Nonresponse adjustment by design?

Barry Schouten

4th Int Workshop on Advances in adaptive and responsive survey design, November 9 and 10, 2015, Manchester

Research questions

1. Is there a theoretical basis for the assertion that a less representative response for X is indicative of a less representative response for other variables, even after adjustment using X?
2. Is there any empirical basis for this assertion?

In other words, should we prefer more balanced/representative data for bias reduction?

RQ1 asks whether such indicators measure product quality or accuracy.
Schouten (technical paper, 2015)

RQ2 asks whether representativeness indicators measure process quality, i.e. quality control point of view.
Schouten, Cobben, Lundquist & Wagner (JRSSA, 2015, forthcoming)

Empirical evidence

Options to investigate empirical support:

1. Simulate adaptive survey designs with reduced effort based on a real data set with a maximal or very extensive effort;
2. Set aside part of the auxiliary variables and treat those as pseudo survey variables;

In the paper, we chose the second option, but literature has reported a few studies under the first option.

Empirical evidence

Approach:

- Collect survey data sets with multiple designs or waves over various countries/institutes;
- Select available auxiliary variables in the data sets and sort them randomly;
- Add variables one by one, and compute (partial) CV and R and remaining nonresponse bias given adjustment on variables already in the models;
- Rank designs/waves within a data set based on their performance;
- Perform a rank test on the inversions in the design preferences when adding the variables one by one;

Conclusion: On combined data sets, the hypotheses of random inversions are rejected and point at consistency in design preferences.

Theoretical conditions - 1

Some considerations:

- Survey nonresponse is not-missing-at-random (NMAR) for survey variables, unless an informed model would exist;
- The nature of the generation of (auxiliary) variables needs to be modelled, because the possible NMAR universe is very “big”;
- If all NMAR mechanisms are seen as equally likely, then ASD may only be useful to improve precision;

How to set up a framework for generating variables on a population?

Theoretical conditions - 2

Two options to model variable generation in a data set:

1. Enumerate (only conceptually) all variables and define a random selection mechanism that applies to surveys;
2. Construct the population as a grid where measurements collapse strata to (much) smaller numbers, and lead to variables with a much smaller number of categories;

Since collinearity is crucial in extending observed data patterns to possible missing data patterns, option 2 seems inevitable.

As a consequence, a population has two important features: the number of population strata and their relative sizes.

Theoretical conditions - 3

Two basic families of variable generating distributions:

- Uniform: Population strata are allocated randomly to a smaller random number of categories;
- Clustered: There are at least two population strata that are always allocated to the same category;

Theoretical conditions - 4

The two types of distributions allow for strong conclusions about expected coefficients of variation and R-indicators, and about the efficacy of ASD:

- Under uniform grouping, a lower CV for one design than another implies a lower expected CV on any other randomly drawn variable on that design;
- Under clustered uniform grouping, the same holds but for any other randomly drawn variable from the same cluster;
- Both conclusions are still valid when ASD is based on minimizing CV;

Estimating population parameters

When $(X_1, X_2, \dots, X_M)^T$ are uniformly generated, then the population diffusion can be estimated. A possible estimation strategy is through the chi-square statistic between pairs of variables

$$\chi^2_{1,2} = \sum_{k=1}^{C_1} \sum_{l=1}^{C_2} \frac{\left(\frac{G_{k,l}}{G} - \frac{G_{k.}}{G}\frac{G_{.l}}{G}\right)^2}{\frac{G_{k.}}{G}\frac{G_{.l}}{G}},$$

and $E(\chi^2_{1,2}|C_1, C_2) = D\ (C_1 - 1)(C_2 - 1)$.

