Statistiska centralbyrån Statistics Sweden

8
Bias indicators, continued

Two bias indicators:

$$(i)\ \ H_3 = cv_m$$

$$(ii)\ \ H_1 = \left|R_{y,m}\right| \times cv_m$$

H_3 is computed from the values $\mathbf{x}_k$; $k \in s$.
It does not depend on the y-variable.

H_1 depends on both y and $\mathbf{x}$

Let us examine the properties of H_1 and H_3 by use of a simulation study.

Monte Carlo simulation

Population of size $N = 832$, derived from Statistics Sweden's **KYBOK** survey (see Session 6).

Information: For every $k \in U$, we know

- membership in one of 4 admin. groups
- the value of a continuous variable

  x = sq.root revenues

Study variable: y = expenditures

Monte Carlo simulation

We used two response distributions, called:
(1) Logit
(2) Increasing exponential

Average response prob.: 86% (for both)

Response probability θ increases with x and with y

Corr. between y and θ :
$\approx$ *0.70* (logit) ; $\approx$ *0.55* (incr. exp.)

Monte Carlo simulation

Measures computed as averages over 10,000 repetitions (s, r) ; size of every s : $n = 300$

$$AveH_3 = \text{Average of } H_3 \times 10^3$$

$$AveH_1 = \text{Average of } H_1 \times 10^3$$

$$RelBias = 100\,[Ave(\hat{Y}_W) - Y]/Y$$

$$Ave(\hat{Y}_W) = \sum_{j=1}^{10,000} \hat{Y}_{W(j)} / 10,000$$

Estimators defined in Session 4

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

REG and SEPREG should be used with caution. We recommend to categorize continuous variables.

Estimators that use categorical auxiliary variables:

EXP	The benchmark estimator
WC	The weighting class estimator , where the classes are the four different types of clergical municipalities
WC2	The weighting class estimator , where the classes are four groups defined by size of *x*-value
TWOWAY	The estimator using the classes in WC and in WC2 (in the + manner; calibration on the margins)

Paths of increasing information:

We go from EXP to WC to TWOWAY

or

from EXP to WC2 to TWOWAY

Response distribution: Logit

Path: EXP to WC to TWOWAY

Estimator	$AveH_3$	$AveH_1$	*RelBias*
EXP	0.0	0.0	5.0
WC	42.1	16.2	2.2
TWOWAY	62.4	26.7	0.4

Response distribution: Logit

Path: EXP to WC2 to TWOWAY

Estimator	$AveH_3$	$AveH_1$	*RelBias*
EXP	0.0	0.0	5.0
WC2	48.4	24.3	0.8
TWOWAY	62.4	26.7	0.4

Here, both $AveH_3$ and $AveH_1$ order the estimators in a correct way with respect to *RelBias*.

Estimators that treat x as a continuous aux. variable, as in REG and SEPREG.

We go from EXP to REG to SEPREG.

Then, for these data, we will see that H_1 does not order REG and SEPREG correctly.

Response distribution: Logit

Estimator	$AveH_3$	$AveH_1$	*RelBias*
EXP	0.0	0.0	5.0
REG	37.3	31.5	-0.5
SEPREG	64.1	30.1	-0.3

In the transition from EXP to REG, we see an ”overadjustment”: relbias goes from positive to negative.

Going from REG to SEPREG (if the required additional info is available), the value of H_1 gives us an indication that ”we have gone too far”: H_1 changes direction (starts to decrease) going from 31.5 to 30.1.

In our experience, an incorrect ordering is more likely to occur when a variable *x* is treated as continuous (as in REG, SEPREG)
than when it is treated as categorized by size of *x* (as in WC2, TWOWAY).

This (and the fact that extreme weights can occur) is why we recommend caution in regard to REG and SEPREG.

(In practice we would not know with certainty that overadjustment has ocurred.)

Response distribution: Increasing exponential

Estimator	$AveH_3$	$AveH_1$	*RelBias*
EXP	0.0	0.0	9.3
WC	48.1	20.8	5.7
WC2	105.6	50.7	0.5
TWOWAY	112.4	51.2	0.4

Estimator	$AveH_3$	$AveH_1$	*RelBias*
EXP	0.0	0.0	9.3
REG	81.5	68.3	-2.6
SEPREG	113.6	58.3	-0.9

A case study

Use of the bias indicators H_3 and H_1 in the Swedish pilot survey on problem gambling

(a telephone interview survey)

Sampling design: STSRS of 2,000 persons (strata: 6 regions × 4 age groups) from Statistics Sweden's Register of the Total Population (RTP)

Overall response rate (unweighted): 50.8 %

(We have written our own programs to compute H_1 and H_3).

Statistics Sweden's RTP and the data base LISA contains many potential auxiliary variables. For example:

Sex, age, income, marital status, nationality, address, type of family, number of children in different age groups, education level, profession, branch of industry, number of days with illness, number of days of unemployment, number of days in early retirement pension, income of capital, and so on

How do we select ?

Preparation:

(i) An initial set of potential auxiliary variables was selected by a subjective procedure

(ii) Aux. variables were used at the sample level (moon variables)

(iii) Continuous variables are used as grouped; all variables used are then grouped.

