An Alternative Nonresponse Adjustment Estimator

Nonresponse is difficult

Imbi Traat

UNIVERSITY OF TARTU

June 16-20, 2019

Why nonresponse is difficult?

$$r \subset s \subset U$$

θ_k-response probability

unknown

Much research

No generally holding principles

π_k- inclusion probability

known

Well-established design-based theory

Full response

Horvitz-Thompson (1952)

$$\hat{Y}_{HT} = \sum_s d_k y_k, \qquad d_k = 1/\pi_k$$

- Design-consistent, unbiased under any design, and any y-variable

Deville and Särndal (1992)

Calibration estimator $\hat{Y}_{CAL} = \sum_s d_k g_k y_k$ is

under any design and any y-variable

- asymptotically design-consistent
- approximately unbiased
- with different distance functions asymptotically equivalent
  - all lead to the calibration estimator based on the lineaar method

Nonresponse

$y_k \in s$ not observed for each *k*

$\hat{Y}_{HT}$ and $\hat{Y}_{CAL}$ not possible

Särndal and Lundström (2005): Book of history and current state of the art for nonresponse

Haziza & Lesage (2016): Common approaches

1) Nonresponse propensity weighting (double expansion, 2phase)
2) Two-step approach: 1)+calibration
3) One-step approach: nonresponse calibration weighting

Nonresponse

Many additional requirements

Little & Vartivarian (2005), Beaumont (2005)

- x_k for estimating response probabilities has to be related to both
  - Response indicator
  - Study variable y_k
- If not related to y_k then will not decrease nonresponse bias

Haziza & Lesage (2016)

- Choice of calibration function has strong effect
- Inappropriate calibration function may lead to biased estimator
  - even in the presence of high association between x_k and y_k
  - sometimes with bias larger than that of unadjusted estimator.

Nonresponse

Many additional requirements

Brick (2013): modelling assumes MAR response

Rubin (1976) MAR: $P(I_k = 1|y_k, \mathrm{x}_k) = P(I_k = 1|\mathrm{x}_k)$

Results do not hold for NMAR (non-ignorable, informative) response

Brick (2013): NMAR response cannot be distinguished from MAR response based on observed data.

Nonresponse balance measures

Improvements over simple response rates

- R-indicator (Schouten et al. 2009)
- Imbalance indicator (Särndal 2011)

NB! Measures with respect to x-vector

- Balnaced with respect to x, may be unbalanced regarding y

- Many y-variables in a survey

Known and unknown estimators

from regression perspective

Target: $\bar{y}_s = \sum_{k\in s} d_k y_k \,/\, \sum_{k\in s} d_k$ (unbiased for population mean)

Notation

$$\mathrm{x}_k : J \times 1$$

$$\exists \mu,\ \mu'\mathrm{x}_k = 1, \forall k,$$

$P= \sum_{k\in r} d_k \,/\, \sum_{k\in s} d_k$-response rate

$$\bar{\mathrm{x}}_r = \frac{\sum_{k\in r} d_k \mathrm{x}_k}{\sum_{k\in r} d_k},\ \bar{\mathrm{x}}_s = \frac{\sum_{k\in s} d_k \mathrm{x}_k}{\sum_{k\in s} d_k}$$

$$\Sigma_r = \sum_{k\in r} d_k \mathrm{x}_k \mathrm{x}_k' \,/\, \sum_{k\in r} d_k,\quad \Sigma_s = \sum_{k\in s} d_k \mathrm{x}_k \mathrm{x}_k' \,/\, \sum_{k\in s} d_k.$$

Regression perspective - Response variable is predicted by $b'x_k$ (fitted value), where b by WLSQ

I. Estimating response probability θ_k

Regressing I_k (response indicator) on $b'\mathrm{x}_k$ gives

$$b' = P\,\bar{\mathrm{x}}_r'\,\Sigma_s^{-1},$$

and the fitted value

$$\hat{\theta}_k = P\,\bar{\mathrm{x}}_r'\,\Sigma_s^{-1}\mathrm{x}_k = P f_k,$$

where

$$f_k = \bar{\mathrm{x}}_r'\Sigma_s^{-1}\mathrm{x}_k$$

Double-expansion estimator

$$\bar{y}_{\mathrm{2fh}} = \frac{\sum_{k\in r} d_k y_k \,/\hat{\theta}_k}{\sum_{k\in s} d_k} = \frac{\sum_{k\in r} d_k y_k \,/f_k}{\sum_{k\in r} d_k}$$

Using fitted values for y_k

Regressing y_k on x_k in *s* gives coefficient vector

$$b_s' = \frac{\sum_{k \in s} d_k y_k \mathrm{x}_k'}{\sum_{k \in s} d_k} \Sigma_s^{-1}$$

Not computable

Replacing means in *s* by means in *r:*

I. In both factors $\longrightarrow$ calibration estimator
II. Only in first factor $\longrightarrow$ f-estimator

Replacing means in both factors

$$b'_s = \frac{\sum_{k\in s} d_k y_k \mathrm{x}'_k}{\sum_{k\in s} d_k} \Sigma_s^{-1} \longrightarrow b'_r = \frac{\sum_{k\in r} d_k y_k \mathrm{x}'_k}{\sum_{k\in r} d_k} \Sigma_r^{-1}$$

