Estimation theory for sample surveys

- **Calibration Estimators Theory**
- **Principle of Calibration Methods**

Cenker Burak METİN

Head of Survey and Questionnaire Design Group

Calibration Estimators Theory

- **Calibration** is a kind of adjustment that makes sample distrubution or sample statistics agree with the population distrubitions or parameters by using auxiliary variables.
- Calibration is defined as the computation of sampling weights by taking into account the auxiliary information subject to the calibration equation(s).
- **The purpose of calibration** is to increase the accuracy of the estimation.
- Although **HT** estimator is an unbiased estimator, it **may give inefficient estimates** especially for the characteristics, which have a negative correlation with the inclusion probabilities. In such situations, **calibration methods would be helpful to get estimators that are more efficient.**

Principle of Calibration Methods

- Calibration gives a single set of weights that can be used for all survey variables.
- The idea behind this is that if **calibration weights, $\mathbf{w_k}$,** fit the estimates of the total parameter values for auxiliary variable(s), then they might be good for other survey variables, as well.
- The main idea of the calibration estimator, $\hat{t}_y^{CAL}$, is to look for the family of estimators that assures the following proper procedures:

1. It should be written as a linear combination of target variable and calibration weights:

$$\hat{t}_y^{CAL} = \sum\nolimits_{k\in s} w_k \, y_k$$

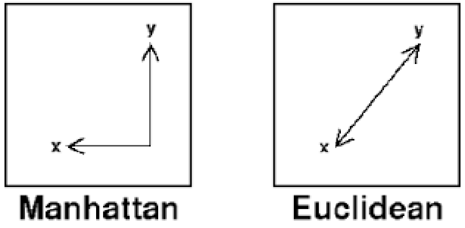
2. It must be optimized at least under the main calibration constraint.

$$\hat{t}_x^{CAL} = \sum_{k\in s} w_k \, \boldsymbol{x}_k = t_{\boldsymbol{x}}$$

3. The set of the calibration weights $\{w_k(s): k \in s\}$ should be close to the set of design weights $\{d_k = 1/\pi_k : k \in s\}$.

Calibration Estimators Theory

- Shortly, the main purpose of the calibration is to obtain weights that make estimated totals become equal to known population totals for auxiliary variables.
- Calibration can deal with a variety of conditions like complex sampling designs, adjustment for non-response and frame errors.
- Many different sets of calibration weights may be constructed in order to guarantee the consistency with auxiliary variable totals.
- **The minimum distance method** is one of the most popular approach to get calibration estimator proposed by Deville and Särndal
- Basically, a distance measure $G_k(w, d)$ is chosen and design weights are modified to calibration weights w_k by minimizing the total distance $\sum_s G_k(w_k, d_k)$.



Calibration Estimators Theory

- The calibration estimator can be shown as

$$\hat{\mathrm{t}}_y^{CAL} = \sum_s w_k y_k$$

where the calibration weight is $w_k = d_k F(q_k \boldsymbol{x}'_k \lambda) = d_k g_k$.

- It consists of design weight and calibration adjustment factor. (Remember $\mathrm{d_k \phi_{k(l)}}$)
- $\mathrm{F(.)}$ is the inverse function of $\mathrm{g(.)}$ which is the derivative of the given distance function $\mathrm{g_k(w, d)} = \partial \mathrm{G_k(w, d)} / \partial \mathrm{w}$ and $\mathrm{q_j}$ is positive scale factor determined by the specialist.
- By solving the main calibration constraint

$$\sum_s d_k \boldsymbol{x_k} \, F(q_k \boldsymbol{x}'_{\boldsymbol{k}} \lambda) = \sum_U \boldsymbol{x_k}$$

it is possible to get Lagrange multipliers λ.

Calibration Estimators Theory

- The most frequently used calibration estimators by NSO's are GREG and raking ratio, and they can also be derived with this approach.
- If the Chi-square distance function is selected as the distance function and showed as

$$G_k(w_k, d_k) = (w_k - d_k)^2/2q_k\, d_k$$

  then for $F_k(q_k u) = F_k(q_k \boldsymbol{x}'_{\boldsymbol{k}} \lambda) = g$ calibration adjustment factor (g-weight) would be denoted as $F_k(q_k u) = 1 + q_k u$.
- This calibration estimator, known as the linear method, turns into GREG estimator for $\mathrm{q}_\mathrm{k} = 1$. Then when Lagrange equations solved calibration weight can be denoted as

$$w_k = d_k(1 + x_k{}'\lambda\,)$$

- Similarly it is possible to get raking ratio estimator by some other calculations.

