Statistiska centralbyrån Statistics Sweden

5
Analysing the bias remaining in the calibration estimator

Important to try to reduce the bias ?

Most of us would say YES, OF COURSE.

A (pessimistic) argument for a NO :
There is no satisfactory theoretical solution; the bias cannot be estimated.
It is always unknown (because the response probabilities unknown).
But we must strive to ***reduce the bias***.
We describe methods for this.

Calibration is not a panacea.

No matter how we choose the aux. vector, the calibration estimator (or any other estimator) will always have a remaining bias .

The question becomes : How do we reduce the remaining bias ?

Answer: Seek ever better $\mathbf{x}_k$

We need procedures for this search (Sessions 6, 7 and 8).

Improved auxiliary vector

will (usually) lead to

reduced bias , reduced variance

Interesting quantities are :

(a) the ***mean squared error***

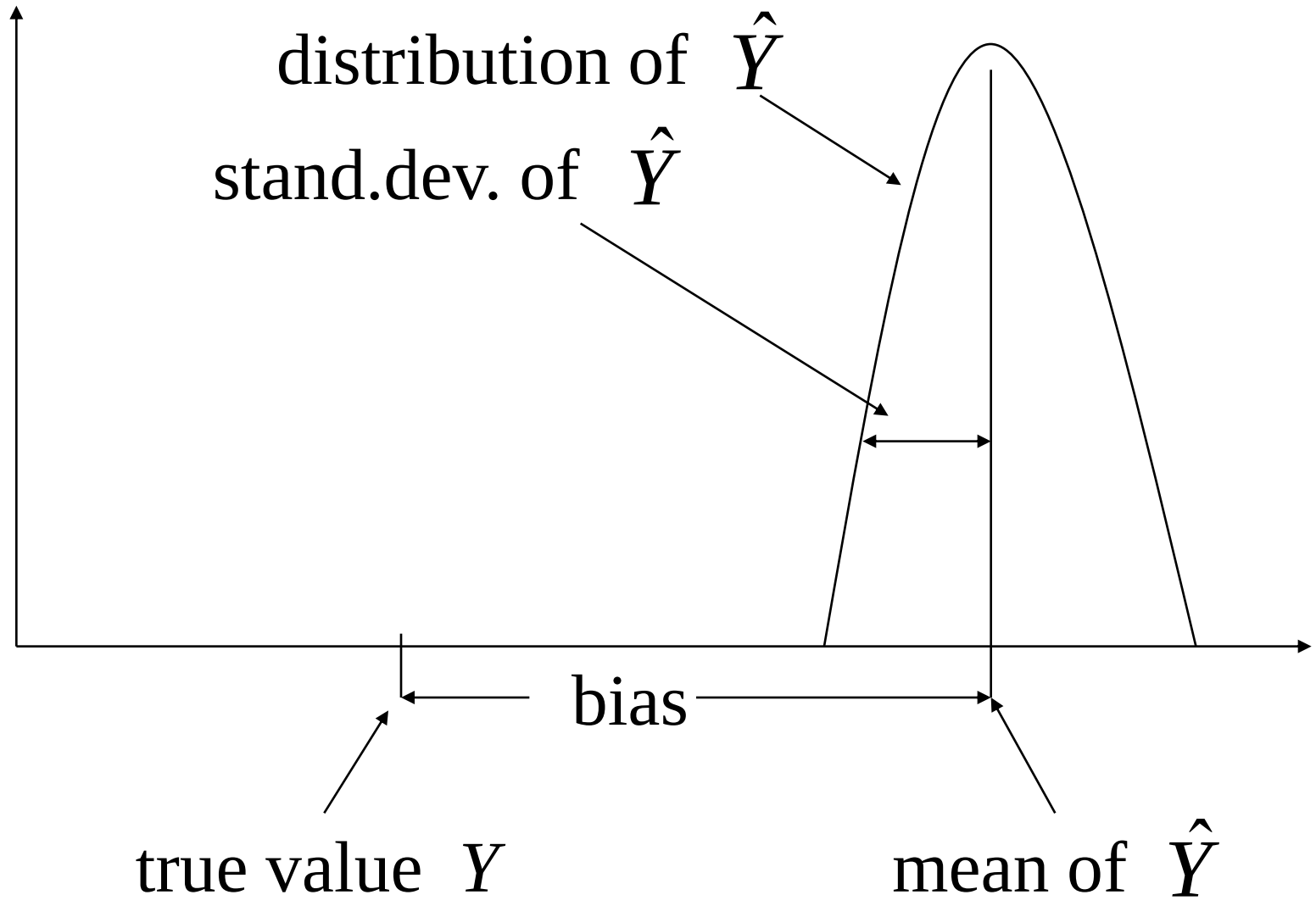
$$\text{MSE} = (\text{Bias})^2 + \text{Variance}$$