The respective fitted values $\hat{y}_k = b'_r \mathrm{x}_k$

Mean of fitted values in *s* is calibration estimator

$$\frac{\sum_{k\in s} d_k b'_r \mathrm{x}_k}{\sum_{k\in s} d_k} = b'_r \bar{\mathrm{x}}_s = \frac{\sum_{k\in r} d_k y_k \mathrm{x}'_k}{\sum_{k\in r} d_k} \Sigma_r^{-1} \bar{\mathrm{x}}_s = \frac{\sum_{k\in r} d_k y_k g_k}{\sum_{k\in r} d_k} = \bar{y}_{\mathrm{CAL}},$$

where calibration weight

$$g_k = \mathrm{x}'_k \Sigma_r^{-1} \bar{\mathrm{x}}_s$$

Check calibration property

Replacing mean in the first factor

$$b'_s = \frac{\sum_{k\in s} d_k y_k \mathrm{x}'_k}{\sum_{k\in s} d_k} \Sigma_s^{-1} \longrightarrow b'_{sr} = \frac{\sum_{k\in r} d_k y_k \mathrm{x}'_k}{\sum_{k\in r} d_k} \Sigma_s^{-1}$$

The respective fitted values $\hat{y}_k = b'_{sr}\mathrm{x}_k$

Mean of fitted values in *r* is f-estimator:

$$\frac{\sum_{k\in r} d_k b'_{sr}\mathrm{x}_k}{\sum_{k\in r} d_k} = b'_{sr}\bar{\mathrm{x}}_r = \frac{\sum_{k\in r} d_k y_k \mathrm{x}'_k}{\sum_{k\in r} d_k}\Sigma_s^{-1}\bar{\mathrm{x}}_r = \frac{\sum_{k\in r} d_k y_k f_k}{\sum_{k\in r} d_k} = \bar{y}_{\mathrm{f}},$$

where calibration weight

$$f_k = \mathrm{x}'_k \Sigma_s^{-1}\bar{\mathrm{x}}_r$$

Strange! Compare with $\bar{y}_{2\mathrm{fh}}$

Scaled f-estimator

Särndal et al. (2018):

Mean of g-weights in *r*:

$$\frac{\sum_{k\in r} d_k g_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in r} d_k \mathrm{x}_k' \Sigma_r^{-1} \bar{\mathrm{x}}_s}{\sum_{k\in r} d_k} = \bar{\mathrm{x}}_r' \Sigma_r^{-1} \bar{\mathrm{x}}_s = 1 \text{ (μ property)}$$

Mean of f-weights in *r*:

$$\frac{\sum_{k\in r} d_k f_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in r} d_k \mathrm{x}_k' \Sigma_s^{-1} \bar{\mathrm{x}}_r}{\sum_{k\in r} d_k} = \bar{\mathrm{x}}_r' \Sigma_s^{-1} \bar{\mathrm{x}}_r = 1 + Q_s \geq 1,$$

where $Q_s = (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s)' \Sigma_s^{-1} (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s)$ – imbalance measure

Scaled f-estimator: $\bar{y}_{\mathrm{SCf}} = \frac{\sum_{k\in r} d_k y_k f_k}{(1+Q_s) \sum_{k\in r} d_k}$

Estimators for a mean

Simple (UNW) $\bar{y}_r = \sum_{k\in r} d_k y_k \,/ \sum_{k\in r} d_k$

Calibration $\bar{y}_{CAL} = \sum_{k\in r} d_k g_k y_k \,/ \sum_{k\in r} d_k$

f-estimator $\bar{y}_f = \sum_{k\in r} d_k f_k y_k \,/ \sum_{k\in r} d_k$

SCf-estimator $\bar{y}_{SCf} = \dfrac{\sum_{k\in r} d_k f_k y_k}{(1+Q_s)\sum_{k\in r} d_k}$

Unbiased $\bar{y}_{UNB} = \dfrac{\sum_{k\in r} d_k y_k/\theta_k}{\sum_{k\in s} d_k}$

