2013 International Workshop for Household Survey Nonresponse, London, UK

Mini paper for Does balancing survey response reduce nonresponse bias?

Barry Schouten[1], Fannie Cobben[2], Peter Lundquist[3] and James Wagner[4]

Recently, various indicators have been proposed as indirect measures of nonresponse error in surveys. The indicators employ available auxiliary variables in order to detect nonrepresentative or unbalanced response. They can be used as quality objective functions in the design of survey data collection. Such designs are called adaptive survey designs, as different subgroups receive different treatments. The natural question is whether the potential decrease in nonresponse bias caused by adaptive survey designs could also be achieved by nonresponse adjustment methods that employ the same auxiliary variables. In this paper, we discuss this important question.

There is no easy way to answer the main question of the paper, as in most cases nonresponse biases on survey target variables are unknown. We circumvent this complication by dividing available auxiliary variables into two sets: a set to be used in the assessment and improvement of indicators of nonrepresentative response and a set to be used in the evaluation of remaining nonresponse bias. We do this in two ways. First, we apply indicators to growing sets of auxiliary variables and investigate whether patterns are consistent, i.e. whether worse indicator values on small models go together with worse values on large models. Second, we perform nonresponse adjustment using growing sets of weighting variables and search for consistency in the remaining biases, i.e. whether larger biases on small models coincide with larger biases on large models.

It is important to stress that the research question is to a large extent empirical. One can easily construct examples where balancing response does not reduce nonresponse bias. If we do find evidence in survey data that balancing helps, it, therefore, does not imply that the indicators have the feature that they detect nonresponse bias on other variables. It merely means that lower quality survey data collection, in the majority of cases, tends to affect the full range of potential variables and that the indicators successfully signal this tendency. Nonetheless, we do provide theoretical considerations that support balancing response through a combination of survey design and estimation.

The strength of the empirical evidence depends on the variety of surveys that are studied and the nature of the auxiliary variables that are input to the indicators. We have selected a wide range of survey data sets from three different countries to find empirical support. We compare the representativeness of response for growing sets of auxiliary variables over different surveys, over different waves of a survey, during data collection and after different survey process steps like establishing contact and obtaining cooperation. In each comparison the auxiliary variables are fixed, but variables are different over different comparisons and different data sets. The data sets that we have selected contain a relatively rich set of auxiliary variables that were linked from registry data.

Our study is somewhat similar to that of Peytcheva and Groves (2009), who investigated whether biases on auxiliary variables covary with biases on survey target variables. They found little evidence for such an association. Our study, however, shows that there is consistency in biases for auxiliary variables, even after adjustment. We do not extrapolate to survey target variables, but do discuss how such a consistency may extend to these variables as well.

Even if our results provide a rationale that adaptive and responsive survey designs are meaningful extensions of traditional survey sampling designs, implementation of such designs in survey practice is not straightforward or easy. It implies a different framework and mindset. We leave it to other papers to recommend on such implementation.

2. Indicators for representative or balanced response

We review indicators that have recently been proposed in the survey methodology literature as measures of nonresponse error. We refer to Särndal and Lundström (2008 and 2010), Särndal (2011), Schouten, Cobben and Bethlehem (2009), Schouten, Shlomo and Skinner (2011) and Shlomo, Skinner

[1] Statistics Netherlands and University of Utrecht, jg.schouten@cbs.nl
[2] Statistics Netherlands, f.cobben@cbs.nl
[3] Statistics Sweden, peter.lundquist@scb.se
[4] University of Michigan, jameswag@umich.edu

and Schouten (2012) for detailed accounts of the indicators and their statistical properties. We also refer to Wagner (2012) for a comparison and taxonomy of indicators.

The representativeness indicator or R-indicator for a variable Z is defined as the transformed standard deviation of the response propensity function ρ_Z

$$R(Z) = 1 - 2S(\rho_Z) . \quad (1)$$

Schouten, Cobben and Bethlehem (2009) introduce this indicator in a design-based context and propose an estimator using logistic regression. The estimator itself is usually referred to as the R-indicator. If one would use linear regression instead of logistic regression, then the R-indicator is equal to the balance indicator BI_2, proposed by Särndal and Lundström (2010). In practice, the choice of link function is, however, rarely influential. The rationale behind the indicators is that an absence of variation implies that response is a random subsample of the full sample.