Calibration Estimators Theory

- For GREG, g-weights may sometimes contain negative values, and in some cases, they may also be far from design weights.
- For raking ratio, g-weights never become negative but calibration weights may be very large compared to the design weights.
- These are undesirable problems for survey statisticians since they will affect the estimate.
- Considering these problems, Deville and Särndal introduced two new calibration estimator (called as **truncated and logit** calibration estimators) where adjustment factors can fall within the bounds set by statisticians.
- By the help of these approaches, you can eliminate the extreme weights while good properties of the estimators are protected.

Calibration Estimators Theory

- Usually, different $F_k(u)$ functions derive different estimators.
- However, for medium and large-scale samples, the estimates produced by these estimators are expected to differ slightly.
- Because, any $\hat{t}_{yCAL}$ estimator is asymptotically equivalent to GREG estimator.

$$AV\left(\hat{t}_y^{CAL}\right) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_{jk}\pi_{kl})\check{\varepsilon}_k\check{\varepsilon}_l$$

$$\hat{V}\left(\hat{t}_y^{CAL}\right) = \sum_{k\in s}\sum_{l\in s}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\right)(g_k\check{e}_k)(g_l\check{e}_l).$$

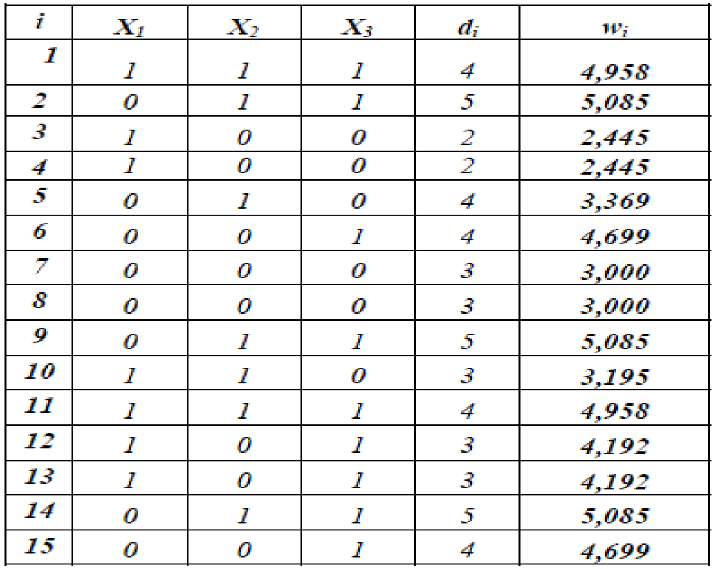
- In these equations, $\check{\varepsilon}_k = \varepsilon_k/\pi_k$ and $\check{e}_k = e_k/\pi_k$ are the weighted representations of population regression errors and the sample regression residuals, respectively.
- The different $g_k = F(q_k\boldsymbol{x}_k'\lambda)$ calibration adjustment factors for alternative calibration estimators will differentiate the variance estimates.

Calibration Estimators Theory

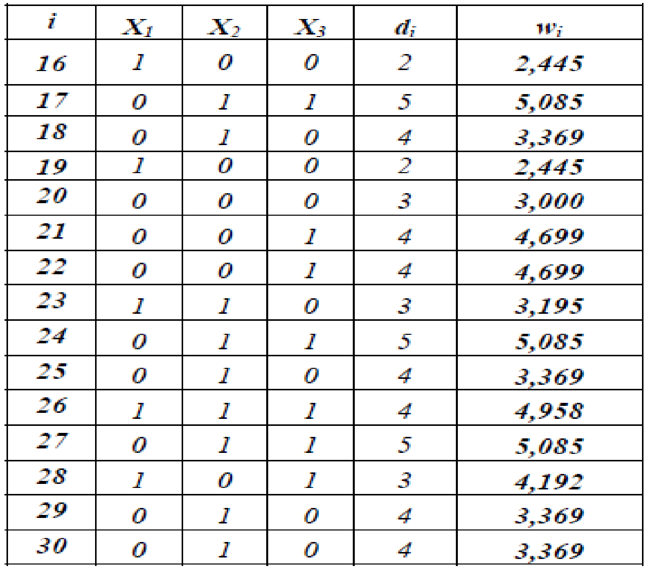
- For the unrestricted calibration estimators (GREG and raking), some disadvantages may arise.
- As a result of the calibration, large variation in weights may occur and variance may increase because of the weights.
- There may be cases where the adjustment factor is extremely greater or less than one and negative adjustment factors may occur for GREG.
- On the other hand, when using distance functions that restrict weights, the solution may not be found due to lack of constraints. (optimization can fail for TRN and LOG)
- The lower and upper bounds should be revised when such a situation is encountered.