The use of H_3 as a tool for stepwise forward selection of variables:

- In each step, the auxiliary vector expands by adding the (grouped) variable causing the largest increase in H_3

- Variables enter ”in the + manner” (”the side-by-side” manner”)

An important requirement in practice :

The same set of weights for each *y*-variable.

Recommended steps for constructing the auxiliary vector:

(i) Make an inventory of potential aux. variables

(ii) Categorize the continuous aux. variables

(iii) Use H_3 in a stepwise forward procedure to find an **x**-vector considered to ”work well” for most of the *y*-variables

(iv) Compute H_1 in a stepwise forward procedure for some very important *y*-variable(s)

Recommended steps for constructing the auxiliary vector (cont.):

(v) Construct a ”compromise” **x**-vector

(vi) Compute the calibrated weights with this **x**-vector

(vii) If some weights are negative or ”too large”, some variables in the **x**-vector may have to be dropped.

To illustrate how the indicators work , we let a register variable play the role of a study variable. Therefore the relative deviation from (RDF) can be computed :

$$RDF = (\tilde{Y}_W - \tilde{Y}_{FUL})/\tilde{Y}_{FUL} \times 10^2$$

In this illustration we use y = Employed as the study variable.

Stepwise forward procedure based on H_3

Auxiliary variable entered	$H_3 \times 10^3$
Education level (3)	186
Cluster of postcode areas (6)	250
Country of birth (2)	281
Income class (3)	298
Age class (4)	354
Sex (2)	364
Urban centre dwelling (2)	374
Level of debt (3)	381
Months with sickness benefits (3)	384
Presence of children (2)	387
Marital status (2)	388
Days unemployed (3)	388

Stepwise forward procedure based on H_3

For comparision: $\tilde{Y}_{EXP} \times 10^{-3} = 4719$; $\tilde{Y}_{FUL} \times 10^{-3} = 4265$

Auxiliary variable entered	$H_3 \times 10^3$	$\tilde{Y}_W \times 10^3$	*RDF*
Education level (3)	186	4520	6.0
Cluster of postcode areas (6)	250	4505	5.6
Country of birth (2)	281	4498	5.5
Income class (3)	298	4369	2.4
Age class (4)	354	4399	3.1
Sex (2)	364	4384	2.8
Urban centre dwelling (2)	374	4378	2.6
Level of debt (3)	381	4364	2.3
Months with sickness benefits (3)	384	4380	2.7
Presence of children (2)	387	4379	2.7
Marital status (2)	388	4379	2.7
Days unemployed (3)	388	4377	2.6

Observations :

- Large change in $\tilde{Y}_W$ for the first few entering variables
- Successive increases in H_3 are large in the early steps, then taper off (as expected).
- It seems hardly motivated to go beyond the eighth variable (Level of Debt)

Let us now consider the selection based on H_1.

Stepwise forward procedure based on H_1

For comparision: $\tilde{Y}_{EXP} \times 10^{-3} = 4719$

Auxiliary variable entered	$H_1 \times 10^3$	$\tilde{Y}_W \times 10^{-3}$
Income class (3)	76	4458
Education level (3)	107	4350
Presence of children (2)	114	4326
Urban centre dwelling (2)	118	4310
Sex (2)	123	4296
Marital status (2)	125	4286
Days unemployed (3)	121	4301
Months with sickness benefits (3)	120	4305
Level of debt (3)	115	4322
Cluster of postcode areas (6)	109	4343
Country of birth (2)	103	4363
Age class (4)	99	4377

Stepwise forward procedure based on H_1

For comparision: $\tilde{Y}_{EXP} \times 10^{-3} = 4719$; $\tilde{Y}_{FUL} \times 10^{-3} = 4265$

Auxiliary variable entered	$H_1 \times 10^3$	$\tilde{Y}_W \times 10^{-3}$	*RDF*
Income class (3)	76	4458	4.5
Education level (3)	107	4350	2.0
Presence of children (2)	114	4326	1.4
Urban centre dwelling (2)	118	4310	1.1
Sex (2)	123	4296	0.7
Marital status (2)	125	4286	0.5
Days unemployed (3)	121	4301	0.9
Months with sickness benefits (3)	120	4305	1.0
Level of debt (3)	115	4322	1.3
Cluster of postcode areas (6)	109	4343	1.8
Country of birth (2)	103	4363	2.3
Age class (4)	99	4377	2.6

Observations

• Order of entry of variables is not the same as with H_3 - this is expected.

• The value of H_1 ”turns around”: increaing at first then starts to decrease

• A very low RDF (= 0.5) is obtained after step 6.

The final auxiliary vector chosen for the *Swedish pilot survey on problem gambling* was the following ”compromise” :

Education level + *Clusters of postcode areas* + *Income group* + *Age group* + *Level of debt*

All variables except *Level of debt* are star variables.

Further recommendations

(i) The auxiliary vector should be robust. If there is a pilot study, the client usually prefers the same vector in the main study as in the pilot study.

(ii) The auxiliary vector should be one that is also likely to reduce the variance