Two indicators have a close similarity to the R-indicator and balance indicator. The first is the coefficient of variation of the response propensity function

$$CV(Z) = \frac{S(\rho_Z(Z))}{\mu} . \quad (2)$$

The indicator (2) can be estimated similar to (1) by dividing over the observed response rate. If the identity link function is used then the estimator has a close similarity to the coefficient of variation of the nonresponse adjustment weights proposed by Särndal and Lundström (2008), which they denote as H_3. Nonresponse adjustment weights can be viewed as smoothed inverse response propensities.

The second indicator that links to the R-indicator and balance indicator is the standardized contrast

$$C(Z) = \frac{S(\rho_Z(Z))}{\rho(1-\rho)} , \quad (3)$$

which is equal to the standardized difference in the expectations of Z for respondents and nonrespondents. Traditionally, the impact of nonresponse on the locations of distributions is decomposed as the product of the contrast and the nonresponse rate $1-\rho$. This product equals the coefficient of variation (2). Also the contrast has a counterpart in the Särndal and Lundström paper (2010); it is denoted by $dist_{R|NR}$.

Groves and Heeringa (2006), Wagner (2008), Särndal (2011) and Schouten, Calinescu and Luiten (2013) propose to differentiate efforts in surveys for different population subgroups in order to maximally reduce bias of estimators based on the survey response within the available survey budget. These designs are termed adaptive or responsive survey designs and resemble adaptive treatment regimes in other areas of statistics. The rationale is that different population subgroups may prefer or react differently to different treatments. The indicators in this section are proposed by some of the authors as quality objective functions in these optimal quality-cost trade-offs. They are applied to a number of candidate designs and the design that has the best indicator value is favoured. Adaptation to the sampled units can be done prior to data collection based on previous waves of the same survey or similar surveys, or during data collection based on observations made on the sampled units. Schouten, Shlomo and Skinner (2011) propose partial R-indicators to identify population subgroups that should be targeted in order to reduce variation in response propensities. Essentially, the variance of response propensities is decomposed into between and within components and the subgroups that have the largest within variances are discarded. The authors define unconditional partial R-indicators, denoted by $P_U(Z \mid X)$, and conditional partial R-indicators, denoted by $P_C(Z \mid X)$, where Z is an element of the auxiliary vector X. $P_U(Z \mid X)$ is defined as the between variance for Z of the response propensity function ρ_X. $P_C(Z \mid X)$ as the within variance attributable to Z given a stratification on X without Z, again of ρ_X. For exact definitions, we refer to Schouten, Shlomo and Skinner (2011). Lundquist and Särndal (2013) define partial imbalance indicators in a comparable fashion.

Can these indicators usefully be applied to improve survey design? We first note that the scaling of the indicators by the response rate in (2) and (3) and the nonresponse rate in (3) implies that the indicators generally lead to different design preferences. Only if the response rate is equal for these different designs, it is true that the choice of indicator is irrelevant. Hence, although they may be interesting in their own right, they cannot be used simultaneously in design decisions. More importantly, however, the indicators are critized for two main deficits. Recently, Beaumont and Haziza (2011) rightfully remarked that the early adaptive and responsive survey design papers restrict attention to bias and ignore variance. Also, in this paper we will focus mostly on bias, because we want to address the other alleged deficit. Although the indicators have subtle differences, they share one important feature: They can be estimated only for auxiliary variables X and not for the variables of interest Y, unless a model is formulated. This feature is the second deficit; balancing response on X may not be meaningful or useful because the missingness on these variables can be accounted for through an adjustment procedure and the real variables of interest remain unaffected. This discussion links strongly to the paper by Andridge and Little (2011) in which missingness is modelled as a function of $Y(X)+\lambda Y$, where $Y(X)$ is the projection of Y on X and λ is a moderating parameter. λ cannot be estimated but allows for a sensitivity analysis. Andridge and Little (2011) do this by computing the fraction of missing information (FMI) for different choices of λ. In adaptive and responsive survey designs the FMI cannot be used in a straightforward way however, as different designs will lead to different λ's and FMI intervals on λ may be broad and overlap.

3. Components of nonresponse bias

In this section, we provide theoretical considerations that support a focus on improving indicator values by design. We formalize the utility of the indicators as process quality indicators. More specifically, we formalize the intuition that a larger variation of response propensities for X corresponds to a larger variation of the true individual response probabilities. Doing so, we capitalize on the existence of an individual response probability.