Calibration Estimators Theory

- **Example:** Suppose there are three auxiliary variables correlated with the research variable and for which population total values are available. Let's say the auxiliary variables are defined as below.



i	X_1	X_2	X_3	d_i	w_i
1	1	1	1	4	4,958
2	0	1	1	5	5,085
3	1	0	0	2	2,445
4	1	0	0	2	2,445
5	0	1	0	4	3,369
6	0	0	1	4	4,699
7	0	0	0	3	3,000
8	0	0	0	3	3,000
9	0	1	1	5	5,085
10	1	1	0	3	3,195
11	1	1	1	4	4,958
12	1	0	1	3	4,192
13	1	0	1	3	4,192
14	0	1	1	5	5,085
15	0	0	1	4	4,699



i	X_1	X_2	X_3	d_i	w_i
16	1	0	0	2	2,445
17	0	1	1	5	5,085
18	0	1	0	4	3,369
19	1	0	0	2	2,445
20	0	0	0	3	3,000
21	0	0	1	4	4,699
22	0	0	1	4	4,699
23	1	1	0	3	3,195
24	0	1	1	5	5,085
25	0	1	0	4	3,369
26	1	1	1	4	4,958
27	0	1	1	5	5,085
28	1	0	1	3	4,192
29	0	1	0	4	3,369
30	0	1	0	4	3,369

X_1:1 for Male, 0 for Female

X_2:1 for Young, 0 for Old

X_3:1 for Blue-collar, 0 for civil servant

$\boldsymbol{X = [\, X_1\ X_2\ X_3\,]' = [30,60,70]'}$

$\hat{X} = \sum_{i=1}^{n} d_i x_i \qquad \hat{X} = [35,68,67]'$

Calibration Estimators Theory

	$\sum_{i=1}^{n} d_i x_i x_i'$	
35	18	21
18	68	33
21	33	67

$$\lambda = \left[\sum_{i=1}^{n} d_i x_i x_i'\right]^{-1} \left(X - \hat{X}\right)$$

	$\left[\sum_{i=1}^{n} d_i x_i x_i'\right]^{-1}$	
0,0366	-0,0054	-0,0088
-0,0054	0,0200	-0,0082
-0,0088	-0,0082	0,0217

$$\left(X - \hat{X}\right) = \begin{bmatrix} -5 \\ -8 \\ 3 \end{bmatrix}, \quad \lambda = \left\{ \begin{array}{|c|c|c|} \hline 0,0366 & -0,0054 & -0,0088 \\ \hline -0,0054 & 0,0200 & -0,0082 \\ \hline -0,0088 & -0,0082 & 0,0217 \\ \hline \end{array} \right\} \begin{bmatrix} -5 \\ -8 \\ 3 \end{bmatrix},$$

$$w_1 = 4\left\{ 1 + (1 \quad 1 \quad 1) \begin{bmatrix} 0,2226 \\ -0,1576 \\ 0,1747 \end{bmatrix} \right\} = 4,958$$

$$\lambda = \begin{bmatrix} 0,2226 \\ -0,1576 \\ 0,1747 \end{bmatrix}$$

Weight Trimming

- It is desirable to avoid assigning extreme weights to any unit in the sample.
- Extremely large and varying weights may cause a significant increase in variance.
- Restricting the extremely large weights control the increase in the variance.
- Calibration weights consist of more variability than design weights.
- Precision of the estimates decrease when large variations in calibration weights
- Extreme weights can produce inefficient estimations.
- **Trimming large weights reduces the variability that depends on weights.**
- Trimming process causes bias while the variance is reduced.
- Therefore, trimming should be done with the assumption of decrease in MSE.

Weight Trimming

- The general application in weight trimming is to keep the weights within some upper and lower bounds.
- Possible to use complex approaches for weight trimming.
- However, researchers generally prefer simple and practical implementation methods.
- Verma, proposed a simple procedure where the main problem is only a limited number of large weights.
- After the calibration step, rescale the calibration weights to average 1 over the sample cases, then any rescaled weights outside the bounds 1/T and T are made equal to these boundary values. As a rule of thumb, the constant T is given a value between 2 and 3.

Example for weight trimming

References

Alkaya, A., & Esin, A., (2005). Calibration estimator. Gazi University Journal of Science, 18(4), 591-601.

Verma, V. (2014). Sampling: An Introduction, University of Siena, Siena, September 2014.

Metin, C. B., Şahin Tekin, S. T., & Özdemir, Y. A. (2021). Restricted calibration and weight trimming approaches for estimation of the population total in business statistics. Journal of Applied Statistics, 48(13-15), 2658-2672.