and

(b) ***proportion of MSE due to squared bias***

$$(\text{Bias})^2 / \{(\text{Bias})^2 + \text{Variance}\}$$

A bad situation : bias > stand. dev.



Bad situation : squared bias represents a large portion of the MSE

$\Rightarrow$ the interval

$$\hat{Y} \pm 1.96 \times \sqrt{\hat{V}(\hat{Y})}$$

estimated stand.dev.

will almost certainly not contain the unknown value Y for which we want to state valid 95% confidence limits.

We know :

Variance

is often small (and tends to 0)

compared to

squared bias (does not tend to 0)

Both bias and variance are theoretical quantities (expectations), stated in terms of values for the whole finite population

Variance can be estimated, but not the bias .

The bias of the calibration estimator

Recall the general definition :

bias =

expected value of estimator

minus

value of parameter under estimation

What is '***expected value***' in our case ?

Before examining the bias in a general way (arbitrary sampling design, arbitrary aux. vector)

let us consider a simple example .

Example: The simplest auxiliary vector

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

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

Weighted respondent mean, expanded by N

We have seen

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

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

$$\bar{y}_U = \frac{1}{N} \sum_U y_k$$ simple unweighted mean

An alternative expression

$$\text{bias}\,(\hat{Y}_{EXP} / N) \approx r_{y\theta} \times cv(\theta) \times S_{yU}$$

where $r_{y\theta}$ is the correlation coeff. between y and θ,

$cv(\theta) = S_{\theta U} / \bar{\theta}_U$ the coeff. of variation of θ

and S_{yU} the stand. dev. of y in U

Suppose the correlation $r_{y\theta}$ between y and θ is **0.6**.

Then

$$\text{bias}(\hat{Y}_{EXP} / N) \approx 0.6 \times cv(\theta) \times S_{yU}$$

If the response probabilities θ do not vary at all, then $cv(\theta) = 0$ and

$$\text{bias}(\hat{Y}_{EXP} / N) \approx 0$$

As long as all elements have the same response prob. (perhaps considerably < 1), $\hat{Y}_{EXP}$ has no bias .

But consider a small value of $cv(\theta)$:

$$cv(\theta) = 0.1$$

Then

$$\text{bias}(\hat{Y}_{EXP} / N) \approx 0.6 \times 0.1 \times S_{yU} = 0.06\, S_{yU}$$

This bias may not seem large, but the crucial question is : How serious is it compared with

$$\text{stand.dev}(\hat{Y}_{EXP} / N) \quad \textbf{?}$$

$$\mathrm{Var}(\hat{Y}_{EXP}/N) \approx \frac{1}{m} S_{yU}^2$$

(a crude approximation; SRS sampling assumed)

Suppose $m = 900$ responding elements

$$\mathrm{stand.dev}(\hat{Y}_{EXP}/N) \approx 0.033\, S_{yU}$$

compared with the much larger

$$\mathrm{bias}(\hat{Y}_{EXP}/N) \approx 0.06\, S_{yU}$$

Then

$$(\text{Bias})^2 / [(\text{Bias})^2 + \text{Variance}] =$$

$$(0.06)^2 / [(0.06)^2 + (1/900)] =$$

$$0.0036/(0.0036 + 0.0011) = \mathbf{77\ \%}$$

Impossible then to make valid inference by confidence interval !