We view auxiliary variables themselves as being sampled from the population of all possible random variables. Suppose a large population consists of G fully homogeneous and equally sized groups, labelled by $g = 1,2,\ldots,G$. All units in group g behave exactly the same in every way, and they have the same response probability for any given survey design. The stratification into the groups itself is not observed, but we do observe categorical variables X_k, $k = 1,2,\ldots,K$, that cluster groups into smaller numbers of groups.

Let us for simplicity look at an 0-1 indicator variable X. Assume that X was constructed by a simple random sample without replacement of size G_X from the set of G groups. Let s_g be the 0-1 indicator that group g was selected. We then have the following definition of X

$$X = \begin{cases} 1 & \forall g, s_g = 1 \\ 0 & \forall g, s_g = 0 \end{cases}, \tag{4}$$

i.e. X is one for all selected groups g and zero otherwise. Since the groups have equal size, the probability that $X = 1$ is equal to G_X / G.

Now, let ρ_g be the response probability of group g, so that the response propensity function $\rho_X(x)$ for X is defined as

$$\rho_X(x) = \begin{cases} \dfrac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g & if \quad x = 1 \\ \dfrac{1}{G - G_X}\sum_{g=1}^{G} (1 - s_g)\rho_g & if \quad x = 0 \end{cases}. \tag{5}$$

In order to investigate the relation between the indicators based on X and those based on the full stratification with the G groups, we consider the expected mean and the expected variance of the response propensity function ρ_X. The mean response propensity can be derived as

$$\bar{\rho}_X = \frac{G_X}{G}\frac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g + (1-\frac{G_X}{G})\frac{1}{G-G_X}\sum_{g=1}^{G}(1-s_g)\rho_g = \frac{1}{G}\sum_{g=1}^{G}\rho_g = \bar{\rho}\,. \qquad (6)$$

From (6) we can conclude that the mean response propensity $\bar{\rho}_X$ is always equal to the mean individual response probability $\bar{\rho}$. Clearly, the expected mean response propensity is then also equal to $\bar{\rho}$. Hence, regardless of the choice of X, the mean response propensity is the mean of the individual probabilities. The variance of ρ_X, $S^2(\rho_X)$, is equal to

$$S^2(\rho_X) = \frac{G_X}{G}(\rho_X(1)-\bar{\rho})^2 + (1-\frac{G_X}{G})(\rho_X(0)-\bar{\rho})^2\,. \qquad (7)$$

The expectation of $\rho_X(x)$ is always equal to $\bar{\rho}$, and, hence, (7) can be rewritten to

$$S^2(\rho_X) = \frac{G_X}{G}Var(\rho_X(1)) + (1-\frac{G_X}{G})Var(\rho_X(0))\,, \qquad (8)$$

where $Var(\rho_X(x))$ is the variance of ρ_X with respect to the sampling design. Since X is constructed using a simple random sample without replacement, $Var(\rho_X(x))$ is equal to

$$Var(\rho_X(x)) = \begin{cases} \frac{1}{G_X}(1-\frac{G_X}{G})S^2(\rho) & if \quad x=1 \\ \frac{1}{G-G_X}\frac{G_X}{G}S^2(\rho) & if \quad x=0 \end{cases}. \qquad (9)$$

Combining (8) and (9) gives

$$S^2(\rho_X) = \frac{1}{G}S^2(\rho)\,, \qquad (10)$$

so that the expected variance is equal to the variance of the individual response probabilities times the population diversity constant $1/G$.

With similar arguments, it can be reasoned that if X is a categorical variable with C categories, then

$$S^2(\rho_X) = \frac{C-1}{G}S^2(\rho)\,. \qquad (11)$$

So for all X, the variance of the response propensity function ρ_X is proportional to the variance of the underlying variance of individual response probabilities. This is a useful finding as it implies that, if for some survey design the R-indicator is smaller or the coefficient of variation is larger than for another survey design, then also the variance of the individual response probabilities is larger. As a consequence, the expected variance of the propensity function resulting from any random draw of subgroups g would be larger for that design too. Although it would not be true that the variance of all propensity functions is larger, there may in fact be various variables that lead to a smaller variance, it must hold that for an arbitrary variable the variance is larger. This conclusion supports the intuition that for surveys with many target variables, one would prefer larger R-indicators or smaller coefficients of variation. It also shows that for single topic surveys, it may actually be the survey target variable itself that is one of the exceptional variables.