2-ph $\bar{y}_{2ph} = \dfrac{\sum_{k\in r} d_k y_k/f_k}{\sum_{k\in r} d_k}$

θ_k - resp. probability

Set-up

Real data of Estonian HH survey

$$n = 1000, m = 600$$

$$logit(\theta) = 5 - HD\_sex + 2HD\_active - 0.0003H\_income$$

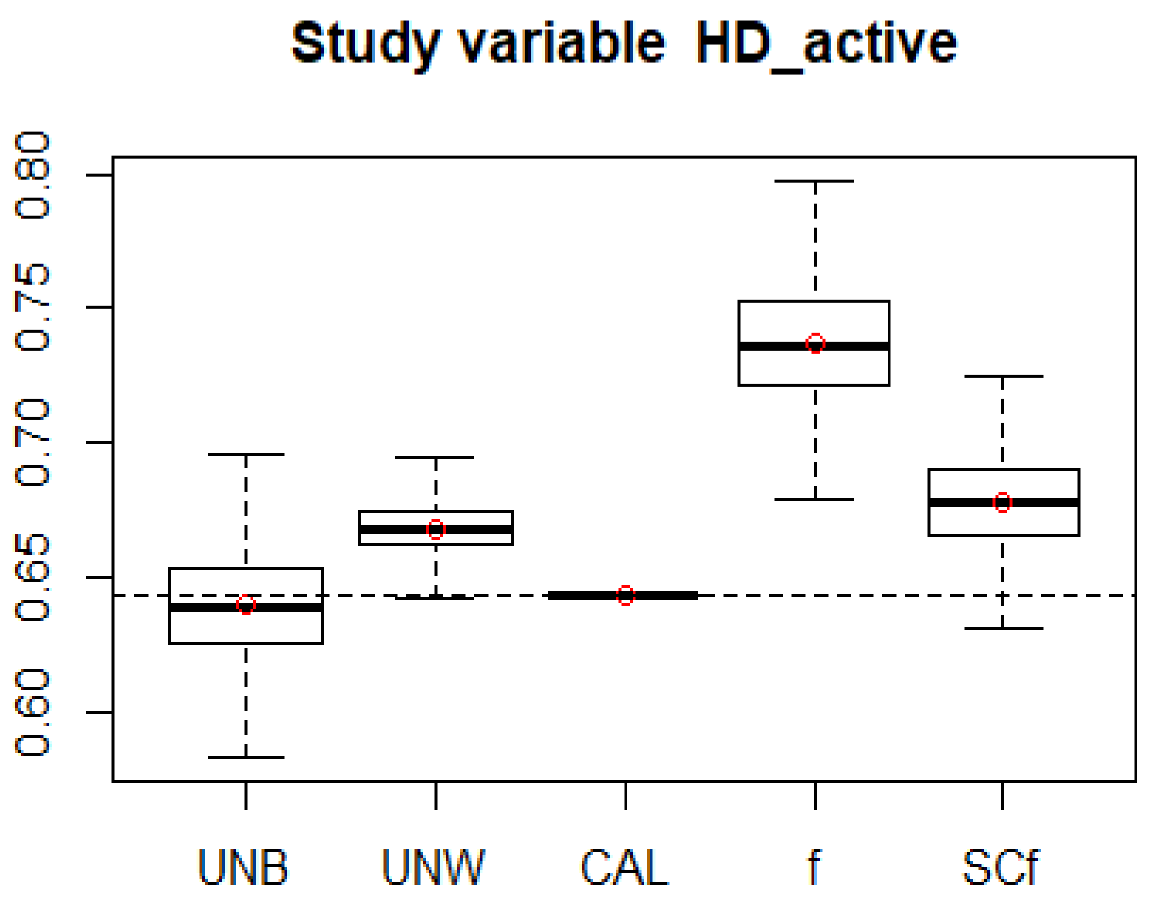
summary(theta)

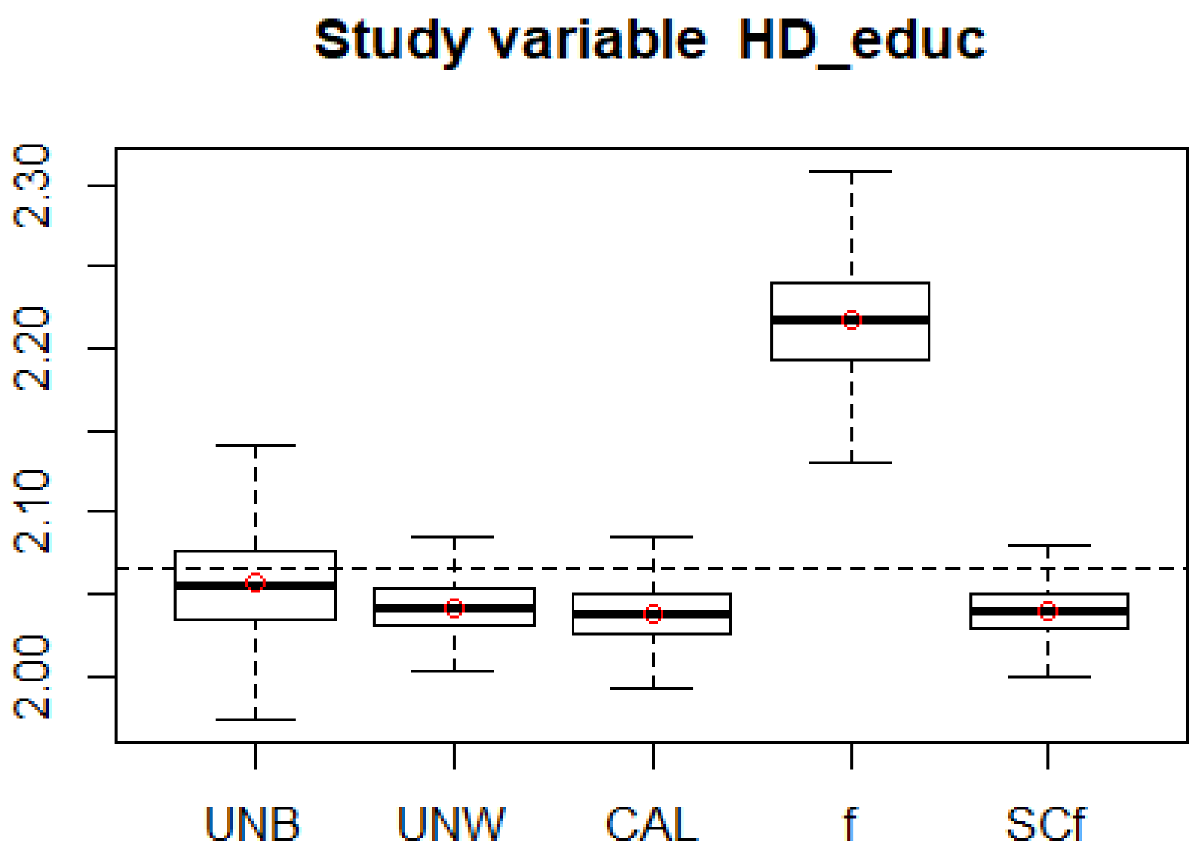
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$0.6\ 10^{-6}$	0.415	0.685	0.600	0.846	0.871

