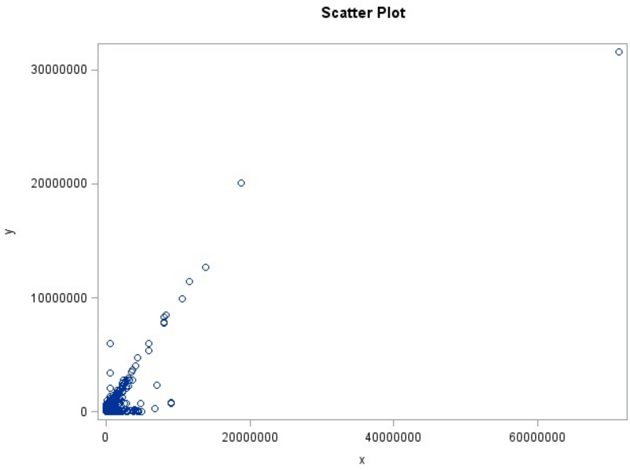
$\hat{B}_0$ denotes the Coakley and Hettmansperger estimator of a regression coefficient and

$$\hat{B}_r = \left(\sum_{i=1}^{n} d_i h_i^{1-\alpha} \hat{u}_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T / c_i \right)^{-1} \cdot \sum_{i=1}^{n} d_i h_i^{1-\alpha} \hat{u}_i \underline{\mathbf{x}}_i y_i / c_i \tag{40}$$

Population under study

- The simulation study investigated two variables: annual revenue was the response or output variable (Y) and VAT information was the only one auxiliary variable.
- Data about the first variable (annual revenue) came from the DG-1 survey. The VAT variable came from the VAT register.
- To conduct the simulation study, a pseudo-population was created (further referred to as the MEETS real dataset), consisting of all enterprises included in the DG-1 survey for which information about the auxiliary variable was available.
- Enterprises which reported zero revenue in the DG-1 survey, were excluded from the dataset.
- Taking advantage of a strong correlation between the pseudo-population and the VAT register, it was possible to match VAT information with records in the pseudo-population.
- The resulting dataset consisted of about 18,000 records containing complete information about the variables under analysis.

Population under study



Totals for y and x variables (MEETS)

Population	$\sum_{i=1}^{N} x_i$	$\sum_{i=1}^{N} y_i$	N
MEETS	1302052161	840579004	18454

- y - annual revenue of enterprise (in thousands of PLN)
- x - VAT (in PLN)

Chosen estimators

Chosen estimators

Five different estimators were taken into account:

- Horvitz-Thompson estimator (HT)
- GREG estimator (GREG)
- GREG estimator whose weight were limited (GREG/L)
- Robust calibration estimator based on Lee formula for r i.e. $r_i = \theta d_i$ (RCL)
- Robust calibration estimator based on Duchesne formula for r i.e. $r_i = d_i u_i^*$ (RCD)
- The following constants were taken into account: $\theta = 0.3$, $\alpha = 1$, $t = 1.5$, $c = 2$, $c_i = 1$, $a = 9$, $b = 0.25$
- $k = 500$ replications were done using simple random 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 two measures:

- The relative bias of estimators:

$$Bias = \frac{E\left(\hat{T}\right) - T_y}{T_y} \quad (41)$$

- The coefficient of variation CV:

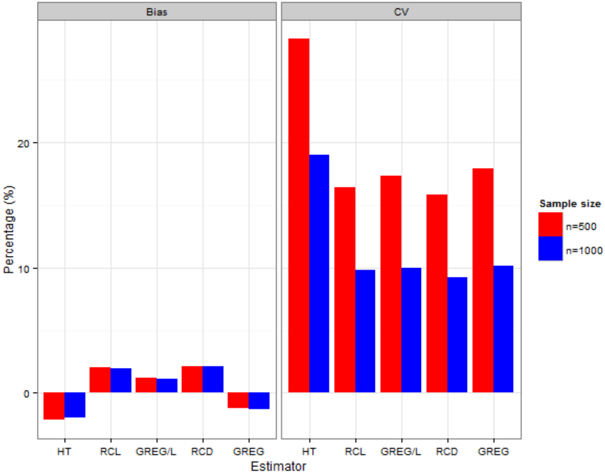
$$CV = \frac{\sqrt{V}}{E\left(\hat{T}\right)} \quad (42)$$

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

Results

	$n = 500$		$n = 1000$	
Estimator	Bias	CV	Bias	CV
HT	-2.15	28.26	-1.95	19.05
GREG	-1.21	17.94	-1.28	10.15
GREG/L	1.17	17.35	1,14	10.02
RCL	2.05	16.46	1.95	9.85
RCD	2.15	15.85	2.10	9.24

Results of the simulation study



Summary

Sumary

- Robust calibration estimators provided a reduction of coefficient of variation compared to other estimators. They were more efficient then the rest of estimators.
- Coefficient of variations, as expected, were smaller for bigger samples for all estimators.
- The robust calibration estimators have comparable bias to the bias of HT and GREG-GREG/L estimators.

Literature

Literature

- Särndal C-E., Lundström S. (2005), "*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.
- Wallgren A., Wallgren B. (2007), "*Register-based Statistics: Administrative Data for Statistical Purposes*", Wiley.
- Deville J-C., Särndal C-E. (1992), "*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.
- Duchesne P. (1999), "*Robust calibration estimators*", Survey Methodology, Vol. 25, No. 1.
- Lee H. (1991), "*Model-based estimators that are robust to outliers*", Proceedings of the 1991 Annual Research Conference, U.S. Bureau of Census.
- Särndal C-E. (2007), "*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.

Thank you very much for your attention!