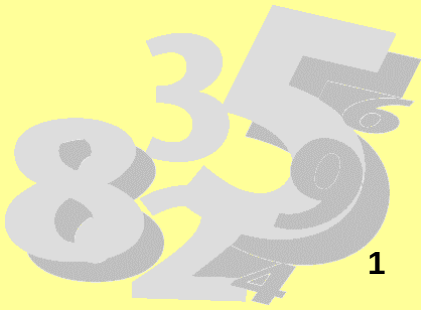
Statistiska centralbyrån Statistics Sweden

# 7
# Bias indicators



# The calibration estimator

is based on a specified auxiliary vector $\mathbf{x}_k$

Intuitively, a better aux. vector leads to smaller bias , smaller variance .

- How do we analyze this in more depth?
- How do we construct the aux. vector ?
- We may have access to ***many*** aux. variables; how do we choose ?
- Primary objective here : reduce bias !

This session and the next are based on two articles :

C.E. Särndal and S. Lundström (2008): Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24, 251-260 .

Same authors (2009): Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Statistics Sweden R&D Report* 2009:1; to appear, *Survey Methodology Journal*.

We consider the estimator $\tilde{Y}_W = \sum_r d_k m_k y_k$

with weights calibrated to the level of the sample :

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k = \begin{pmatrix} \sum_s d_k \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$$

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$$

$\mathbf{x}_k^*$ star vector (population info) known $k \in s$

$\mathbf{x}_k^\circ$ moon vector (sample info) known $k \in s$

with the weighting factor

$$m_k = \underbrace{(\sum_s d_k \mathbf{x}_k)'}_{\text{row vector}} \underbrace{(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}}_{\text{inverted matrix}} \underbrace{\mathbf{x}_k}_{\text{col. vector}}$$

The value $m_k$, defined for $k \in s$, depends

- on the sampling design through $d_k$
- on the outcome $s$ of the sampling
- on the outcome $r$ of the response phase
- on the choice of aux. vector $\mathbf{x}_k$

$$\hat{Y}_W = \sum_r d_k v_k y_k \qquad \text{calibrated to} \qquad \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$$

$$\tilde{Y}_W = \sum_r d_k m_k y_k \qquad \text{calibrated to} \qquad \begin{pmatrix} \sum_s d_k \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$$

They have the *same nearbias* .

An auxiliary variable $x$ is

*equally important* for controlling the bias

when it is of « the star type » $\mathbf{x}_k^*$

as when it is of « the moon type » $\mathbf{x}_k^\circ$

We assume the following background:

A number of potential auxiliary variables
are at our disposal : age, sex, occupational group,
and perhaps many others

Our objective: From these variables (star or moon),
build an efficient auxiliary vector,
likely to give low NR bias

We need tools for comparing different **x**-vectors

Objective : Compare different **x**-vectors

for the calibration estimator $\tilde{Y}_W = \sum_r d_k m_k y_k$

A "worst possible scenario" : We have no other choice than ***the trivial x-vector***

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

This case is useful as a ***basis of comparison***

The trivial aux. vector $\mathbf{x}_k = 1 \quad \text{for all} \quad k$

The weights are $m_k = 1/P \quad \text{for all} \quad k$

$$P = \frac{\sum_r d_k}{\sum_s d_k}, \quad \text{the (weighted) response rate}$$

The estimator is $\tilde{Y}_W = \hat{N}\, \bar{y}_{r;d} = \hat{Y}_{EXP}$

with $\text{nearbias}(\hat{Y}_{EXP}) = N(\bar{y}_{U;\theta} - \bar{y}_U)$

which can be very large.

Another basis of comparison is
the case of full response :

$y_k$ available for $k \in s$

Estimator : $\tilde{Y}_{FUL}$

for ex. $\tilde{Y}_{FUL} = \hat{Y}_{HT}$ without bias

or $\tilde{Y}_{FUL} = \hat{Y}_{GREG}$ almost without bias

In the presence of non-response,
these estimators are *hypothetical*, not computable.