We return to the

General calibration estimator

For a specified ***auxiliary vector*** $\mathbf{x}_k$

with corresponding ***information*** $\mathbf{X}$,

let us evaluate its bias.

(Derivation, see the book.)

The Calibration Estimator : Its bias

$$\hat{Y}_W = \sum_r w_k y_k$$

with

$$w_k = d_k v_k = d_k (1 + \boldsymbol{\lambda}_r' \mathbf{x}_k)$$

$$\boldsymbol{\lambda}_r' = \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$$

Calibration estimator
close approximation to its bias

$$\text{bias}(\hat{Y}_W) \approx \quad \text{nearbias}(\hat{Y}_W)$$

where

$$\text{nearbias}(\hat{Y}_W) = \quad -\sum_U (1-\theta_k) e_{\theta k}$$

with $$e_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_{U;\theta}$$

$$\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k \mathbf{x}_k y_k$$

$$\text{nearbias}(\hat{Y}_w) = -\sum_U (1-\theta_k)\, e_{\theta k}$$

It is a general formula :

- valid for any sampling design
- valid for any aux. vector
- it is a close approximation (verified in simulations)
- does not depend on the sampling design (the d_k)

$e_{\theta k} = y_k - \mathbf{x}_k' \mathbf{B}_{U;\theta}$ looks like a regression residual

Under a simple condition on $\mathbf{x}_k$:

$$\text{nearbias}(\hat{Y}_W) = \left(\sum_U \mathbf{x}_k\right)' \left(\mathbf{B}_{U;\theta} - \mathbf{B}_U\right)$$

Shows **nearbias** as a function of the difference between two different regression coefficients .

Interpretation: NR causes systematic error in the estimated regression relationship (reason: 'non-random selection'). We would like to estimate the ordinary regression coefficient $\mathbf{B}_U$, but what we can obtain is an estimate of $\mathbf{B}_{U;\theta}$, which is unfortunately affected (biased) by the NR .

Comments

- Detailed derivation of **nearbias**, see the book
- For given auxiliary vector, **nearbias** is the same for any sampling design, but depends on the (unknown) response prob's
- **nearbias** is a function of certain regression residuals (not the usual regression residuals)
- The **variance** does depend on sampling design

Comments

- The nearbias formula makes no distinction between "star variables" and "moon variables"

- In other words, for bias reduction, an *x*-variable is *equally important* when it carries info to the pop. level (included in $\mathbf{x}_k^*$) as when it carries info ***only*** to the sample level (included in $\mathbf{x}_k^\circ$)

Surprising conclusion, perhaps.

But for variance, the distinction can be important.

Example: Let x_k be a continuous aux. variable

- Info at *population level* : $\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)'$

$\Rightarrow$ N and $\sum_U x_k$ known

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

- Info at *sample level* only : $\mathbf{x}_k = \mathbf{x}_k^\circ = (1, x_k)'$

$$\Rightarrow \hat{N} = \sum_s d_k \text{ and } \sum_s d_k x_k \text{ computable}$$

$$\Rightarrow \hat{Y}_W = \hat{N}\{\bar{y}_{r;d} + (\bar{x}_{s;d} - \bar{x}_{r;d})\hat{B}_{r;d}\}$$

where $\bar{x}_{s;d} = \sum_s d_k x_k / \hat{N}$

The two estimators differ, but same **nearbias** .

- Can nearbias be zero? (Would mean that the calibration estimator is almost unbiased.)

Answer : Yes .