x-vector of dim. 4 $HD\_sex \times HD\_active$

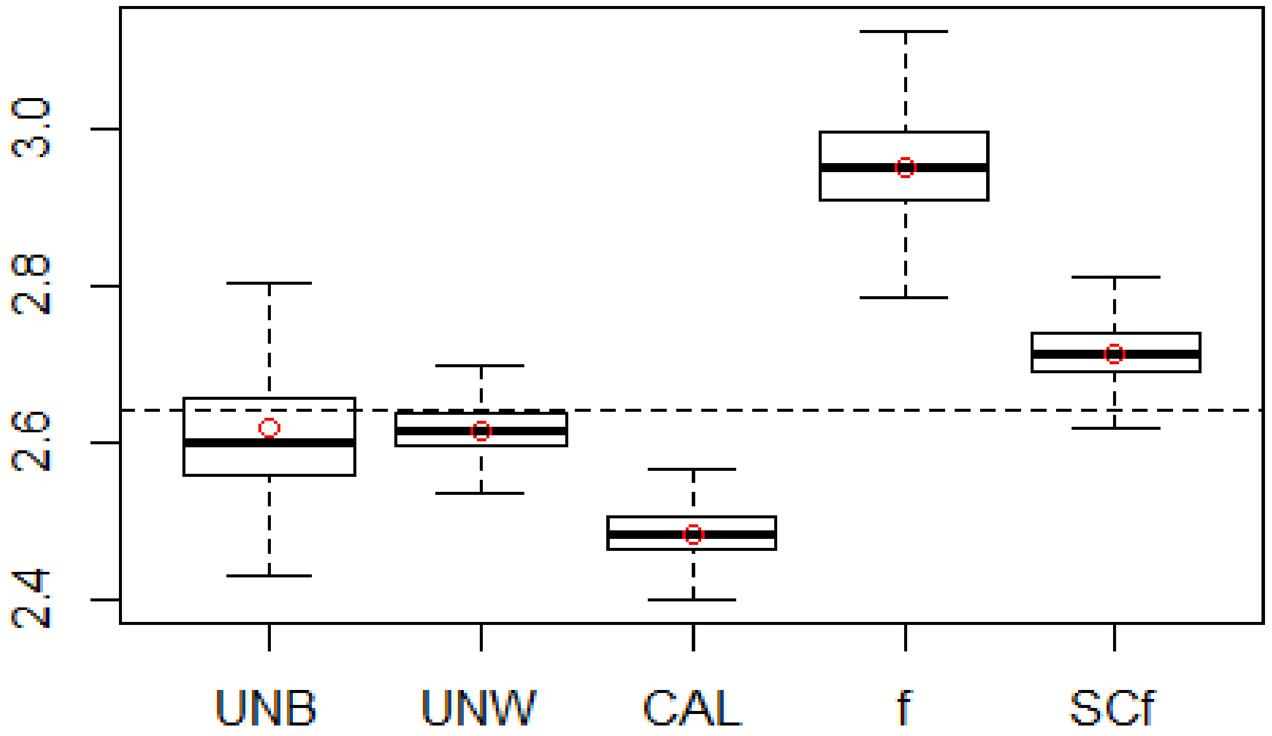
s fixed, 1000 repetitions of *r*

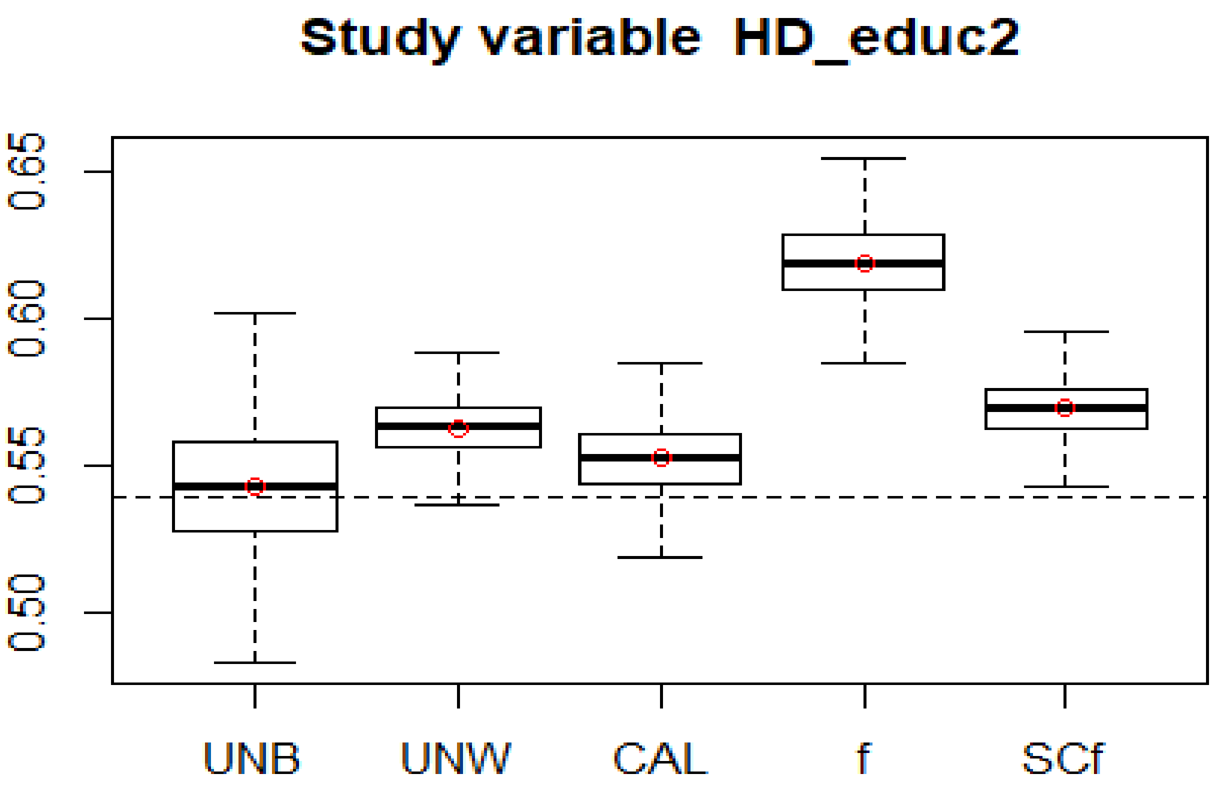
Want to be close to sample mean $\bar{y}_s$