$\tilde{Y}_{FUL}$ full response, thus hypothetical, (almost) unbiased

$\tilde{Y}_{EXP}$ trivial $\mathbf{x}-$vector

$\tilde{Y}_{W}$ calibration on a (much) better $\mathbf{x}-$vector

Define $relbias = \dfrac{\text{nearbias}(\tilde{Y}_{W})}{\text{nearbias}(\tilde{Y}_{EXP})}$

The *bias ratio* is defined as an estimate of *relbias*. It is computable for a fixed survey outcome $(s\ ,\ r)$ , assumed quite large

$$\text{bias ratio} \ = \ \frac{\tilde{Y}_W - \tilde{Y}_{FUL}}{\tilde{Y}_{EXP} - \tilde{Y}_{FUL}}$$

Hence, for fixed outcome $(s, r)$ we have

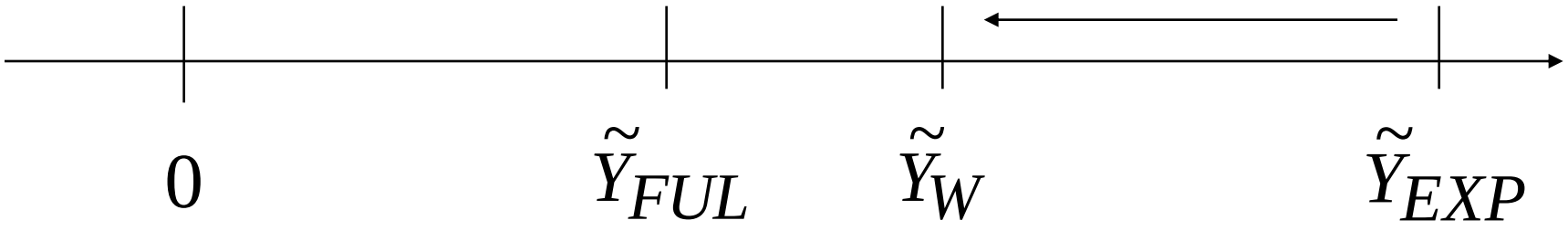
$$\text{bias ratio} = \frac{\tilde{Y}_W - \tilde{Y}_{FUL}}{\tilde{Y}_{EXP} - \tilde{Y}_{FUL}} = 1 - \frac{\tilde{Y}_{EXP} - \tilde{Y}_W}{\tilde{Y}_{EXP} - \tilde{Y}_{FUL}}$$

$\tilde{Y}_{EXP} - \tilde{Y}_W$ computable

but $\tilde{Y}_{FUL}$ unknown so bias ratio unknown

Objective: Choose **X**-vector to get a ***large distance*** $\left|\tilde{Y}_{EXP} - \tilde{Y}_W\right|$

Interpretation, for fixed $r$ and $s$ :

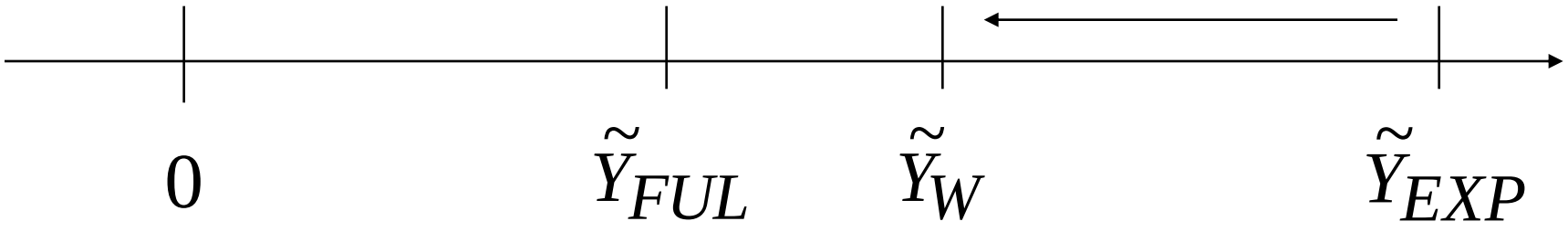



The tendency when **x** more and more powerful :