Calibration Estimator usage in Statistical Offices

- **Using auxiliary information to adjust weights**
- **Poststratification**
- **Raking Ratio (Iterative Proportional Fitting)**

Cenker Burak METİN

Head of Survey and Questionnaire Design Group

Using auxiliary information to adjust weights

- **Use of auxiliary variables in the sampling design**
  - ✓ Quantitative variable: Sampling with unequal probabilities
  - ✓ Qualitative variable: Stratified sampling
- **Use of auxiliary variables in the estimator**
  - ✓ Quantitative variable: Ratio estimator
  - ✓ Quantitative variable: Regression estimator
  - ✓ Qualitative variable: Post-stratification estimator
- **When was the first auxiliary info used historically?**
  - ✓ John Graunt (1662) "Rule of Three" if AB=CD then D = AB/C
  - ✓ P.S. Laplace (1786) ratio estimator

Using auxiliary information to adjust weights

- **By using auxiliary information, we can**
  - ✓ Gain in precision
  - ✓ Correct for nonresponse effects
  - ✓ Calibrate on known population

    (or other external source) values

- **Auxiliary variables are the variables that are external to the survey.**
  - ✓ The values of the auxiliary variable are available for all population units
  - ✓ The values of the auxiliary variable are available for all sample units
  - ✓ The values of the auxiliary variable are available for respondents and on an aggregated population level.

Using auxiliary information to adjust weights

RATIO ESTIMATOR

The ratio estimator is a statistical parameter and is expressed as the ratio of the mean of two random variables.

- ✓ Ratio estimates are biased
- ✓ Because rate estimates are asymmetrical, symmetrical tests such as the t-test should not be used to construct confidence intervals.
- ✓ Ratio estimation may be used to adjust for nonresponse.

For ratio estimation to apply, two quantities y_k and x_k must be measured on each sample unit.

Let's we want to estimate population total parameter in U.

$$\boxed{t_y = \sum_{k \in U} y_k .}$$

A sample s is selected in U by using $p(s)$. Data is collected for y_k for $k \in S$.

RATIO ESTIMATOR

Suppose now that we have an auxiliary variable x with the following known information:

- $x_k \geq 0, k \in S.$
- $t_x = \sum_U x_k \, (population \ total \ of \ x).$

$$\boxed{B = \frac{t_y}{t_x} = \frac{\overline{y_U}}{\overline{x_U}}}$$

where B is a ratio estimator.

Ratio estimator for the total Y if s is a SRS of size n with the design weight $d_k = \frac{N}{n}$:

$$\boxed{\hat{t}_y^{RAT} = t_x \frac{\sum_{k \in S} d_k y_k}{\sum_{k \in S} d_k x_k} = t_x \frac{(\frac{N}{n}) \sum_{k \in S} y_k}{(\frac{N}{n}) \sum_{k \in S} x_k} = t_x \frac{\sum_{k \in S} y_k}{\sum_{k \in S} x_k}}$$

RATIO ESTIMATOR

Ratio estimator for the total Y if s is a STR of size n, with $d_k = N_h/n_h$, if $k \in s_h$:

$$\hat{t}_y^{RAT} = t_x \frac{\sum_h \sum_{k \in s_h} d_k y_k}{\sum_h \sum_{k \in s_h} d_k x_k} = t_x \frac{\sum_h (\frac{N_h}{n_h}) \sum_{k \in s_h} y_k}{(\frac{N_h}{n_h}) \sum_{k \in s_h} x_k}$$

Notation as a weight adjustment

$$\hat{t}_y^{RAT} = t_x \frac{\hat{t}_y}{\hat{t}_x} = \frac{t_x}{\hat{t}_x} \sum_{k \in S} d_k y_k = \sum_{k \in S} \frac{t_x}{\hat{t}_x} d_k y_k = \sum_{k \in S} g_k d_k y_k = \sum_{k \in S} w_k y_k$$

RATIO ESTIMATOR

Design or sampling weight, selection with $p(s)$: d_k

Weight adjustment or g-weight:

$$\boxed{g_k = \frac{t_x}{\hat{t}_x} = \frac{t_x}{\sum_{k\in s} d_k x_k}}$$

Final weight: $w_k = d_k g_k$.

Note: g_k depends on s. $\sum_{k\in s} w_k x_k = t_x$ $(\,calibration)$.

Bias: $\hat{t}_y^{RAT}$ is slightly biased.

The bias is of order $\frac{1}{n}$ it can be large for small n.