Study variable HD_active
0.60
0.65
0.70
0.75
0.80
UNB
UNW
CAL
f
SCf

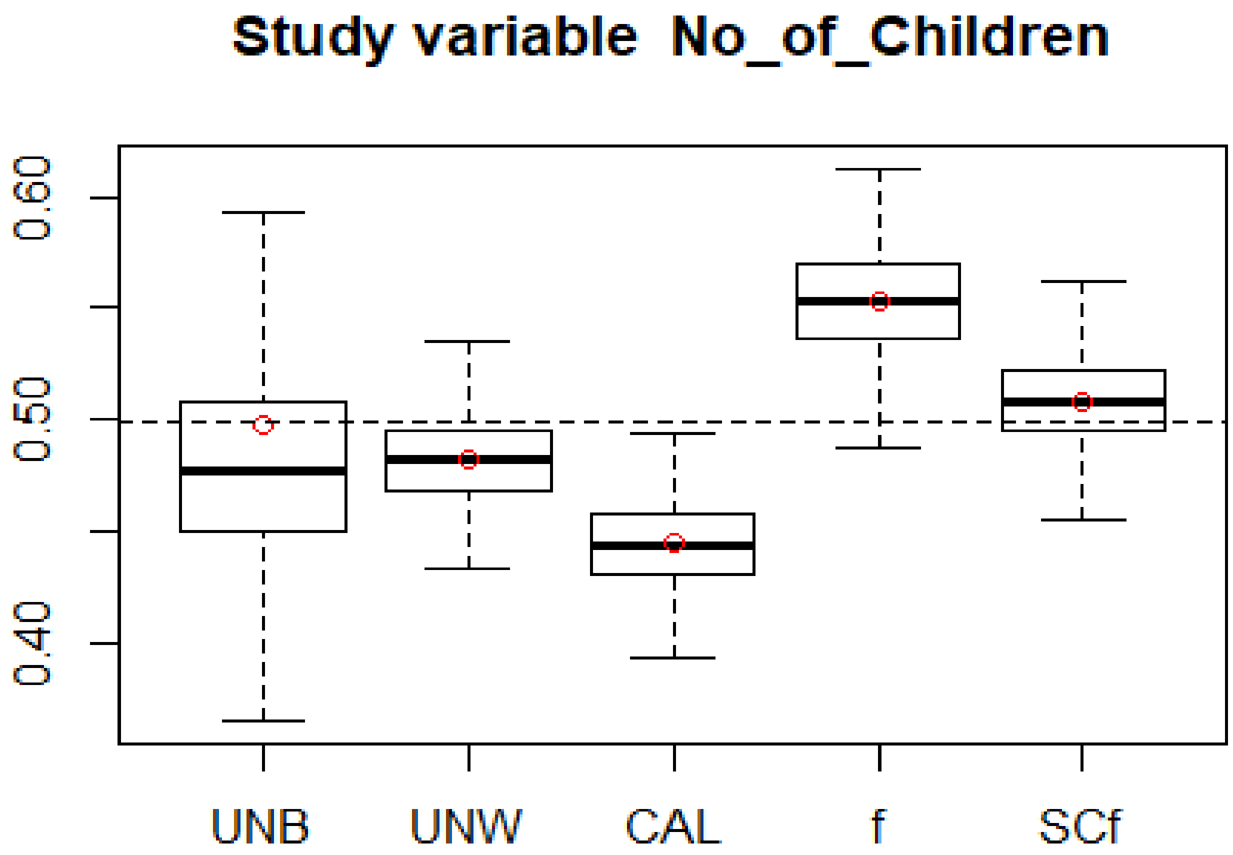

Study variable HD_educ
