UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M09/18

Methodology

Estimation Of An Indicator Of The Representativeness Of Survey Response

Natalie Shlomo, Chris Skinner, Barry Schouten

Abstract

Nonresponse is a major source of estimation error in sample surveys. The response rate is widely used to measure the quality of a sample survey associated with nonresponse. It is, however, inadequate as an indicator because of its limited relation with nonresponse bias. Schouten et al. (2009) proposed an alternative indicator, which they refer to as an indicator of representativeness or R-indicator. This indicator measures the variability of the probabilities of response for units in the population. This paper develops methods for the estimation of this R-indicator assuming that values of a set of auxiliary variables are observed for both respondents and nonrespondents. In particular, we consider the bias of point estimators proposed by Schouten et al. (2009) and propose bias adjustments and linearization variance estimators. The proposed procedures are evaluated in a simulation study and their use is illustrated in an application to two business surveys at Statistics Netherlands.

Estimation of an Indicator of the Representativeness of Survey Response

Natalie Shlomo[1], Chris Skinner[1], Barry Schouten[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, United Kingdom

[2] Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands

Abstract

Nonresponse is a major source of estimation error in sample surveys. The response rate is widely used to measure the quality of a sample survey associated with nonresponse. It is, however, inadequate as an indicator because of its limited relation with nonresponse bias. Schouten et al. (2009) proposed an alternative indicator, which they refer to as an indicator of representativeness or R-indicator. This indicator measures the variability of the probabilities of response for units in the population. This paper develops methods for the estimation of this R-indicator assuming that values of a set of auxiliary variables are observed for both respondents and nonrespondents. In particular, we consider the bias of point estimators proposed by Schouten et al. (2009) and propose bias adjustments and linearization variance estimators. The proposed procedures are evaluated in a simulation study and their use is illustrated in an application to two business surveys at Statistics Netherlands.

Key words: nonresponse; quality; representative; response propensity; sample survey.

1. Introduction

One of the most important sources of estimation error in surveys is nonresponse. Survey organisations need indicators of such error for a variety of purposes, for example to compare different surveys, to monitor changes in a repeated survey over time or to monitor changes during the fieldwork of a single survey, perhaps to inform decisions such as when to end fieldwork. An indicator which is widely used for such purposes is the response rate, where a higher response rate is taken to indicate higher quality. However, there has been much recent empirical research (see e.g. Groves (2006), Groves and Peytcheva (2008), Heerwegh, Abts and Loosveldt (2007) and references therein) which concludes that the response rate is insufficient as an indicator to measure the potential error arising from nonresponse. Since sample sizes are usually large in surveys, the key feature of such error is typically nonresponse bias. However, the empirical evidence suggests that the response rate is only a weak predictor of nonresponse bias. There is therefore much interest in survey organisations in the development of alternative indicators (Groves et al., 2008).

In this paper, we consider an indicator proposed by Schouten et al. (2009). The basic idea is that nonresponse bias depends critically on the contrast between the characteristics of respondents and nonrespondents. This contrast can be assessed in terms of the probability of a unit responding to the survey. If all units in the population share the same probability of responding then no nonresponse bias will result and the response mechanism may be viewed as 'representative'. The indicator proposed by Schouten, Cobben and Bethlehem (2009), termed the R-indicator ('R' for representativeness), measures the extent to which the response probabilities vary. An advantage of this indicator (shared by the response rate) for various practical applications is that it provides a single measure for the whole survey. It should be recognized that nonresponse bias is defined in relation to a specific population parameter (and hence one or more survey variables). Thus, for any one (multipurpose) survey there may be a very large number of nonresponse biases. It would be feasible to construct indicators which are parameter-specific (Groves et al., 2008, Wagner, 2008), but here we suppose the requirement is for a single indicator for the whole survey.

Further discussion of the rationale and applications of the R-indicator is provided by Cobben and Schouten (2007), Schouten and Cobben (2007) and Schouten et al. (2009). The purpose of this paper is to consider in more detail some of the estimation issues associated with the R-indicator. In particular, we consider the bias of point estimators proposed by Schouten et al. (2009) and propose bias adjustments and linearization variance estimators. We evaluate these proposed procedures in a simulation study and demonstrate the application of these procedures in real business surveys.

