Statistiska centralbyrån Statistics Sweden

6
Selecting the most relevant auxiliary information

Auxiliary information can be used both

at the design stage

and

at the estimation stage

The design stage

Commonly used sampling designs

- Simple random sampling (SRS)
- Stratified simple random sampling (STSRS)
- Cluster sampling
- Two-stage sampling
- Probability-proportional-to-size

The estimation stage

Two important steps in building the auxiliary vector:

(i) making an inventory of potential auxiliary variables

(ii) selecting the most suitable of these variables and preparing them for entry into the auxiliary vector

Inventory of potential auxiliary variables

Example of an extensive data source:
Sweden's Total population register (TPR) : A complete listing of the population of individuals (around 9 million)

Some of the variables in TPR:

Unique personal identity number, name and adress, date of birth, sex, marital status, country of birth and taxable income.

Recall:

If the nonresponse is considerable and not counteracted by effective adjustment then

(i) the squared bias term is likely to dominate the MSE

(ii) the possibilies for valid statistical inference are reduced; valid confidence intervals cannot be computed

Guidelines for the construction of an auxiliary vector

Principle 1: The auxiliary vector (or the instrument vector) should explain the inverse response probability, called the response influence (based on Condition 1)

Principle 2: The auxiliary vector should explain the main study variables (based on Condition 2)

Principle 3: The auxiliary vector should identify the most important domains

Principle 1 fulfilled:

The bias of the calibration estimates reduced for ***all*** study variables

Principle 2 fulfilled:

The bias is reduced in the estimates for the main study variables, and the variance is also reduced

Principle 3 fulfilled:

For the main domains, both bias and variance will be reduced

The general formula for the nearbias (Session 5) can guide our search for a powerful auxiliary vector. It also answers the question:

When is the nearbias = 0, for a given estimator ?

Let us look at some traditional estimators.

The **x**-vector is a 'star vector' in most of these examples.

Prospects for zero nearbias with traditional estimators

Expansion estimator: $\hat{Y}_{EXP} = N\ \bar{y}_{r;d}$

Auxiliary vector: $\mathbf{x}_k = 1$

Zero nearbias if

(i) $\phi_k = a$ for all $k \in U$ (Condition 1)

(ii) $y_k = \alpha$ for all $k \in U$ (Condition 2)

Weighting class estimator: $\hat{Y}_{WC} = \sum_{p=1}^{P} \hat{N}_p y r_{p;d}$

Population weighting adjustment estimator:

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

Aux. vector $\mathbf{x}_k = \boldsymbol{\gamma}_k$ = class indicator vector

Moon vector for $\hat{Y}_{WC}$, star vector for $\hat{Y}_{PWA}$

Zero nearbias if

(i) $\phi_k = a_p$ for all $k \in U_p$ or if

(ii) $y_k = \beta_p$ for all $k \in U_p$

Regression estimator:

$$\hat{Y}_{REG} = N\{y_{r;d} + (x_U - x_{r;d})\hat{B}_{r;d}\}$$

Auxiliary vector: $\mathbf{x}_k = (1, x_k)'$

Zero nearbias if

(i) $\phi_k = a + bx_k$ or if

(ii) $y_k = \alpha + \beta x_k$

Separate regression estimator:

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

Auxiliary vector: $\mathbf{x}_k = (\boldsymbol{\gamma}'_k, x_k \boldsymbol{\gamma}'_k)'$

Zero nearbias if

(i) $\phi_k = a_p + b_p x_k$ or if

(ii) $y_k = \alpha_p + \beta_p x_k$ for all $k \in U_p$

Two-way estimator:

$\hat{Y}_{TWOWAY}$ (expression rather complicated)

Auxiliary vector: $\mathbf{x}_k = (\boldsymbol{\gamma}'_k, \boldsymbol{\delta}'_k)'$

$\boldsymbol{\gamma}$ indicates classes p=1,…, P;

$\boldsymbol{\delta}$ indicates classes h=1,…, H

Zero nearbias if

(i) $\phi_k = a_p + b_h$ or if

(ii) $y_k = \alpha_p + \beta_h$

Conclusion: Among those estimators,

Best suited for fulfilling Principle 1:
SEPREG or TWOWAY

Best suited for fulfilling Principle 2:
The same two

Worst : For both Principles: EXP.

Monte Carlo simulation

10,000 SRS samples
each of size $n = 300$ drawn from
experimental population of size $N = 832$,
constructed from actual survey data :
Statistics Sweden's **KYBOK** survey

Elements (clergical municipalities) classified into four administrative groups; sizes: 348, 234, 161, 89

Monte Carlo simulation

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

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

  x = sq.root revenues

We can use some or all of that info.

