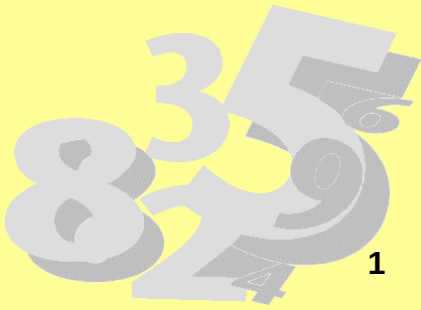
Statistiska centralbyrån Statistics Sweden

4
Traditional estimators as special cases of the calibration approach



The family of calibration estimators

includes many

‘traditional estimator formulas’

Let us look at some examples.

An advantage of the calibration approach:

We need not any more think in terms of ‘traditional estimators’ with specific names.

All of the following examples are special cases of the calibration approach,
corresponding to simple formulations
of the auxiliary vector $\mathbf{x}_k$

The simplest auxiliary vector

$$\mathbf{x}_k = \mathbf{x}_k^* = 1 \quad \text{for all } k$$

The corresponding information is weak :

$$\sum_U \mathbf{x}_k = \sum_U 1 = N$$

Calibrated weights (by the general formula) :

$$w_k = d_k \times \frac{N}{\sum_r d_k}$$

$$\hat{Y}_W = N\,\bar{y}_{r;d} = N\frac{\sum_r d_k y_k}{\sum_r d_k} = \hat{Y}_{EXP}$$

known as the ***Expansion estimator***

Notation for weighted means and other weighted quantities:

Example: $\bar{y}_{r;d} = \dfrac{\sum_r d_k y_k}{\sum_r d_k}$

r is the set of elements for the computation

d is the weighting

The simplest auxiliary vector

$$\mathbf{x}_k = \mathbf{x}_k^* = 1$$

In particular, for SRS (n sampled from N); m out of n respond

$$w_k = \frac{N}{n}\frac{n}{m} = \frac{N}{m}$$

sampling NR adjustment

The simplest auxiliary vector

$$\mathbf{x}_k = \mathbf{x}_k^* = 1 \quad \text{for all } k$$

$$\Rightarrow \quad \hat{Y}_W = \hat{Y}_{EXP} = N\,\bar{y}_{r;d}$$

- a weak auxiliary vector: $\mathbf{x}_k = 1$

  it recognizes no differences among elements

- the bias is usually large

One can show, for any sampling design,

$$\text{bias}(\hat{Y}_{EXP}) / N \approx \bar{y}_{U;\theta} - \bar{y}_U$$

Note the
difference between two means :

The ***theta-weighted mean*** $\bar{y}_{U;\theta} = \dfrac{\sum_U \theta_k y_k}{\sum_U \theta_k}$

The ***unweighted mean*** $\bar{y}_U = \dfrac{\sum_U y_k}{N}$

The bias of the expansion estimator

The ***theta-weighted population mean***
can differ considerably from
the ***unweighted population mean***,
(both of them unknown),
so **bias** can be very large.

These means differ considerably when y
and θ have high correlation.

Comment on the Expansion Estimator

Despite an often large nonresponse bias, the ***expansion estimator*** is (surprisingly enough) often used by practitioners and researchers in social science.

This practice cannot be recommended.

The classification vector ("gamma vector")

Elements classified into P dummy-coded groups

$$\gamma_k = (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk})'$$

$$= (0,\ldots,1,\ldots,0)'$$

The only entry '1'
identifies the group (out of P possible ones)
to which element k belongs

The classification vector

Typical examples:

- Age groups
- Age groups by sex (complete crossing)
- Complete crossing of >2 groupings
- Groups formed by intervals
  of a continuous *x*-variable

The classification vector

as a star vector

$$\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k = (0,\ldots,1,\ldots,0)'$$

The associated information :

The vector of population class frequencies

$$\sum_U \mathbf{x}_k^* = (N_1,\ldots, N_p,\ldots, N_P)'$$

Calibrated weights (by the general formula) :

$$w_k = d_k \times \frac{N_p}{\sum_{r_p} d_k} \quad \text{for all } k \text{ in group } p$$

The classification vector

as a star vector : $\mathbf{x}_k = \mathbf{x}_k^* = \boldsymbol{\gamma}_k$

The calibration estimator takes the form

$$\hat{Y}_W = \sum_{p=1}^{P} N_p \bar{y}_{r_p;d} \quad = \hat{Y}_{PWA}$$

known as the

Population Weighting Adjustment estimator

Population Weighting Adjustment estimator

A closer look :

with

$$\hat{Y}_{PWA} = \sum_{p=1}^{P} N_p \bar{y}_{r_p;d}$$

$$\bar{y}_{r_p;d} = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k}$$ = weighted group y-mean for respondents

N_p = known group count in the population

The classification vector

as a moon vector

$$\mathbf{x}_k = \mathbf{x}_k^{\circ} = \gamma_k = (0,\ldots,1,\ldots,0)'$$

Information for calibration :

the unbiasedly ***estimated*** class counts

$$\hat{N}_p = \sum_{s_p} d_k, \quad p = 1,2,\ldots, P$$

The general formula gives the weights

$$w_k = d_k \times \frac{\sum_{s_p} d_k}{\sum_{r_p} d_k} \quad \text{for all } k \text{ in group } p$$