We introduce the theoretical framework and define response propensities in Section 2. The R-indicator is defined at the population level in Section 3. The relation of the R-indicator to non-response bias is discussed in Section 4. Point estimation of the R-indicator using sample data is considered in Section 5. The bias of the point estimator and bias adjustment, variance estimation and confidence intervals are considered in Section 6. A simulation study and results of that study are described in Section 7 and results from real datasets are demonstrated in Section 8. Finally, we conclude and discuss future work in Section 9.

2. Preliminaries and Response Propensities

We suppose that a sample survey is undertaken, where a sample s is selected from a finite population U. The units in U are labelled $i = 1, 2, \ldots, N$, with the sizes of s and U denoted n and N, respectively. A probability sampling design is employed, where s is selected with probability $p(s)$. The first order inclusion probability of unit i is denoted π_i and $d_i = \pi_i^{-1}$ is the design weight.

The survey is subject to unit nonresponse, with the set of responding units denoted r, so $r \subset s \subset U$. We denote summation over the respondents, sample and population by Σ_r, Σ_s and Σ_U, respectively. Let R_i be the response indicator variable so that $R_i = 1$ if unit i responds and $R_i = 0$, otherwise. Hence, $r = \{i \in s; R_i = 1\}$.

We define the *response propensity* ρ_i as the conditional expectation (under a model) of R_i given the values of specified variables and survey conditions (Little, 1986, 1988). If it is necessary to clarify the conditioning, we write $\rho_i = \rho_X(x_i) = E_r(R_i \mid X_i = x_i)$ to denote the conditional expectation of R_i given that the vector of variables X_i for unit i takes the value x_i. Here $E_r(.)$ denotes expectation with respect to model underlying the response mechanism. We assume that R_i is defined for each population unit $i \in U$, so that nonresponse is what Rubin (1987) refers to as 'stable', and ρ_i is also defined for all $i \in U$. We also assume that the R_i for different units are independent, conditional on the specified variables and survey conditions. We shall further assume that the sampling design and the nonresponse process are 'unconfounded' (Rubin, 1987) so that the probability of selecting s remains $p(s)$, whatever the values of the $R_i, i \in U$. Thus, it is assumed that nonresponse does not depend on the configuration of the sample.

3. Representativeness Indicator

The variation in the response propensities may be viewed as reflecting the 'representativeness' of the nonresponse. Schouten et al. (2009) define response to be (strongly) *representative* if the response propensities are the same for all units in the population, corresponding to the notion of missing completely at random (MCAR) given the variables which are conditioned upon when defining ρ_i. They define a representativeness indicator, termed the *R-indicator* and denoted R_ρ, in terms of the population standard deviation of the response propensities: $S_\rho = \sqrt{(N-1)^{-1}\sum_U (\rho_i - \bar{\rho}_U)^2}$, where $\bar{\rho}_U = \sum_U \rho_i / N$. In order facilitate the interpretation of the indicator, they define it in terms of S_ρ as follows:

$$R_\rho = 1 - 2S_\rho \tag{3.1}$$

where this transformation of S_ρ ensures that $0 \le R_\rho \le 1$ since it may be shown that $S_\rho \le \sqrt{\bar{\rho}_U (1-\bar{\rho}_U)} \le 0.5$. The value $R_\rho = 1$ indicates the most representative response, where the ρ_i display no variation, and the value 0 indicates the least representative response, where the ρ_i display maximum variation.