When it comes to the bias of the standard estimators (response mean inverse propensity weighting, GREG and double robust estimators) it can be shown that

$$\sqrt{CV^2(\rho) - CV^2(\rho_X)} = \sqrt{\frac{S^2(\rho)}{\mu^2} - \frac{S^2(\rho_X)}{\mu^2}} = \frac{S(\rho)}{\mu}\sqrt{\frac{G-C+1}{G}} = \frac{S(\rho_X)}{\mu}\sqrt{\frac{G-C+1}{C-1}}, \qquad (12)$$

which is the omnipresent term in the bias intervals of all estimators.

If we could assume that the set of auxiliary variables X_k, $k = 1,2,\ldots,K$, consists of independent random draws of subgroups, then we could estimate G and $S^2(\rho)$. The parameter G may be estimated from the maximal covariance found among the X_k or using the first eigenvalue in a factor model. The variance $S^2(\rho)$ can be estimated by the average of the propensity function variances. Of course, the discussion in this section is conceptual, auxiliary variables cannot be considered as independent, random draws of population subgroups. However, models for nonresponse are often being critized for the lack of relevant, explanatory variables; standard variables like age or gender may have proved to be indicative of homogeneity in the population, they were certainly not picked to model response.

4. Testing the reduction of nonresponse bias

Even when some theoretical considerations, as laid out in the previous section, would advise to aim at improving indicator values through design, it would make a much stronger case if empirical results support such an endeavour. For this reason, we explored a wide range of survey data. We evaluated the validity of the preferences of the indicators using a multiple sample location rank test. In the test, we randomly divided the set of auxiliary variables into two groups: an evaluation set and a validation set. We tested the null hypothesis that indicator values and nonresponse biases for evaluation sets are not indicative of indicator values and nonresponse biases for validation sets.

For each indicator we rank the designs within a comparative data set based on growing models of auxiliary variables. We start by ranking the designs on X_1, then add X_2, and continue to add variables until the whole vector $X_v = (X_{v,1}, X_{v,2}, \ldots, X_{v,M_v})^T$ is included. Hence, the included auxiliary variables function as the evaluation set and the omitted auxiliary variables as the validation set. The evaluation set grows with each step, while the validation set shrinks. If we assume that indicators computed on evaluation sets are not predictive of indicators based on validation sets, then it holds that the different rankings are independent. The total number of pair wise inversions needed to go from the first ranking to the last ranking is the sum of independent numbers or pair wise inversions. A small total number of pair wise inversions implies clustered preferences, i.e. the indicator shows a consistent picture when different variables are considered. The multiple sample location rank test is based on the numbers of pairwise inversions. It is possible to test per comparative data set and to test all comparative data sets simultaneously, i.e. summing the individual test statistics over all data sets.

There are two basic assumptions underlying to the rank test: 1) the ranks do not produce any ties, and 2) auxiliary variables are independent. The rank test assumes that the indicators and biases have a continuous measurement level and do not produce ties. The indicators and biases are continuous but they are random variables and are subject to imprecision. As a result, two indicator values and two biases may not be statistically different at a certain significance level. Given a significance level, the indicators and biases do produce ties. It is, however, not straightforward how to account for the standard errors of the indicators and biases in the rank test without making assumptions on the probability distributions of the indicators and biases over variables. We, therefore, accepted that the tests will be conservative and we selected survey data sets that have modest to large sample sizes. The second assumption is more fundamental as it links to the independence assumption in the rank test. It is not true that the auxiliary variables are independent, and any dependence between the auxiliary variables may lead to spurious consistency in the rankings of indicator values. For this reason we also ranked designs within data sets based on conditional partial R-indicators (see Schouten, Shlomo and Skinner 2011) and on remaining nonresponse bias after adjustment for the GREG estimator, adding variables to the adjustment one by one.