- Under what condition(s) ***?***

Answer : There are 2 conditions, each sufficient to give **nearbias** = 0.

- Can we expect to satisfy these conditions in practice ?

Answer: No. We can try to reduce the bias.

Conditions for **nearbias = 0**

In words : $\text{nearbias}(\hat{Y}_W) = 0$

under either of the follwing conditions:

Condition 1 : The influence ϕ has
perfect linear relation to the aux. vector

Condition 2 : The study variable y has
perfect linear relation to the aux. vector

One can show that nearbias = 0 under each of the following conditions

Condition 1: $\phi_k = \lambda' \mathrm{x}_k \quad \text{for all} \quad k \in U$

$$\Rightarrow \quad \left(\sum_U \mathbf{x}_k\right)' \left(\mathbf{B}_{U;\theta} - \mathbf{B}_U\right) = 0$$

and $\quad \mathbf{nearbias} = 0$

Condition 2 : $y_k = \beta' \mathrm{x}_k \quad \text{for all} \quad k \in U$

$$\Rightarrow \quad \mathbf{B}_{U;\theta} = \mathbf{B}_U \quad \text{and} \quad \mathbf{nearbias} = 0$$

Condition 1

nearbias = 0 (for any *y*-variable) if the influence ϕ has perfect linear relation to the auxiliary vector :

$$\text{nearbias}\,(\hat{Y}_W) = 0 \quad \text{if, for all } k \text{ in } \boldsymbol{U},$$

$$\phi_k = \frac{1}{\theta_k} = 1 + \boldsymbol{\lambda}' \mathbf{x}_k$$

for some constant vector $\boldsymbol{\lambda}$

Comments :

1. The requirement $\phi_k = 1 + \boldsymbol{\lambda}' \mathbf{x}_k$ must hold for all $k \in \boldsymbol{U}$.
2. It is ***not a model***. (A model is something you assume as a basis for a statistical procedure.) It is a population property.
3. It requires the influence to be linear in $\mathbf{x}_k$
4. If it holds, **nearbias** = 0

Condition 2

nearbias = 0 if the specific study variable y has ***perfect*** linear relation to the aux. vector

$$\text{nearbias}\,(\hat{Y}_W) = 0 \quad \text{if, for all} \quad k \in \text{U},$$

$$y_k = \boldsymbol{\beta}' \mathbf{x}_k$$

for some constant vector $\boldsymbol{\beta}$

Condition 2

Note :

$$y_k = \boldsymbol{\beta}' \mathbf{x}_k \quad \text{for all} \quad k \in U$$

is ***not a model***.

It is a population property saying that

nearbias = 0

if the y-variable has perfect linear relation to the aux. vector.

Comment

We have found that

$$\text{nearbias}\,(\hat{Y}_W) = 0$$

1. if the influence ϕ has ***perfect*** linear relation to the aux. vector

2. if the *y*-variable has ***perfect*** linear relation to the aux. vector .

Comment

There are many *y*-variables in a survey :

- One for every socio-economic concept measured in the survey
- One for every domain (sub-population) of interest

To have nearbias = 0 for the whole survey requires that ***every one*** of the many *y*-variables must have perfect linear relation to the auxiliary vector.

Not easy (or impossible) to fulfill.

Comment

Therefore,
the first condition is the more important one

If satisfied, then $\text{nearbias}\,(\hat{Y}_W) = 0$

for ***every one*** of the many *y*-variables

Can the statistician

control

the remaining bias ?

make nearbias smaller ?

Can the bias be controlled ?

We would like to ***come close*** to ***one or both*** of :

1. the influence ϕ has ***perfect*** linear relation to the aux. vector
2. every *y*-variable of interest has ***perfect*** linear relation to the aux. vector

We propose diagnosic tools called ***indicators*** (Sessions 7 and 8).

Questions that we shall consider in the following sessions :

What aux. vector should we use?
How do we evaluate different choices of aux. vector ?