4. Relation of Indicator to Non-response Bias

The R-indicator, R_ρ , may also be motivated in terms of nonresponse bias. Suppose that the target of inference is a population mean $\theta = N^{-1}\sum_U y_i$ of a survey variable, taking value y_i for unit i and observed only for $i \in r$. A standard design-weighted estimator of θ is $\hat{\theta} = \sum_s d_i R_i y_i / \sum_s d_i R_i$. The bias of $\hat{\theta}$ as an estimator of θ may be evaluated by taking expectations with respect to both the random sampling mechanism and the conditional distribution of R_i given the specified variables and conditions used to define ρ_i. These expectations are denoted E_s and E_r respectively. We assume, for now, that the specified variables include y_i so that it may be treated as fixed. We then have:

$$E_r E_s(\hat{\theta}) = E_r E_s\left(\sum_{i\in s} d_i R_i y_i / \sum_{i\in s} d_i R_i\right) \approx \sum_{i\in U} \rho_i y_i / \sum_{i\in U} \rho_i \, , \tag{4.1}$$

where the approximation is for large samples and we have used the assumption that the sampling and response mechanisms are unconfounded. Hence the bias depends on nonresponse only via ρ_i. It follows that

$$
\begin{aligned}
Bias(\hat{\theta}) &\approx \sum_{i \in U} \rho_i (y_i - \theta) / \sum_{i \in U} \rho_i \\
&= corr_{\rho y} S_\rho S_y / \overline{\rho}_U , \qquad (4.2)
\end{aligned}
$$

where $corr_{\rho y} = (N-1)^{-1} \sum_{i \in U} (\rho_i - \overline{\rho}_U)(y_i - \theta) / S_\rho S_y$ and $S_y^2 = (N-1)^{-1} \sum_{i \in U} (y_i - \theta)^2$.

Expression (4.2) is also obtained in Bethlehem (1988) and Särndal and Lundström (2005). An upper bound for the absolute bias can thus be expressed in terms of the R-indicator by

$$
| Bias(\hat{\theta}) | \leq S_\rho S_y / \overline{\rho}_U = \frac{(1 - R_\rho) S_y}{2 \overline{\rho}_U} \qquad (4.3)
$$

A standardized measure, which is free of y is given by:

$$
B = \frac{(1 - R_\rho)}{2 \overline{\rho}_U} \qquad (4.4)
$$

5. Estimation of R-indicator

We suppose that the data available for estimation purposes consists first of the values $\{y_i ; i \in r\}$ of the survey variable (or, more generally, a vector of survey variables), observed only for respondents. Secondly, we suppose that information is available on the values $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ of a vector X_i of auxiliary variables for all sample units, i.e. for both respondents and non-respondents. We refer to this as *sample-based auxiliary information*. This is a key assumption and is natural if, for example, the variables making up x_i are available on a register. Other possible assumptions about the availability of auxiliary information are discussed in section 9.

Since y_i is only observed for respondents, the response propensity conditional on y_i is generally inestimable without further assumptions. Instead, we propose to take ρ_i in the definition of R_ρ in (3.1) as conditional on x_i, i.e. to set $\rho_i = \rho_X(x_i) = E_r(R_i \mid X_i = x_i)$.

Nonresponse is *missing at random*, denoted MAR (Little and Rubin, 2002), if R_i is conditionally independent of y_i given x_i. In this case, we have $E_r(R_i \mid y_i, x_i) = E_r(R_i \mid x_i)$ and $\rho_i = \rho_X(x_i) = \rho_{YX}(y_i, x_i)$ and so y_i may implicitly be included in the conditioning set. Hence the argument used to obtain the bias bound in (4.3) still applies if MAR holds. The bias bound and the R-indicator itself may, however, be too conservative. If MAR holds then $\rho_Y(y_i) = E_r[\rho_X(x_i) \mid y_i]$ and:

$$\begin{aligned}\text{var}(\rho_i) &= \text{var}[\rho_X(x_i)] = \text{var}\{E[\rho_X(x_i) \mid y_i]\} + E\{\text{var}[\rho_X(x_i) \mid y_i]\} \\ &= \text{var}[\rho_Y(y_i)] + E\{\text{var}[\rho_X(x_i) \mid y_i]\} \qquad (5.1)\end{aligned}$$

The first term on the right hand side of (5.1) represents the variation of the conditional probabilities $\rho_Y(y_i)$, which we should ideally like to use in the R-indicator. The second term represents additional variation which is unrelated to non-response bias and may be viewed as redundant variability, i.e. noise, in the ρ_i relative to what we are interested in.