Under the assumption that the $\hat{\chi}^2_P(m_1, m_2)$ follow independent chi-square distributions for all pairs of variables, then the ML estimator is

$$\widehat{D} = \frac{\sum_{m_1=1}^{M-1} \sum_{m_2=m_1+1}^{M} \hat{\chi}^2_P(m_1,m_2)/n}{\sum_{m_1=1}^{M-1} \sum_{m_2=m_1+1}^{M} (C_{m_1} - 1)(C_{m_2} - 1)}.$$

Discussion

Questions for discussion:

- Sensible to construct theoretical conditions for the efficacy of ASD?
- Is it meaningful to think about the random, possibly clustered, generation of variables?
- The diversity (number of strata) and diffusion (variation in stratum sizes) parameters of a population may be estimated from panel data. Useful?

Other issues:

- Many variables are needed to get a precise signal of a prefered design. The empirical study confirms this requirement;
- Random measurement error leads to spurious diversity;

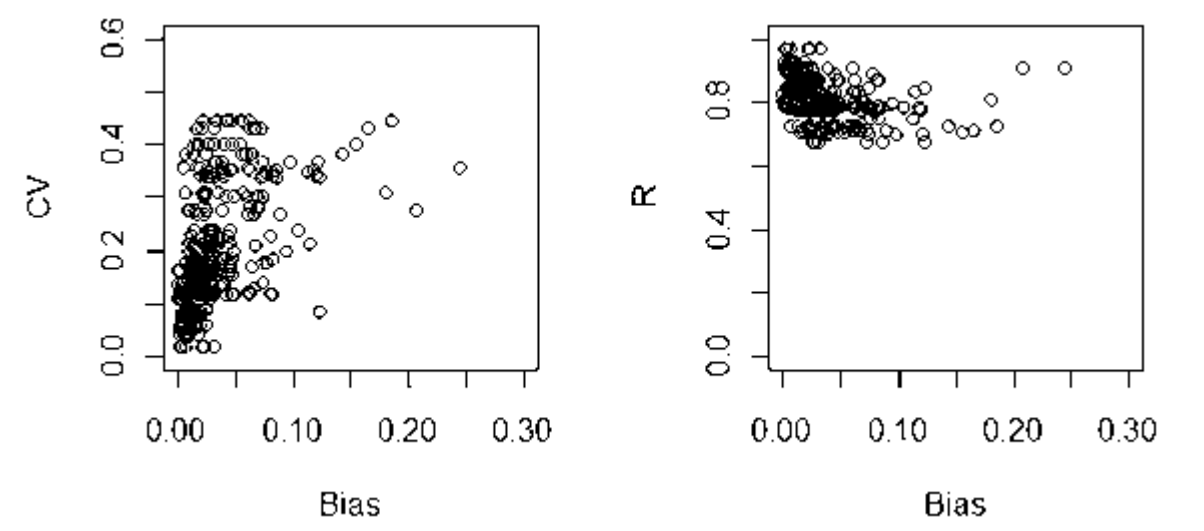
Empirical evidence

Rank test applied to rank inversions in design preferences when adding auxiliary variables one at a time (14 data sets).

	Number of rank inversions		
	Expected	Observed	p-value
Stat Netherlands	189.5	97	<0.001
Stat Sweden/ISR	118.5	97	0.02
All	308	194	<0.001

Empirical evidence

Coefficient of variation and R-indicator plotted against the remaining nonresponse bias after adjustment.



Theoretical evidence

Coefficient of variation for five mode designs in the Crime Victimization Survey for ten auxiliary variables, averaged and for all variables combined.

	1	2	3	4	5	6	7	8	9	10	Av	All
W	0.21	0.14	0.07	0.28	0.07	0.07	0.01	0.18	0.15	0.05	0.12	0.36
M	0.18	0.16	0.05	0.14	0.06	0.06	0.04	0.19	0.05	0.04	0.10	0.29
F	0.09	0.13	0.00	0.00	0.14	0.05	0.01	0.11	0.04	0.13	0.07	0.23
W→F	0.06	0.08	0.01	0.08	0.10	0.08	0.01	0.11	0.06	0.10	0.07	0.18
M→F	0.08	0.09	0.02	0.09	0.05	0.03	0.04	0.10	0.05	0.04	0.06	0.16

Theoretical evidence

Coefficient of variation for five mode designs in the Crime Victimization Survey after selection on Cramèr's V with respect to three survey variables.