2.00
2.10
2.20
2.30
UNB
UNW
CAL
f
SCf

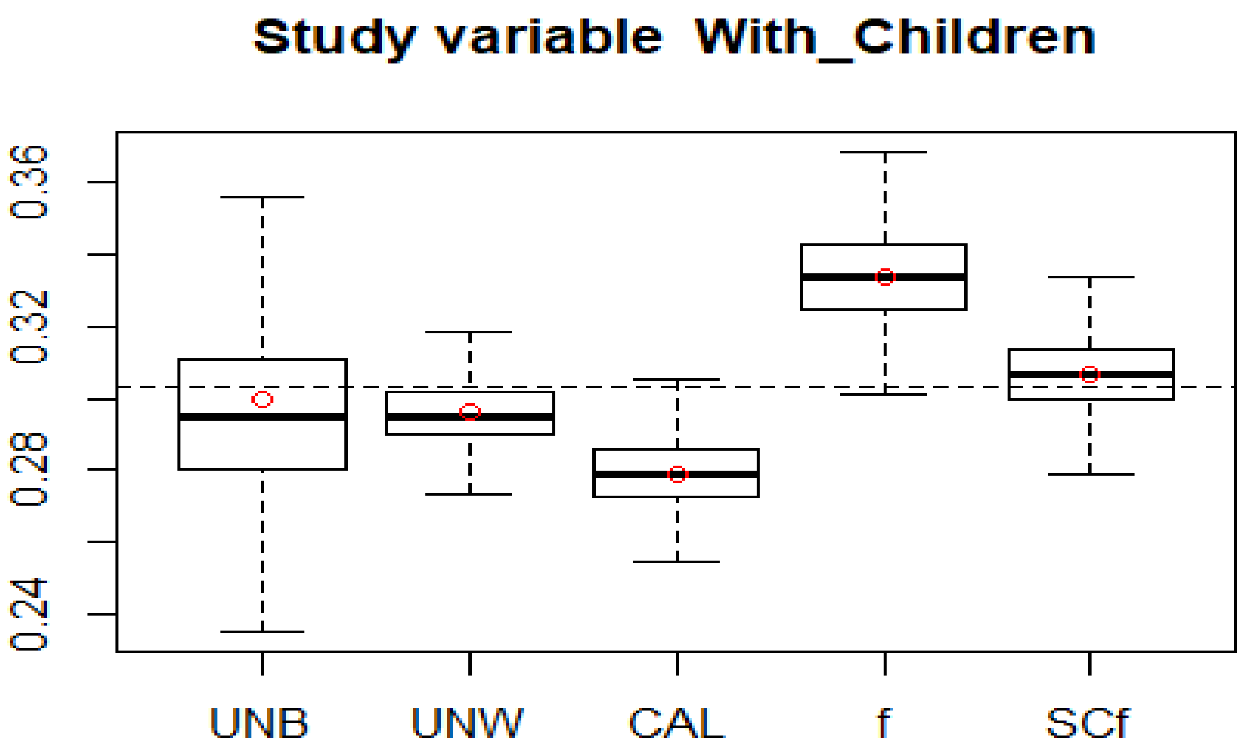
Study variable H_size