One special case occurs when nonresponse is missing completely at random (MCAR) so that it is independent of both x_i and y_i. In this case, both $\rho_X(x_i)$ and $\rho_Y(y_i)$ are constant so that both terms on the right hand side of (5.1) are zero.

Hence, there is no variability in the ρ_i and this does, albeit in a degenerate way, capture the fact that there is nothing in the nonresponse process that will lead to nonresponse bias for estimation related to y_i.

If nonresponse is not MAR then (5.1) no longer holds. Instead, $\rho_i = \rho_X(x_i)$ will represent a smoothed version of $\rho_{YX}(y_i, x_i)$ and it is not necessarily the case that $\mathrm{var}(\rho_i)$ will be at least as large as $\mathrm{var}[\rho_Y(y_i)]$. Thus, we may fail to capture relevant features of the nonresponse process in the ρ_i. In particular, if R_i is conditionally independent of x_i given y_i then $\mathrm{var}[\rho_Y(y_i)]$ will necessarily be at least as large as $\mathrm{var}(\rho_i)$, i.e. $\mathrm{var}[\rho_X(x_i)]$ (following a parallel argument to the MAR case). It may be argued therefore that it is desirable to select the auxiliary variables constituting x_i in such a way that the MAR assumption holds as closely as possible. In any case, it must be emphasized that our definition of $\rho_i = \rho_X(x_i)$ relates to a specific choice of auxiliary variables x_i. A different choice would generally result in a different ρ_i.

We noted in section 2 that we define the response propensity conditional on the survey conditions that apply when the data are collected. We do not make this conditioning explicit in our notation, but it is crucial to recognize this conditioning since, as we noted in Section 1, one of the objectives of constructing R-indicators is to be able to compare the representativeness of different surveys and such comparisons becomes challenging when the definition of the response propensity for any one survey is dependent on the conditions with which that survey has been implemented, for example upon the modes of data collection, the choice of interviewers, the way these interviewers were trained and work and the contact strategy. Even for a single survey repeated at different points in time, such conditions may well not remain constant.

5.2 Nonresponse models

In order to estimate the R-indicator, we first estimate the response propensities, $\rho_i = E(R_i \mid x_i)$. To do this, we assume that ρ_i depends on x_i in a parametric way via:

$$\mathrm{g}(\rho_i) = x_i'\beta , \tag{5.2}$$

where $\mathrm{g}(.)$ is a specified link function, β is a vector of unknown parameters and x_i may involve the transformation of the original auxiliary variables for the purpose of model specification. In particular, we shall consider the logit link function $g(\rho) = \log[\rho/(1-\rho)]$ leading to the logistic regression model.

We propose to estimate β by maximum pseudo likelihood (Skinner, 1989) i.e. β is estimated by $\hat{\beta}$, which solves:

$$\sum\nolimits_s d_i[R_i - g^{-1}(x_i'\beta)]x_i = 0 \tag{5.3}$$

where $g^{-1}(.)$ is the inverse of the link function. One reason for using the design weights here is because the objective is to estimate an R-indicator which provides a descriptive measure for the population.

The response propensity ρ_i is then estimated by:

$$\hat{\rho}_i = g^{-1}(x_i'\hat{\beta}) . \tag{5.4}$$

5.3 Estimation of R-indicator

As in Schouten et al. (2009), we propose to estimate R_ρ by:

$$\hat{R}_\rho = 1 - 2\hat{S}_\rho \ , \tag{5.5}$$

where $\hat{S}_\rho^2 = (N-1)^{-1}\sum_s d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$, $\hat{\rho}_i$ is defined in (5.4), $\hat{\bar{\rho}}_U = (\sum_s d_i \hat{\rho}_i)/N$ and N may be replaced by $\sum_s d_i$ if it is unknown.

6. Bias and Confidence Intervals

6.1 Bias and Bias Adjustment

We now consider the bias properties of the estimator $\hat{R}_\rho$ defined in (5.5). We shall assume that the vector of auxiliary variables x_i is given so that no bias can arise from specifying the 'wrong' set of auxiliary variables. We note, nevertheless, that the choice of auxiliary variables is a critical decision in practice and we shall illustrate empirically in section 7 how the R-indicator can depend on this choice.