If the sample is of fixed size:

$$\boxed{Bias = E\left(\hat{t}_y^{RAT}\right) - t_y \approx t_y\left(\frac{var(\hat{t}_x)}{{t_x}^2} - \frac{cov\left(\hat{t}_x, \hat{t}_y\right)}{t_x t_y}\right)}$$

RATIO ESTIMATOR

For SRS:

$$\boxed{Bias \approx t_y\left(1-\frac{n}{N}\right)\frac{1}{n}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)}$$

$$var(\hat{t}_x) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}S_x^2, \qquad with\ S_x^2 = \frac{1}{N-1}\sum_U (x_k-\bar{X})^2$$

$$cov(\hat{t}_x, \hat{t}_y) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}S_{xy}, \qquad with\ S_{xy} = \frac{1}{N-1}\sum_U (x_k-\bar{X})(y_k-\bar{Y})$$

If s is a SRS and n large, the bias equals zero if and only if

$$\frac{S_x^2}{\bar{X}^2} = \frac{S_{xy}}{\bar{X}\bar{Y}} \ or\ \frac{S_x^2}{S_{xy}} = \frac{\bar{X}}{\bar{Y}}$$

RATIO ESTIMATOR

Let the linear regression $y_k = a + bx_k + \varepsilon_k, k \in U$.

Var(ε_k)=σ^2.

The estimated parameters are: $b = \frac{S_{xy}}{S_x^2}$ $and\ a = \bar{Y} - b\bar{X}$.

Therefore: the bias equals zero if and only if b $= \frac{\bar{Y}}{\bar{X}}$, i.e. if the intercept a equals zero.

Ratio estimation is most appropriate if a straight line through the origin summarizes the relationship between x_k and y_k and if the variance of y_k about the line is proportional to x_k.

Under these conditions, $\hat{B}$ is the weighted least squares regression slope for the line through the origin with weights proportional to $1/x_k$—the slope $\hat{B}$ minimizes the sum of squares

$$\boxed{\sum_{k \in S} \frac{1}{x_k}\left(y_k - \hat{B}x_k\right)^2}$$

Using auxiliary information to adjust weights

REGRESSION ESTIMATOR

As with the ratio estimator, the regression estimator uses auxiliary variables that are correlated with the variable of interest to improve the precision of estimates of the mean and total of a population.

- ✓ We need auxiliary information.
- ✓ Total of the auxiliary quantitative variables used will be estimated exactly: $\hat{\boldsymbol{t}}_{\boldsymbol{xREG}} = \boldsymbol{t}_{\boldsymbol{x}}$ (consistency-respect of known totals.)
- ✓ The collection phase has been completed (use of auxiliary variables at the estimation stage).

Regression Estimator for SRS:

Quantities y_k and x_k are measured in each sample units. We aim at estimating the total $\boldsymbol{t}_{\boldsymbol{y}} = \sum_{\boldsymbol{k \in U}} \boldsymbol{y}_{\boldsymbol{k}}$ in U.

- ✓ x_k is a quantitative variable.
- ✓ The total $\boldsymbol{t}_{\boldsymbol{x}} = \sum_U x_k$ is known.

Regression estimator takes advantage of the correlation of x and y in the population.

The higher the correlation, the better it works.

REGRESSION ESTIMATOR

The coefficient of correlation between x and y is: $\boxed{\rho_{xy} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}}$ where

$$\boxed{S_{xy} = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{X})\,(y_k - \bar{Y}) \qquad S_x^2 = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{X})^2}$$

Model Assisted Approach:

It is supposed that a straight line regression model between y_k and the auxiliary variable x_k would provide a good fit.

The linear relation between the y variable and the x variable :

$$\boldsymbol{y_k = a + bx_k + E_k,}$$

where

$$\boldsymbol{\bar{E} = \frac{\sum_{k\in U} E_k}{N} = 0.}$$

$a\ and\ b$ are chosen in order to minimize $(\sum_{k\in U} {E_k}^2)$.

REGRESSION ESTIMATOR

According to the model, we have:

$$t_y = aN + bt_x$$

Where a and b are the ordinary least square regression coefficient calculated on U.

Proof:

$$\begin{aligned} t_y = \textstyle\sum_{k\in U} y_k &= \textstyle\sum_{k\in U}(a + bx_k + E_k) \\ &= a\textstyle\sum_{k\in U} 1 + b\sum_{k\in U} x_k + \sum_{k\in U} E_k \\ &= aN + bt_x \end{aligned}$$

Regression Estimator with a SRS:

The regression estimator can be written:

$$\hat{t}_y^{REG} = N\bar{y} + \hat{b}(t_x - N\bar{x})$$

$$\hat{t}_y^{REG} = \hat{t}_y^{HT} + \hat{b}(t_x - \hat{t}_x^{HT})$$

Expression as a weighted linear estimator (weights adjustment):

$$\hat{t}_y^{REG} = \sum_{k\in S} g_k(s)\frac{y_k}{\pi_k} = \sum_{k\in S} w_k(s)y_k$$

where $g_k(s) = 1 + \frac{n}{n-1}\frac{(\bar{X}-\bar{x})(x_k-\bar{x})}{S_x^2}$

GENERALIZED REGRESSION ESTIMATOR (GREG)