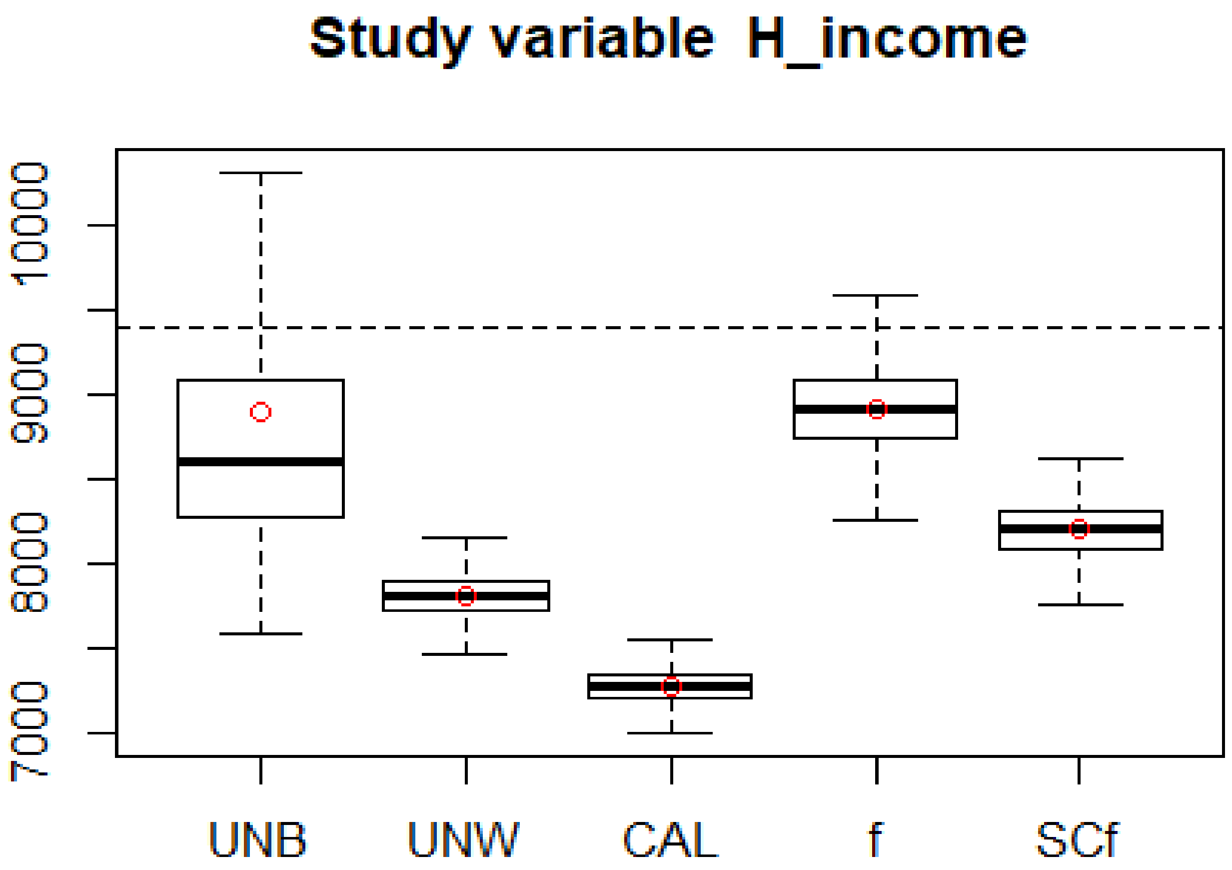


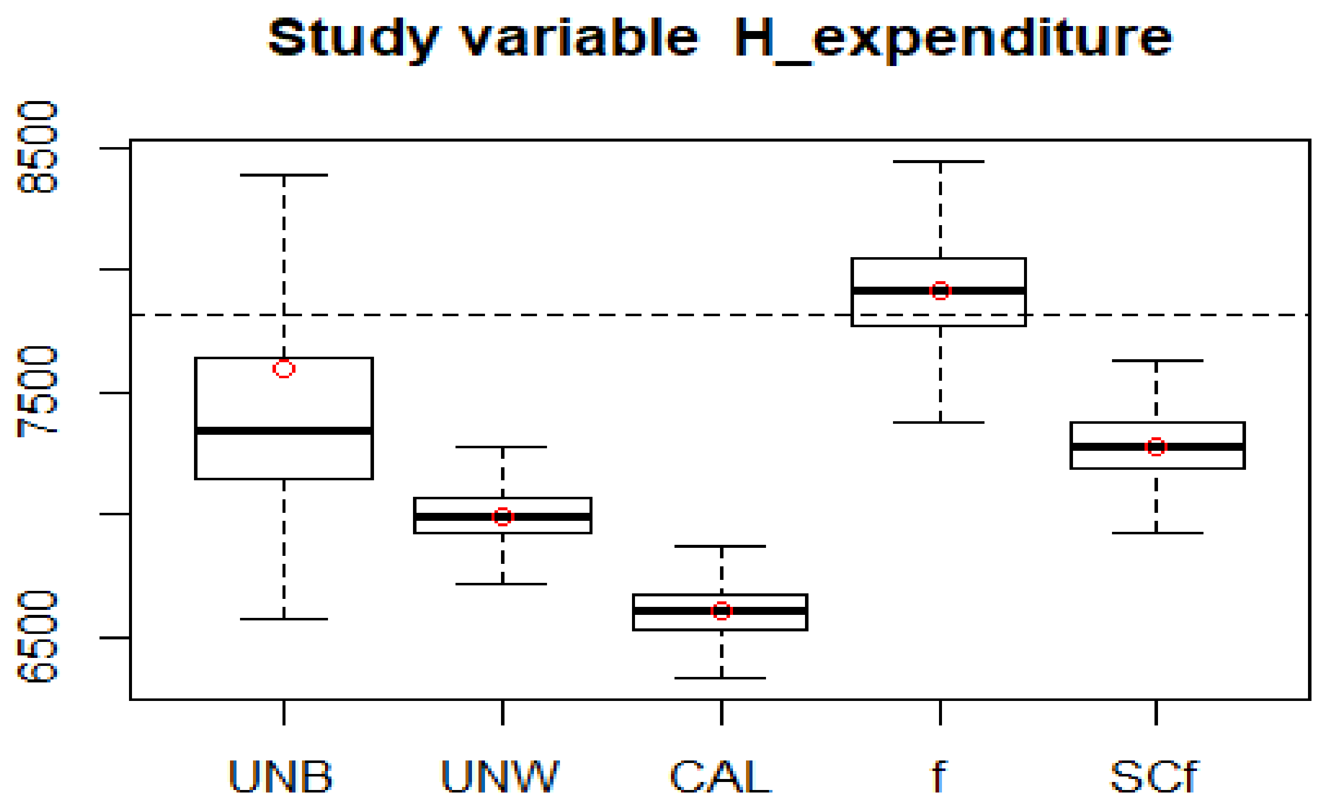
Study variable HD_educ2
0.65
0.60
0.55
0.50
UNB
UNW
CAL
f
SCf