Even if the vector of auxiliary variables is given, bias can arise from misspecification of the nonresponse model in (5.2). We first consider defining the bias with respect to the sampling mechanism, holding the R_i fixed. Under this source of random variation, the pseudo MLE $\hat{\beta}$ is approximately unbiased for the 'census' parameter β_U which solves

$$\sum_U [R_i - g^{-1}(x_i'\beta)]x_i = 0 \tag{6.1}$$

(Skinner, 2003). The approximation here is with respect to an asymptotic framework, with a sequence of samples and populations with n and N increasing. This census parameter implies a corresponding response propensity $\rho_{iU} = g^{-1}(x_i'\beta_U)$ and R-indicator $R_{\rho U}$, defined in terms of these propensities. We then have $E_s(\hat{R}_\rho) \approx R_{\rho U}$. The difference $R_{\rho U} - R_\rho$ may be viewed as the bias arising from model misspecification.

Instead of defining the bias with respect to just sampling variation we could also consider the response mechanism. In a parallel way, we may write $E_r E_s(\hat{R}_\rho) \approx R_{\rho U0}$, where $R_{\rho U0}$ is the R-indicator defined in terms of the response propensities $\rho_{iU0} = g^{-1}(x_i{}'\beta_{U0})$ and β_{U0} is the solution of:

$$\sum_U [\rho_{i0} - g^{-1}(x_i{}'\beta)]x_i = 0. \tag{6.2}$$

where $\rho_{i0} = E_r(R_i \mid x_i)$ is the true response propensity given x_i and we suppose that $g(\rho_{i0})$ is not necessarily linear in x_i, as in (5.2), i.e. the latter model may be misspecified. See Annex 1 for further discussion. Thus, $R_{\rho U0} - R_\rho$ may be viewed as the bias (with respect to both sampling variation and the response mechanism) arising from model misspecification. We may expect that $R_{\rho U} - R_{\rho U0} = O_p(N^{-0.5})$ so that there will usually be negligible difference in practice between the two measures $R_{\rho U} - R_\rho$ or $R_{\rho U0} - R_\rho$ of bias.

In principle, one might consider ways of assessing either of these measures of bias, perhaps by comparing the results of using the parametric model in (5.2) with those for some kind of non-parametric regression. We do not pursue this approach further here, however. Instead we consider the finite sample bias $E(\hat{R}_\rho) - R_{\rho U}$, treating $R_{\rho U}$ as the parameter of interest, which is equivalent to assuming that the nonresponse model in (5.2) is correctly specified. We might anticipate that the finite sample bias of $\hat{R}_\rho$ will be non-negligible, since $\hat{R}_\rho$ is defined via the variance of the $\hat{\rho}_i$ and we might expect sampling variation in these quantities to inflate this variance.

We approximate this finite sample bias of $\hat{R}_\rho$ by first considering the bias of $\hat{S}^2_\rho$. We derive in Annex 2 the following approximation:

$$E_p E_r(\hat{S}^2_\rho) \approx S^2_\rho + \lambda_1 + \lambda_2$$

where $\lambda_1 = E_s\{N^{-1}\sum_s d_i V_r(\hat{\rho}_i)\}$, $\lambda_2 = -\text{var}_s(\overline{\rho}_s) + 2N^{-1}\overline{\rho}_U \text{cov}(\hat{N}_s, \overline{\rho}_s)$, $\hat{N}_s = \sum_s d_i$ and $\overline{\rho}_s = N^{-1}\sum_s d_i \rho_i$ so that $\lambda_1 + \lambda_2$ represents the approximate bias of $\hat{S}^2_\rho$.

We then propose a bias-corrected estimator of R_ρ:

$$\tilde{R}_\rho = 1 - 2\tilde{S}_\rho. \tag{6.3}$$

where $\tilde{S}^2_\rho = \hat{S}^2_\rho - \hat{\lambda}_1 - \hat{\lambda}_2$ and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are estimators of λ_1 and λ_2 respectively.