$\tilde{Y}_W$ ***moves away*** from $\tilde{Y}_{EXP}$ (bad; very biased)

in the direction of $\tilde{Y}_{FUL}$ (good, unbiased)

For fixed $r$ and $s$ :




$\tilde{Y}_W$ moves away from $\tilde{Y}_{EXP}$

The standardized distance $\dfrac{\left|\tilde{Y}_{EXP} - \tilde{Y}_W\right|}{\hat{N} \times S_y}$ smaller and smaller

We denote it by $H_1$ and use it as a *bias indicator*

$S_y = S_{y|r;d} = \text{standard deviation of } y$

## Bias indicator

$$H_1 = \frac{\left|\tilde{Y}_{EXP} - \tilde{Y}_W\right|}{\hat{N} \times S_y}$$ is computable

gets larger as the **x**-vector improves

Objective : Construct **x**-vector to make $H_1$ large

What is a typical value of $H_1 = \dfrac{\left|\tilde{\tilde{Y}}_{EXP} - \tilde{\tilde{Y}}_W\right|}{\hat{N} \times S_y}$ ?

As little as 10% is "a good value" for $H_1$

$$\Rightarrow \quad \frac{\tilde{\tilde{Y}}_W}{\hat{N}} = \frac{\tilde{\tilde{Y}}_{EXP}}{\hat{N}} - 0.10 \times S_y$$

We have moved away 0.1 stand. dev. from the poor estimate

Suppose $H_1 = \dfrac{\left|\tilde{Y}_{EXP} - \tilde{Y}_W\right|}{\hat{N} \times S_y} = 0.1$

so that $\dfrac{\tilde{Y}_W}{\hat{N}} = \dfrac{\tilde{Y}_{EXP}}{\hat{N}} - 0.10 \times S_y$

It can be a ***very large move*** compared with $\dfrac{S_y}{\sqrt{n}}$

for ex. $n = 40{,}000 \Rightarrow \dfrac{S_y}{\sqrt{n}} = \dfrac{S_y}{200} = 0.005 \times S_y$

## The weighting factors $m_k$ in $\tilde{Y}_W = \sum_r d_k m_k y_k$

$$m_k = \underbrace{(\sum_s d_k \mathbf{x}_k)' \sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}}_{\textbf{row vector}} \quad \underbrace{\mathbf{x}_k}_{\text{column}}$$

For given survey outcome $(s\ ,\ r)$

$m_k$ is computable for all $k \in s$ ,

and $m_k$ has a *mean*, a *variance*,

a *correlation* with $y_k$ , and other characteristics,

computable on $r$ , or on $s$

Expression 1 (not shown here) for

the stand. distance $H_1 = \dfrac{\left|\tilde{Y}_{EXP} - \tilde{Y}_W\right|}{\hat{N} \times S_y}$

We have $H_1 = cv_m \times \left|R_{y,m}\right|$

$cv_m$ = coeff. of variation of $m$

$R_{y,m}$ = correl. coeff. between $y$ and $m$

For ex. $R_{y,m} = -0.5$, $cv_m = 0.2 \Rightarrow H_1 = 0.1$

## Properties of $cv_m$

- The trivial aux. vector $\mathbf{x}_k = 1$ gives $cv_m = 0$

- Typical range for other **x**-vectors:

$$0.1 < cv_m < 0.6$$

- When new variables are added to the aux. vector, the effect is ***an increase in*** $cv_m$ (compare $R^2$ in regression analysis).