Study variable: y = expenditures

Monte Carlo simulation

In this experiment, 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

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

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

$$\text{Variance} = \frac{1}{9,999} \sum_{j=1}^{10,000} \left[\hat{Y}_{W(j)} - Ave(\hat{Y}_W)\right]^2 \times 10^{-8}$$

Monte Carlo simulation ; logit response

Estimator	RelBias	Variance
Expansion (EXP)	5.0	69.6
Weighting Class (WC)	2.2	59.4
Population Weighting Adjustment (PWA)	2.2	37.1
Regression (REG)	-0.6	9.5
Separate Regression (SEPREG)	-0.2	8.1
Two-Way Classification (TWOWAY)	0.5	21.7

Monte Carlo simulation ; increasing exp. response

Estimator	RelBias	Variance
Expansion (EXP)	9.3	70.1
Weighting Class (WC)	5.7	57.7
Population Weighting Adjustment (PWA)	5.7	36.3
Regression (REG)	-2.7	8.1
Separate Regression (SEPREG)	-0.8	7.4
Two-Way Classification (TWOWAY)	0.5	20.3

What do we learn from the simulations ?

Bias ↓ when the auxiliary vector 'gets better' (more informative)

Variance also ↓ , as expected

For ex., SEPREG clearly uses much more information than EXP

We want to be more precise about 'informative' This will follow .

The search for a powerful auxiliary vector

Simple traditional tools

Principle 1: Nonresponse analysis

Principle 2: Analysis of important target variables, one at a time.

New and more advanced tools

Indicators H_1 and H_3 that measure the simultaneous effect of several auxiliary variables.

H_3 is a tool to realize Principle 1.

H_1 takes into consideration both Principle 1 and Principle 2.

This is discussed in Sessions 7 and 8.

Simple tool for Principle 1

Nonresponse analysis

Nonresponse analysis

Example: The Survey on Life and Health

(postal survey; Statistics Sweden)

Age group	18-34	35-49	50-64	65-79
Response rate (%)	54.9	61.0	72.5	78.2

Country of birth	Nordic countries	Other
Response rate (%)	66.7	50.8

Income class (in thousands of SEK)	0-149	150-299	300-
Response rate (%)	60.8	70.0	70.2

Marital status	Married	Other
Response rate (%)	72.7	58.7

Education level	Level 1	Level 2	Level 3
Response rate (%)	63.7	65.4	75.6

Conclusions from this nonresponse analysis:

- The response propensities vary quite a lot between groups

- Without adjustment weighting, one can expect a disturbingly large nonresponse bias

- Some of the presumptive auxiliary variables are related, for example, income and education level. What is the simultaneous effect? Should both be used, or just one of them?

We seek an indicator for Principle 1 that gives us information on the simultaneous effect of the auxiliary variables, rather than one aux. variable at a time. This leads us to consider the indicator H_3 described in Session 7.

Simple tool for Principle 2

Analysis of important target variables
(one at a time)

Analysis of important target variables

Example: The Survey on Life and Health

Four important dichotomous study variables (attributes) are :

(a) Poor health

(b) Avoiding staying outdoors after dark

(c) Difficulties in regard to housing

(d) Poor personal finances

Auxiliary variable: Sex

Attribute	Male	Female
(a)	7.5	8.9
(b)	7.8	21.1
(c)	2.6	2.4
(d)	19.6	19.8

Auxiliary variable: Age class

Attribute	18-34	35-49	50-64	65-79
(a)	4.3	6.6	10.6	10.9
(b)	11.8	11.4	14.3	23.4
(c)	5.9	2.8	1.0	0.8
(d)	31.0	26.6	12.5	9.6

Auxiliary variable: Country of birth

Attribute	Nordic countries	Other
(a)	8.0	11.7
(b)	14.7	18.3
(c)	2.4	4.2
(d)	19.2	28.5

Auxiliary variable: Income group (in thousands of SEK)

Attribute	0-149	150-299	300-
(a)	10.0	7.2	4.0
(b)	18.6	12.6	8.1
(c)	3.8	1.5	1.0
(d)	25.3	16.5	6.9

Auxiliary variable: Marital status

Attribute	Married	Other
(a)	8.2	8.2
(b)	13.8	16.3
(c)	1.1	4.3
(d)	14.1	26.5

Auxiliary variable: Education level

Attribute	Level 1	Level 2	Level 3
(a)	10.5	7.3	4.6
(b)	19.1	12.6	12.9
(c)	1.7	3.2	1.8
(d)	17.5	21.6	16.8

Conclusions from the analysis of important target variables: A somewhat mixed pattern :

- Sex important for explaining variable (b)

- Marital status important for variable (d)

- Age class and country of birth important for most of the four variables

- Income group and education level are both important, but seem to give almost the same information

Question arising : What is the simultaneous effect of these aux. variables?

In Session 7 we present the indicator H_1 that measures the simultaneous effect of the variables. It takes into account both Principle 1 and Principle 2.

Rekommendations for simple analysis

(i) Make an inventory of potential aux. variables

(ii) Categorize the continuous aux. variables

(iii) Carry out the simple analysis as shown in the preceding example

(iv) Use all (or a subset) of the aux. variables; compute the calibrated weights. The subset is determined more or less subjectively.

Issues in an analysis of the weights

Some weights too large?

- Could make the estimate for some domains too large

- The variance estimator may deteriorate

Some weights negative?

- Most users dislike negative weights

If some weights are negative or ”too large”, drop the aux. variable that has the least effect.