The classification vector

as a moon vector : $\mathbf{x}_k = \mathbf{x}_k^{\circ} = \boldsymbol{\gamma}_k$

In particular for SRS sampling :

$$w_k = \frac{N}{n}\frac{n_p}{m_p} \quad \text{for all } k \text{ in group } p .$$

Design weight

NR adjustment by inverse of group response rate

The classification vector

as a moon vector

$$\mathbf{x}_k = \mathbf{x}_k^{\circ} = \gamma_k = (0,\ldots,1,\ldots,0)'$$

$$\hookrightarrow \hat{Y}_W = \sum_{p=1}^{P} \hat{N}_p \bar{y}_{r_p;d} = \hat{Y}_{WC}$$

known as

Weighting Class estimator

Weighting Class estimator

$$\hat{Y}_{WC} = \sum_{p=1}^{P} \hat{N}_p \bar{y}_{r_p;d}$$

Class sizes not known but estimated: $\hat{N}_p = \sum_{s_p} d_k$

$$\bar{y}_{r_p;d} = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k} = \text{weighted group } y\text{-mean for respondents}$$

A continuous x-variable

for example, x_k = income ; y_k = expenditure

$$\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)' \qquad \text{Info}: \quad \sum_U \mathbf{x}_k = (N, \sum_U x_k)'$$

The (simple) Regression Estimator

$$\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)'$$

calibrated weights given by :

$$d_k v_k = d_k \times N \left(\frac{1}{\sum_r d_k} + \frac{\bar{x}_U - \bar{x}_{r;d}}{\sum_r d_k (x_k - \bar{x}_{r;d})^2} (x_k - \bar{x}_{r;d}) \right)$$

The calibration estimator takes the form

$$\hat{Y}_W = N \left(\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})\, B_{r;d} \right) = \hat{Y}_{REG}$$

regression coefficient

The (simple) Regression Estimator

A closer look :

$$\hat{Y}_{REG} = N\left(\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})\, B_{r;d}\right)$$

with

$$\bar{x}_{r;d} = \sum_r d_k x_k \,/\, \sum_r d_k$$

$\bar{y}_{r;d}$ analogous y-mean

$$B_{r;d} = \frac{\sum_r d_k (x_k - \bar{x}_{r;d})(y_k - \bar{y}_{r;d})}{\sum_r d_k (x_k - \bar{x}_{r;d})^2}$$

regression of y on x

Combining
a classification and a continuous x-variable

Information about ***both***

(i) the classification vector

$$\boldsymbol{\gamma}_k = (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk})'$$

$$= (0,\ldots,1,\ldots,0)'$$

and

(ii) a continuous variable with value x_k

Known group counts ***and*** group totals for a continuous variable

The vector formulation :

$$\mathbf{x}_k = \mathbf{x}_k^* = (\boldsymbol{\gamma}_k', \; x_k \boldsymbol{\gamma}_k')'$$

$$(\gamma_{1k}, \ldots, \gamma_{pk}, \ldots, \gamma_{Pk}, x_k\gamma_{1k}, \ldots, x_k\gamma_{pk}, \ldots, x_k\gamma_{Pk})'$$

Information for $p = 1, \ldots, P$: N_p and $\sum_{U_p} x_k$

gives the ***SEPREG*** (separate regression) estimator

The Separate Regression Estimator

$$\hat{Y}_W = \sum_{p=1}^{P} N_p \left\{ \bar{y}_{r_p;d} + \left(\bar{x}_{U_p} - \bar{x}_{r_p;d} \right) B_{r_p;d} \right\} = \hat{Y}_{SEPREG}$$

Marginal counts for a two-way classification

P groups for classification 1 (say, age by sex)

H groups for classification 2 (say, profession)

$$\mathbf{x}_k = \mathbf{x}_k^* =$$

$$= (\gamma_{1k},\ldots,\gamma_{pk},\ldots,\gamma_{Pk},\delta_{1k},\ldots,\delta_{hk},\ldots,\delta_{H-1,k})'$$

$$= (0,\ldots,1,\ldots,0 \;\; 0,\ldots,1,\ldots,0)'$$

For computation, vector dimension = $P + H - 1$

Note : -1 necessary to invert the matrix

Gives the two-way classification estimator

Marginal counts for a two-way classification

Note : We are calibrating
on the marginal counts of the P-groups
(classification 1)
and
on the marginal counts of the H-groups
(classification 2)

We say that the classifications are
"in the + relationship" (not "in the × relationship")

The formula for the two-way classification estimator is not a simple one.

List of ‘traditional estimators’

(We shall refer to them later.)

Expansion (EXP)
Weighting Class (WC)
Population Weighting Adjustment (PWA)
Regression (REG)
Separate Regression (SEPREG)
Two-Way Classification (TWOWAY)

Comment : No need to give individual names to the traditional estimators.
All are calibration estimators.
For example, although known earlier as 'regression estimator',

$$\hat{Y}_{REG} = N\,\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})\,B_{r;d}$$

is now completely described as the calibration estimator for the vector $\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)'$

Note: Our list does not include the ratio estimator for nonresponse

$$\sum_U x_k \frac{\sum_r d_k y_k}{\sum_r d_k x_k}$$

It is not a good choise, compared with REG.