By Expression 1 , bias indicator $H_1 = cv_m \times \left| R_{y,m} \right|$

The relation between $y$ and $\mathbf{x}$ –
does it play a role ?
$y$ and $\mathbf{x}$ more or less correlated -
what is the effect ?
We expect higher correlation
to decrease the bias

Expression 2 (not shown here) for the stand. distance

We have $H_1 = cv_m \times R_{y,\mathbf{x}} \times \left| R_{D,C} \right|$

where

$R_{y,\mathbf{x}}$ = coef. multiple corr. between $y$ and $\mathbf{x}$

$R_{D,C}$ also interpretable as a regression coefficient

So $0 \le R_{y,\mathbf{x}} \le 1$ ; $\left| R_{DC} \right| \le 1$ and usually $cv_m \le 0.6$

$$H_1 = cv_m \times R_{y,\mathbf{x}} \times \left| R_{D,C} \right|$$

for ex. $H_1 = 0.3 \times 0.8 \times 0.5 \;\; = \;\; 12\%$

In our experience, $H_1$ seldom $>$ 20%

For ex. compare vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ ;

suppose $\mathbf{x}_1$ gives $H_1 = 6\%$

and that $\mathbf{x}_2$ gives $H_1 = 12\%$

Then the adjustment with $\mathbf{x}_2$ is - 0.12 stand dev. :

$$\frac{\tilde{Y}_W}{\hat{N}} = \frac{\tilde{Y}_{EXP}}{\hat{N}} - 0.12 \times S_y$$

with $\mathbf{x}_2$ it is only $-0.06$ stand. dev.

The standardized difference

$$H_1 = cv_m \times R_{y,\mathbf{x}} \times \left| R_{D,C} \right|$$

Properties when we add a new variable $x$

to the **x**-vector :

- $\text{cv}_m$ increases
- $R_{y,\mathbf{x}}$ increases
- $|R_{D,C}|$ does not necessarily increase, but may be fairly constant

$\Rightarrow$ $H_1$ does not necessarily increase

# Building the auxiliary vector

Suppose a supply of *x*-variables is available for the survey. *Our task* : Build an efficient aux. vector from this supply.

Options :

- Use all available *x*-variables
- Use a selection of *x*-variables

  Selection can be
  - by stepwise forward inclusion
  - by stepwise backward elimination

# Building the auxiliary vector

- Stepwise forward

  Start with the trivial vector $\mathbf{x}_k = 1$;
  add one *x*-variable at a time

- Stepwise backward

  Start with all available *x*-variables ;
  eliminate one at a time

The procedure must be based on a
*criterion for selection* at each step

# Criteria for the selection of *x*-variables

Three possibilities are :

1. $H_1 = cv_m \times \left|R_{y,m}\right| = cv_m \times R_{y,\mathbf{x}} \times \left|R_{D,C}\right|$
2. $H_2 = cv_m \times R_{y,\mathbf{x}}$
3. $H_3 = cv_m$

Advantage of $H_3$ : Independent of $y$ .

$H_1$ and $H_2$ are « tailor-made » for a specific $y$ .

We use $H_3$ (preferred) and $H_1$

## The procedure

« stepwise forward selection »

Specify first the criterion $H$ to be used for the selection

Step 0 : Trivial vector $\mathbf{x}_k = 1$ all $k \in U$

Step 1 : Chose "the best"

Step 2 : Chose "the best", given the Step 1variable

and so on

## Stepwise forward

Start with the trivial vector $\mathbf{x}_k = 1$ (Step 0); add one $x$-variable at a time

Step 1. Compute the criterion $H$ for all vectors of the form $(1, x_k)$, where $x_k$ is one of the available $x$-variables. If there are $J$ available $x$-variables, we get $J$ values of $H$. Keep the $x$-variable that gives the largest of these values.

## Stepwise forward

Step 2. Given Step 1, compute the $J - 1$ new values of $H$; add (to the vector in the preceding step) the variable with the largest value of $H$.

and so on, in Steps 3, 4, …

(A test of the significance of the change can perhaps be developed.)

Note 1 : In practice, the $x$-variables
are often categorical .
That is, an $x$-variable with 5 categories
(mutually exclusive and exhaustive)
implies a vector of dimension 4.

So in effect 4 variables enter at the same time.
One category is deleted to avoid singularity.

Session 8 illustrates the stepwise forward
procedure using the criteria $H_3$ and $H_1$ .

Note 2 : The value of the criterion $H_3$ *increases* in every step forward.

This is not necessarily the case with the criterion $H_1$
But our algorithm is such that the selection
is in every step based on the largest value of $H_1$.