An estimator of λ_1 is $\hat{\lambda}_1 = N^{-1}\sum_s d_i \hat{V}_r(\hat{\rho}_i)$, where $\hat{V}_r(\hat{\rho}_i)$ is an estimator of $V_r(\hat{\rho}_i)$ and N may be replaced by $\hat{N}_s$ if it is unknown. We propose to use the estimator $\hat{V}_r(\hat{\rho}_i)$ given in Annex 1. In the case of constant weights $d_i = N/n$ this gives:

$$\hat{\lambda}_1 = n^{-1}\sum_{i\in s} \nabla h(x_i'\hat{\beta})^2 x_i'[\sum_{j\in s} \nabla h(x_j'\hat{\beta}) x_j x_j']^{-1} x_i,$$

where $\nabla h(x_i'\hat{\beta}) = \exp(x_i'\hat{\beta})/[1+\exp(x_i'\hat{\beta})]^2$.

The second term λ_2 may, in general, be estimated using design-based variance estimation methods. In the case of constant weights the term $\hat{N}_s$ is constant so λ_2 reduces to $\lambda_2 = -\mathrm{var}_s(\bar{\rho}_s)$. Under simple random sampling, we may write $\lambda_2 = -(n^{-1} - N^{-1})S_\rho^2$. It follows that a bias corrected estimator of S_ρ^2 in the case of simple random sampling is:

$$\tilde{S}_\rho^2 = \hat{S}_\rho^2 - \hat{\lambda}_1 - \hat{\lambda}_2 = (1 + n^{-1} - N^{-1})\hat{S}_\rho^2 - n^{-1}\sum_{i \in s} \nabla h(x_i'\hat{\beta})^2 x_i' [\sum_{j \in s} \nabla h(x_j'\hat{\beta}) x_j x_j']^{-1} x_i . \quad (6.4)$$

6.2 Standard Errors and Confidence Intervals

A linearization variance estimator for $\hat{R}_\rho$ is derived in Annex 3 in terms of a variance estimator $v(\hat{S}_\rho^2)$ of $\hat{S}_\rho^2$, assuming that a logistic regression model is fitted and holds. A confidence interval for R_ρ with level $1-\alpha$ is given by $1 - 2\sqrt{\tilde{S}_\rho^2 \pm z_{\alpha/2} v(\hat{S}_\rho^2)^{0.5}}$.

7. Simulation Study of the Properties of the estimated R-indicators

7.1 Design of Simulation Study

In this section, we carry out a simulation study to assess the sampling properties of the estimation procedures described in section 6. The study is based on repeated samples drawn from a file (representing itself a 20% sample) from the 1995 Israel Census. The file contains 753,711 individuals aged 15 and over in 322,411 households. The samples are drawn using designs intended to be similar to some standard household and individual surveys carried out at national statistics institutes. We use the following sample designs in the simulations:

- Household Survey – similar to a Labour Force Survey where the sample units are households and all persons over the age of 15 in the sampled households are interviewed. Typically a proxy questionnaire is used and therefore there is no individual non-response within the household. In addition, we assume that every household has an equal probability to be included in the sample.

- Individual Survey - similar to a Social Survey where the sample units are individuals over the age of 15. We assume equal inclusion probabilities.

For each type of survey, we carried out a two-step design to define response probabilities in the census file. In the first step, we determined probabilities of response based on explanatory variables that typically lead to differential non-response based on our experiences of working with survey data collection. A response indicator was then generated for each unit in the population file. In the second step, we fit a logistic regression model and generate a 'true' response propensity for each unit in the population as predicted by the model. The dependent variable for the logistic model is the response indicator and the independent variables of the model the explanatory variables used in the first step (described below). This two-step design ensures that we have a known model generating the response propensities in the population and therefore can assess model misspecification besides the sampling properties of the indicators.

The explanatory variables used to generate the response probabilities are the following:

- Household Survey – Type of locality (3 categories), number of persons in household (1,2,3,4,5,6+), children in the household indicator (yes, no).