Design	$C_V > 0.10$						$C_V > 0.15$		
	Y_1		Y_2		Y_3		Y_1	Y_2	Y_3
	Av	All	Av	All	Av	All	Gender	NA	Age
W	0.06	0.10	-	-	0.15	0.29	0.07	-	0.21
M	0.05	0.08	-	-	0.14	0.24	0.05	-	0.18
F	0.07	0.14	-	-	0.11	0.19	0.00	-	0.09
W→F	0.06	0.10	-	-	0.09	0.16	0.01	-	0.06
M→F	0.03	0.05	-	-	0.07	0.13	0.02	-	0.08

Nonresponse adjustment

Frequently used estimators for population mean		Approximate bias
Response mean (HT-estimator):	$\bar{y}_{RM} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$	$B_{RM} = \frac{\mathrm{cov}(Y, \rho)}{\bar{\rho}}$
Inverse Propensity Weighting:	$\bar{y}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\rho_X(X_i)}$	$B_{IPW} = \frac{\mathrm{cov}(Y, \frac{\rho}{\rho_X})}{\bar{\rho}}$
Generalised Regression:	$\bar{y}_{GREG} = \bar{y}_{RM} + \beta(\bar{x}_n - \bar{x}_{RM})$	$B_{GREG} = \frac{\mathrm{cov}(Y - \beta X, \rho)}{\bar{\rho}}$
Double-robust:	$\bar{y}_{DR} = \bar{y}_{IPW} + \beta(\bar{x}_n - \bar{x}_{IPW})$	$B_{DR} = \frac{\mathrm{cov}(Y - \beta X, \frac{\rho}{\rho_X})}{\bar{\rho}}$

Bias intervals under NMAR

Given correctly specified link functions between X, Y and ρ, and given the population regression parameters, the IPW, GREG and DR estimators center the NMAR bias interval around 0;

The width of the NMAR interval is approximately equal to

$$\frac{2S(Y)S(\rho)}{\bar{\rho}}\sqrt{1-\mathrm{cor}^2(Y,\beta X)}\sqrt{1-\mathrm{cor}^2(\rho,\rho_X)}$$

and, hence, proportional to

$$\sqrt{(\mathrm{CV}^2(\rho)-\mathrm{CV}^2(\rho_X))R^2(X,Y)}$$

with CV the coefficient of variation and R^2 the proportion of unexplained variance.

Setting 1 – uniform grouping

Theorem: If X is generated from a uniform grouping distribution, then

$$\mathrm{ECV}^2(\rho_X) = \frac{\mathrm{EC}-1}{\mathrm{G}-1}\mathrm{CV}^2(\rho).$$

Corollaries:

- When CV(design=1) < CV(design=2) for *X*, then in expectation CV(design=1) is also smaller for any other randomly drawn variable;
- For remaining bias after nonresponse adjustment, it holds that

$$\sqrt{CV^2(\rho) - ECV^2(\rho_X)} = \sqrt{\frac{G-EC}{G-1}}\,CV(\rho) = \sqrt{\frac{G-EC}{EC-1}}\,ECV(\rho_X).$$

Setting 2 – clustered uniform grouping

Theorem: If X is generated from a clustered, uniform grouping distribution, then it holds that $\mathrm{ECV}^2(\rho_X) = \frac{\mathrm{EC}-1}{\mathrm{G}-1}\frac{{S_{B,p}}^2(\rho)}{\bar{\rho}}$, with ${S_{B,p}}^2(\rho)$ a between variance and ${S_{B,p}}^2(\rho) \leq S^2(\rho)$.

Corollary: Theorem is motivation for acceptance-rejection schemes in which subsets of variables are selected.

1. Accept uniformly generated *X* when Cramer's V is larger than a specified threshold, $C_V(Y, X) > \gamma$;
2. Randomly draw variables from the subset of variables that relate to nonresponse (paradata);
3. Randomly draw variables from the subset of variables that relate to the survey variables;