- GREG is a model-assisted estimator designed to improve the accuracy of estimations using auxiliary information.
- The GREG estimator ensures that sample estimates for auxiliary variables are consistent with known totals.
- When auxiliary information is appropriate at unit or domain level; the GREG estimator can be used to reduce the variance of the estimates by taking advantage of the relationship between the target variable and the auxiliary variable.
- Main idea is to construct assistant model and predict $\hat{y}_k = \boldsymbol{x_k}'\widehat{\boldsymbol{B}}$ values for all $k \in U$

  where $\widehat{\boldsymbol{B}} = \left(\sum_{k\in s} \frac{\boldsymbol{x_k x_k}'}{\sigma_k^2 \pi_k}\right)^{-1} \sum_{k\in s} \frac{\boldsymbol{x_k} y_k}{\boldsymbol{\sigma_k^2 \pi_k}}$
- By using predicted $\hat{y}_k$ values, we can get design-unbiased estimator for t_y.

$$\hat{t}_y^{GREG} = \sum_{k\in U} \hat{y}_k + \sum_{k\in s} d_k (y_k - \hat{y}_k)$$

$$= \sum_{k\in s} d_k y_k + \left(\sum_{k\in U} \hat{y}_k - \sum_{k\in s} d_k \hat{y}_k\right)$$

GENERALIZED REGRESSION ESTIMATOR (GREG)

- **Aim:** Making the residuals $y_k - \hat{y}_k$ very small with the help of a fit model to get accurate estimations.
- Modelling is the basis of the GREG approach.
- **Linear GREG, which is also a calibration estimator**, will be discussed here.

$$\hat{t}_y^{GREG} = (\textstyle\sum_{k\in U} \boldsymbol{x_k})' \boldsymbol{B_{s;dq}} + \sum_{k\in s} d_k \left(y_k - \boldsymbol{x'_k} \boldsymbol{B_{s;dq}}\right)$$

$$\boldsymbol{B_{s;dq}} = (\textstyle\sum_{k\in s} d_k q_k \boldsymbol{x_k} \boldsymbol{x'_k})^{-1} (\sum_{k\in s} d_k q_k \boldsymbol{x_k} y_k)$$

- As a standard, q_k =1 for all units. The choice of q_k generally has a limited effect on the accuracy
- Nearly unbised estimations occurs for any value of q_k (except for extremely large/small values).
- A close fitting linear regression of y on x holds the key to a small variance for $\hat{t}_{yGREG}$
- The g_k, which we have previously defined as the calibration adjusment factor (g-weight), is obtained in the GREG approach as

$$g_k^{(GREG)} = 1 + \left(t_x - \hat{t}_y^{HT}\right)' \hat{T}^{\,-1}\, x_k/\sigma_k^2 \quad \text{(independent from y)}$$

$$\hat{T} = \sum\nolimits_s x_k x_k^{'}/\sigma_k^2 \pi_k$$

GENERALIZED REGRESSION ESTIMATOR (GREG)

- GREG works as a regression estimator if one quantitative variable is used in the SRS

- Similarly, the ratio estimator is a special case of GREG when a single auxiliary variable is available and can be supported by a model, and the variance of the target variable is assumed to be a linear function $V(y) = \sigma^2 x$ of the auxiliary variable.

- Poststratification estimator is also special form of GREG estimator.

- Disadvantages of GREG :
  - ✓ May increase variance by causing large variation in weights.
  - ✓ There may be situations where the g-weights is too large or small and negative values.
  - ✓ Although asymptotically unbiased, it may be biased in small sample sizes.
  - ✓ The GREG estimator can be very sensitive in the presence of outliers.

POSTSTRATIFICATION

What is post-stratification?

Age, gender, education, occupation, etc. qualitative auxiliary variables such as can be used in the estimation phase. One of the well-known and frequently used examples of this is the post-stratification estimator. As in the Stratified Sampling method, the auxiliary variable information is not known on a unit basis before the sample selection; If the strata of a unit is determined in the sample survey, these strata are post-strata.

The main idea: Post-stratification is to separate the population into homogeneous strata according to the information from the sample.

Aim: Reduce variance.

POSTSTRATIFICATION

Post-stratification is applied in the following cases:

- ✓ If the necessary stratification variable is not available to classify and rank each unit
- ✓ When the stratification variable is available but cannot be used
- ✓ Post-stratification, although drawn proportionally from the sample population, can be used in subclasses.