- Individual Survey – Type of locality (3 categories), number of persons in household (1,2,3,4,5,6+), children in the household indicator (yes, no), income group (15 groups), sex (male, female) and age group (9 groups).

Samples of size n were drawn from the Census population of size N at different sampling fractions 1:50, 1:100, and 1:200. For each sample drawn, a sample response indicator was generated from the 'true' population response probability. The overall response rate was 82% for the household survey and 78% for the individual survey. Response propensities and the R-indicator were then estimated from the sample. Two choices of auxiliary variables were considered, first the 'true' variables employed to generate the response propensities and, second, a simpler set of variables, intended to represent a possible misspecified model.

7.2 Results

Simulation means of $\hat{R}_\rho$, defined in (5.5), and its bias corrected version $\tilde{R}_\rho$, defined in (6.3), obtained from repeated samples drawn from a Household Survey at different sampling rates and for two different models are reported in Table 1. Corresponding results for the Individual Survey are presented in Table 2. The results

[PLACE TABLE 1 HERE]

[PLACE TABLE 2 HERE]

for the 'true' model provide evidence of downward bias in $\hat{R}_\rho$, with the (absolute) size of the bias increasing as the sample size decreases. This is as expected. Sampling error tends to lead to overestimation of the variability of the estimated response propensities and this leads to underestimation of the R-indicator. We observe that the bias correction reduces the (absolute) bias of $\hat{R}_\rho$ when the true model holds (although there is some evidence of over-correction in Table 2 which does not disappear as the sample size increases). The bias correction decreases (in absolute value) with the increase in sample sizes and tends to stabilize $\hat{R}_\rho$.

Using a less complex logistic model to estimate response probabilities results in a 'smoothing' of the probabilities and hence an increase in the value of the R-indicator. We include in Tables 1 and 2 values of $R_{\rho U0}$, which is the R-indicator for the logistic model for the reduced set of auxiliary variables which best fits the response propensities generated by the 'true' model (for the full set of auxiliary variables) in the population. Treating $R_{\rho U0}$ as the parameter of interest, we observe that the bias adjustment does reduce the (absolute) bias for the household survey but not necessarily for the individual survey, where the bias correction can lead to overestimation.

Simulation means of the linearization variance estimator (see section 6.2) are compared in Tables 3 and 4 with the simulation variances (calculated across the replicated samples) of $\hat{R}_\rho$ for the household and individual surveys, respectively.

[PLACE TABLE 3 HERE]

[PLACE TABLE 4 HERE]

The linearization variance estimator is seen to be approximately unbiased across the range of conditions represented in these tables.

Figures 1 and 2 present box plots comparing $\hat{R}_\rho$ and its bias adjusted version $\tilde{R}_\rho$ for the Household and Individual Survey simulation respectively when fitting the 'true' logistic regression model. The gains from the bias adjustment are evident.

[PLACE FIGURE 1 HERE]

[PLACE FIGURE 2 HERE]

8. Application to Real Surveys

We demonstrate R-indicators on business surveys undertaken for the 2007 Dutch Short Term Statistics (STS) for retail and industry. Table 5 provides a brief description of the two surveys.

[PLACE TABLE 5 HERE]

In the table, the survey response rates are given for 15, 30, 45 and 60 days of fieldwork. After 30 days STS needs to provide data for monthly statistics. We examine both a small set of auxiliary variables consisting of business size class (based on number of employees) and business sub-type. For the full auxiliary set we added VAT 2006 as collected by the Tax Board. Table 6 provides the results of the bias adjusted R-indicators, 95% confidence intervals and the standardized maximal bias (obtained by plugging estimated response propensities into (4.4)) after 15, 30, 45 and 60 days of fieldwork for each of the business surveys. Figures 3 and 4 provide plots of the bias-adjusted R-indicators against the response rates at each of the reporting times for the STS Industry and STS Retail respectively.