5. Application to survey data sets

Our empirical illustration is based on a wide variety of survey data sets from three countries. The p-values are estimated for 14 datasets in two ways: 1) using quantiles based on independence as in section 4.2, and 2) using quantiles of simulated empirical probability distributions. The adjusted p-values based on the simulated distributions must, however, be interpreted with care as these are based on relatively simple models for the generation of auxiliary variables.
Table 5.1 contains the p-values. As expected, the unadjusted p-values are small for the R-indicator, coefficient of variation and the contrast. The adjusted (simulated) p-values are much larger and rarely have values smaller than usual significance levels. The unadjusted and adjusted p-values for the partial R-indicator and nonresponse biases are in most datasets similar in size, indicating that for these indicators p-values are robust for collinearity in the selected auxiliary variables. For the partial R-indicator five out of the 14 datasets have an unadjusted p-value smaller than 0.05 and of these five values four are smaller than 0.01. For the nonresponse bias three values are smaller than 0.05 and these are also smaller than 0.01.

Table 5.1: Observed numbers of inversions, expected numbers of inversions and p-values for various comparative datasets. R is the R-indicator, CV the coefficient of variation, C the contrast, Pc the conditional partial R-indicator and B the remaining bias of the GREG estimator.

Dataset	*p-value based on independence*					*p-value based on simulation*				
	R	*CV*	*C*	*Pc*	*B*	*R*	*CV*	*C*	*Pc*	*B*
HS	0.03	0.00	0.03	0.32	0.18	0.65	0.42	0.75	0.48	0.17
CVS	0.50	0.03	0.50	0.97	0.12	0.93	0.41	0.92	0.96	0.12
HS – CVS	0.14	0.00	0.20	0.96	0.00	0.98	0.70	1.00	1.00	0.00
LFS	0.00	0.01	0.03	0.82	0.82	0.22	0.62	0.77	0.88	0.78
SCS	0.00	0.00	0.00	0.00	0.00	0.17	0.02	0.04	0.07	0.01
SCSASD	0.06	0.06	0.06	0.06	0.06	0.27	0.27	0.27	0.07	0.02
LISS	0.00	0.00	0.00	0.03	0.00	0.00	0.07	0.12	0.27	0.00
STS-IND	0.01	0.01	0.01	0.72	0.72	0.09	0.12	0.11	0.74	0.65
STS-RET	0.01	0.03	0.03	0.50	0.88	0.09	0.41	0.39	0.53	0.82
LCS	0.00	0.00	0.00	0.00	0.07	0.18	0.00	0.56	0.01	0.08
PPS	0.00	0.00	0.00	0.00	0.10	0.06	0.00	0.70	0.00	0.13
SCA	0.08	0.01	0.08	0.81	0.35	1.00	1.00	0.98	0.92	0.31
NSFG	0.01	0.01	0.01	0.01	0.77	0.24	0.29	0.29	0.04	0.68
HRS	0.04	0.04	0.04	0.36	0.36	0.50	0.56	0.52	0.46	0.27

Table 5.2 contains the observed numbers of inversions and corresponding p-values for the partial R-indicator and nonresponse bias when multiple datasets are combined into one overall test. Three combinations of datasets are combined: all nine datasets from Statistics Netherlands, all five datasets from Stat Sweden and ISR Michigan, and all 14 datasets. In all cases the p-values are smaller than 0.05, and with one exception they are much smaller. The overall test, thus, indicates that the total observed numbers of inversions are much smaller than expected when design preferences per variable would be random.

Table 5.2: Expected numbers of inversions, observed numbers of inversions and p-values for combined datasets from Statistics Netherlands, from Stat Sweden and ISR Michigan and from all institutes. Pc is the conditional partial R-indicator and B the remaining bias of the GREG estimator.

	Number of inversions			*p-value*	
	Expected	*Pc*	*B*	*Pc*	*B*
Stat Netherlands	189.5	142	97	<0.001	<0.001
Stat Sweden/ISR	118.5	66	97	<0.001	0.02
All	308	208	194	<0.001	<0.001

While individual datasets do not point strongly at consistency in design preferences, their combination does indicate that nonresponse affects multiple variables simultaneously, even when adjusting for multicollinearity. This conclusion must be viewed with some care as it is still based on 14 datasets with a specific selection of auxiliary variables. Nonetheless, given the wide range of surveys these results do provide incentive to balance response on auxiliary variables by design regardless of any adjustment afterwards.

Points for discussion

- Can available auxiliary variables in nonresponse analyses be treated as random selections of the universe of variables on a target population?
- Do the results of the study provide empirical support for adaptive and responsive survey designs?
- Would others be interested in joining in in the evaluation, i.e. to apply the rank test to their survey data sets?