For estimating the total $\boldsymbol{t_y \sum_{k \in U} y_k}$ in U, sample s is selected in U by using $p(s)$.

Data is collected for $y_k, k \in s$.

Suppose now that we have the following auxiliary information for a given partition

$U = U_1 \cup \ldots \cup U_h \cup \ldots \cup U_H$ of the population U.

- ✓ $N_H, h = 1, \ldots, H$ (population counts in the post-strata)
- ✓ $z_{hk} = 1$ if $k \in U_h$, 0 otherwise, for $k \in s$, $h = 1, \ldots, H$.

Partition: $U = U_1 \cup \ldots \cup U_h \cup \ldots \cup U_H$, with N_h the number of observation in U_h, $h = 1, \ldots, H$.

POSTSTRATIFICATION

The population total may be expressed as:

$$\boxed{t_y = \sum_h t_{y_h} = \sum_h N_h \bar{Y}_h}$$

N_h is known for $h = 1, \ldots, H$.

The mean $\bar{Y}_h$ is estimated with $k \ \in s$.

Using the estimator $\hat{\bar{Y}}_h = \hat{t}_{y_h} / \hat{N}_h$, we define the post-stratified estimator $\hat{t}_{yPOST}$ of Y:

$$\boxed{\hat{t}_y^{POST} = N_h \hat{\bar{Y}}_h = \sum_h N_h \frac{\hat{t}_{y_h}}{\hat{N}_h}}$$

Same for the post-stratified estimator $\hat{\bar{Y}}_{post}$ of $\bar{Y}$:

$$\boxed{\hat{\bar{Y}}_{POST} = \sum_h \frac{N_h}{N} \hat{\bar{Y}}_h = \sum_h \frac{N_h}{N} \frac{\hat{t}_{y_h}}{\hat{N}_h}}$$

with $\hat{N}_h = \sum_{k \in s_h} d_k$.

POSTSTRATIFICATION

Notes:

- ✓ $\boldsymbol{n_h}$, the number of observations in $s \cap U_h$ **is a random number**,
- ✓ $\hat{\bar{Y}}_h$ is not defined if $n_h = 0$.

Post-Stratification for SRS and Relation to Stratification

If is a SRS of size n with a post-stratification:

$$\boxed{\hat{t}_y^{POST} = \sum_h N_h \frac{\hat{t}_{y_h}}{\hat{N}_h} = \sum_h N_h \frac{\sum_{s_h} \frac{N}{n} y_k}{N \frac{n_h}{n}} = \sum_h N_h \frac{\sum_{s_h} y_k}{n_h}}$$

where n_h is a random value, $E(n_h) = n\frac{N_h}{N}$

If s is a STR sample of size n with the allocation n_h, $h = 1, \dots, H$:

$$\boxed{\hat{t}_y = \sum_h N_h \hat{\bar{Y}}_h = \sum_h N_h \bar{y}_{sh} = \sum_h N_h \frac{\sum_{s_h} y_k}{n_h}}$$

where n_h is a constant value.

POSTSTRATIFICATION

NOTATION AS A WEIGHT ADJUSTMENT

General case:

$$\boxed{\hat{t}_y^{POST} = N_h \hat{\bar{Y}}_h = \sum_h \frac{N_h}{\hat{N}_h} \sum_{k \in s_h} \boldsymbol{d_k} y_k = \sum_k \boldsymbol{g_k} \boldsymbol{d_k} y_k = \sum_k \boldsymbol{w_k}\, y_k}$$

Design or sampling weight, selection with $p(s)$: d_k.

Weight adjustment or g-weight: $\boldsymbol{g_k} = \frac{N_h}{\hat{N}_h} = \frac{N_h}{\sum_{k \in s_h} d_k}$ if $k \in s_h$.

Final weight: $w_k = \boldsymbol{d_k} \boldsymbol{g_k}$

Notes: g_k depends on s; $\hat{N}_h = n_h N / N_h$ for SRS; and $\sum_{s_h} w_k = N_h$ (calibration).

POSTSTRATIFICATION

For s a SRS of size n with a post-stratification based on N_h, $h = 1, \dots, H$, we have the approximate variance:

$$\boxed{Var\left(\hat{t}_y^{POST}\right) \approx N^2\left[\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_h \frac{N_h}{N} S_h^2 + \left(1-\frac{n}{N}\right)\frac{1}{n^2}\sum_h \frac{N-N_h}{N} S_h^2\right]}$$

First part: variance for a STR with proportional allocation $n_h = nN_h/N$.

Second part: variability due to the random size n_h. Small in comparison with the first part $(1/n^2)$.