Study variable No_of_Children
0.40
0.50
0.60
UNB
UNW
CAL
f
SCf



Study variable With_Children
0.24
0.28
0.32
0.36
UNB
UNW
CAL
f
SCf


Study variable H_income
10000
9000
8000
7000
UNB
UNW
CAL
f
SCf



Study variable H_expenditure
8500
7500
6500
UNB
UNW
CAL
f
SCf

Absolute Relative Bias - ARB

$$\mathrm{ARB} = \frac{|E_{rep}\hat{\bar{y}} - \bar{y}_s|}{\bar{y}_s},$$

where

$\hat{\bar{y}}$ - our estimator of sample mean $\bar{y}_s$

Absolute Relative bias

	UNB	UNW	CAL	f	SCf
HD_active	0.0047	0.0394	0.0000	0.1466	0.0542
HD_sex	0.0033	0.2902	0.0000	0.4923	0.5325
HD_educ	0.0038	0.0111	0.0129	0.0737	0.0125
H_size	0.0078	0.0092	0.0589	0.1177	0.0281
No_of_Children	0.0036	0.0348	0.1101	0.1078	0.0185
HD_educ1	0.0036	0.0077	0.0262	0.0692	0.0167
HD_educ2	0.0076	0.0443	0.0252	0.1492	0.0569
HD_educ3	0.0143	0.0851	0.0721	0.0258	0.1039
With_Children	0.0098	0.0233	0.0786	0.1009	0.0120
H_big	0.0021	0.0059	0.0565	0.1461	0.0540
H_income	0.0525	0.1691	0.2260	0.0506	0.1270
H_transfer	0.0239	0.1135	0.1296	0.0122	0.0917
H_expenditure	0.0277	0.1047	0.1551	0.0127	0.0682

Average ARB for entire survey

Average over all 11 study variables (2 first were neglected)

UNB	UNW	CAL	f	SCf
0.0142	0.0553	0.0865	0.0787	0.0536

Conclusions

- No uniformly best (unbiased) estimator
  - for all study variables
  - for all response mechanisms

Particular study – NMAR response

- One-step calibration was good only under 1:1, or very strong, relationship between y and x
- SCf was best for the entire survey
- UNW was te second best
- f-estimator was good for the income-related study variables
- Here CAL was the worst

Conclusions

Nonresponse is difficult

Due to missingness it is not possible to

- Evaluate whether response is NMAR or MAR
- Test which estimator is best for particular study variable

References

Beaumont (2005). Survey Methodology, 31, 227-231

Brick (2013). Journal of Official Statistics, 29, 329-353

Deville and Särndal (1992). J. of American Statistical Association, 87, 376-382

Haziza & Lesage (2016). Journal of Official Statistics, 32, 129-145

Horvitz-Thompson (1952). J. of American. Statistical Association, 47, 663-685

Little & Vartivarian (2005). Survey Methodology, 31, 161-168

Rubin (1976). Biometrika, 63, 581-592

Särndal and Lundström (2005). Estimation in Surveys with Nonresponse. Wiley

Särndal (2011). Journal of Official Statistics, 27, 1-21

Särndal, Traat, Lumiste (2018). Statistics in Transition,19,183-200

Schouten, Cobben, Betlehem (2009). Survey Methodology, 35,101-113

UNIVERSITY OF TARTU

1632

Thank you!

unitartu

tartuuniversity