If n is large, we can estimate the variance as:

$$\boxed{\widehat{Var}\left(\hat{t}_y^{POST}\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_h \frac{N_h}{N} s_h^2 = \left(1-\frac{n}{N}\right)\frac{N}{n} N_h s_h^2}$$

where:

$$\boxed{s_h^2 = \frac{1}{n_h - 1}\sum_{k \in s_h} \left(y_k - \bar{y}_{s_h}\right)^2}$$

i.e. H-T variance for STR with proportional allocation.

Raking Ratio (Iterative Proportional Fitting)

Raking is a poststratification method that may be used when poststrata are formed using more than one variable, but only the marginal population totals are known. It is also known as the **iterative proportional fitting** approach, is the process of fitting marginal population total estimates to the known marginal population iteratively. It is also called as **"raking",** “**rim weighting”** and “**marginal calibration”** in the literatüre.

In the implementation of the method;

- ✓ The consistency of the sample row totals and the population row totals is ensured.
- ✓ The corrected sample column totals and population column totals are consistent.
- ✓ The first process is repeated for the sample row totals that are corrupted.
- ✓ Afterwards, the second process is repeated for the sample column totals that are corrupted. The process is repeated until the population totals converge.

Raking Ratio (Iterative Proportional Fitting)

$N_{rc}^{(0)} = \widehat{N}_{rc} \; \forall \; r = 1,2,\dots,R; c = 1,2,\dots,C$

$$N_{rc}^{(2t-1)} = N_{rc}^{(2t-2)} \frac{N_{r.}}{\sum_c N_{rc}^{(2t-2)}}, r = 1,2,\dots,R; c = 1,2,\dots,C$$

$$N_{rc}^{(2t)} = N_{rc}^{(2t-1)} \frac{N_{.c}}{\sum_c N_{rc}^{(2t-1)}}, r = 1,2,\dots,R; c = 1,2,\dots,C$$

Rows and columns are adjusted to population totals by iterative operations, respectively, with each operation denoted by t= 1, 2, 3, ...

Here, in practice, the sum of the design weights is generally used for its $\widehat{N}_{rc}$ estimation. The total estimator for the target variable, with population estimates $N_{rc}^{(F)} = \widetilde{N}_{rc}$ at the final stage of convergence:

$$\hat{t}_y^{IPF} = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{\widetilde{N}_{rc}}{n_{rc}} \sum_{j=1}^{n_{rc}} y_{rcj}$$

EXAMPLE

Let’s say SESRIC have telephone numbers of 1.000 people which is our population. We have no other information without the total age-gender cross-table given below. We randomly select 100 telephone number and dial-up to ask whether they are satisfied or not for SESRIC activities. We also get their gender and age groups which is given below. There is no nonresponse and results show that 15 of the young-male are satisfied for SESRIC activities. Please, **estimate for the population** how many percent of young-male are satisfied for SESRIC activities? (By HT Estimator, Poststratified Estimator and Raking Ratio)

Population

	Male	Female	Total
Young	245	215	460
Middle age	140	140	280
Old	110	150	260
Total	495	505	1000

Sample

	Male	Female	Total
Young	24	16	40
Middle age	15	19	34
Old	12	14	26
Total	51	49	100

Considering that the samples were selected with the SRS, the design weight was calculated for all units

$d = N/n = 10;$

HT

POST

Estimation of the population total of young men : $N_{11}^{(0)} = \widehat{N}_{11} = 24 * 10 = 240.$

Known population total of young people : $N_{1.} = 460.$

Young people's total estimate from the sample : $\sum_c N_{1.}^{(0)} = 400$

$$N_{11}^{(1)} = N_{11}^{(0)} \frac{N_{1.}}{\sum_c N_{1.}^{(0)}} = 240 \frac{460}{400} = 276$$

Similarly, $N_{12}^{(1)}$, $N_{21}^{(1)}$,… calculated.

$$N_{11}^{(2)} = N_{11}^{(1)} \frac{N_{.1}}{\sum_r N_{.1}^{(1)}} = 276 \frac{495}{519{,}53} = 262{,}97$$

When the repetition process is continued, it is seen that the population row and column totals are approached, that is, they converge to the marginal distributions. Assuming the necessary compliance was achieved in the 4th iteration, the procedure was interrupted. The values obtained as a result of the process are given below:

4th Iteration

	Male	Female	Total
Young	264,77	195,15	459,92
Middle age	116,67	163,38	280,05
Old	113,56	146,47	260,03
Total	495,00	505,00	1000

References

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. John Wiley & Sons.

Metin, C. B.(2020) Örneklemede ağırlıklandırma prosedürleri ve kalibrasyon yaklaşımı.

Ardilly, P., & Tillé, Y. (2006). *Sampling methods: Exercises and solutions*. Springer Science & Business Media.