[PLACE TABLE 6 HERE]

[PLACE FIGURE 3 HERE]

[PLACE FIGURE 4 HERE]

The samples for the business surveys are large and hence the confidence intervals are small with widths between 1% and 1.5%. The R-indicator for STS retail after 30 days fieldwork drops almost 7% when VAT is added to the auxiliary information. For STS industry the decrease is much smaller. Apparently, the size of VAT in the previous year does not relate to response very strongly. Without the VAT information the retail respondents have a higher R-indicator than the industry respondents. When VAT is added this picture changes and the retail respondents score worse. STS retail shows a reduction in the R-indicator as the response rates increase for the small set of auxiliary variables. The main survey item of the STS surveys is monthly turnover (subdivided over different activities). As VAT in a previous year can be expected to correlate strongly to turnover in the running year, it is important that representativeness is good with respect to VAT. The main conclusion is that for Industry, the R-indicator goes up after 30 days, suggesting response representativeness is still improving and one would ideally wait longer than 30 days before producing statistics. For Retail, the R-indicator is lower, suggesting that response is less representative than for Industry, but there is very little change when data collection is prolonged. Hence, it does not pay off to wait longer than 30 days considering the composition of the response. The only reason to do so would be that the risk of nonresponse bias as reflected by the maximal bias is still decreasing as responses are coming in.

9. Discussion

In this paper we have considered a new indicator, called the R-indicator, designed to reflect the potential estimation error arising from nonresponse. The indicator is defined at the population level and we have developed methods for its estimation using sample data, including methods of bias adjustment and variance estimation. The approximate validity of these methods has been demonstrated via simulation. We have also demonstrated how the indicator may be used in real business surveys.

The indicator has been defined with respect to a set of auxiliary variables. A key assumption has been that these variables are measured on both respondents and nonrespondents. This assumption may be reasonable in some survey settings. For example, rich auxiliary information is available at Statistics Netherlands from a population register. However, in other survey settings, the availability of unit-level auxiliary information on nonrespondents may be very limited. Instead, aggregate information on the population totals of auxiliary variables may be available. We are addressing the estimation of R-indicators using such information in subsequent work.

Acknowledgements

This research was undertaken as part of the RISQ (Representativity Indicators for Survey Quality) project, funded by the European 7th Framework Programme (FP7), as a joint effort of the national statistical institutes of Norway, the Netherlands and Slovenia and the Universities of Leuven and Southampton. We should like to thank Li-Chun Zhang, Jelke Bethlehem, Mattijn Morren and Ana Marujo for their contributions.

Annex 1. Variance of $\hat{\rho}_i$ for logistic regression model

For the logistic regression model, write $h(\eta) = g^{-1}(\eta) = \exp(\eta)/[1+\exp(\eta)]$. The estimating equations in (5.3) may then be expressed as:

$$\sum\nolimits_s d_i[R_i - h(x_i'\beta)]x_i = 0 . \qquad \text{(A1.1)}$$

Let $\hat{\beta}$ solve (A1.1). Then in large samples we may approximate the distribution of $\hat{\beta}$ with respect to the sampling design (c.f. Skinner, 1989) by the distribution of :

$$\hat{\beta} \approx \beta_U + I(\beta_U)^{-1}\sum\nolimits_s d_i[R_i - h(x_i'\beta_U)x_i] , \qquad \text{(A1.2)}$$

where β_U is defined in (6.1), $I(\beta) = \sum\nolimits_s d_i \nabla h(x_i'\beta)x_i x_i'$ is the information matrix and $\nabla h(\eta) = \partial h(\eta)/\partial \eta = h(\eta)[1-h(\eta)]$. In particular, the variance of $\hat{\beta}$ with respect to the sampling design is in large samples

$$V_s(\hat{\beta}) \approx I(\beta_U)^{-1}V_s\{\sum\nolimits_s d_i[R_i - h(x_i'\beta_U)]x_i\}I(\beta_U)^{-1} \qquad \text{(A1.3)}$$

and, since $\hat{\rho}_i = h(x_i'\hat{\beta})$ from (5.4), we have

$$V_s(\hat{\rho}_i) \approx \nabla h(x_i'\beta_U)^2 x_i' V_s(\hat{\beta})x_i = \nabla h(x_i'\beta_U)^2 x_i' I(\beta_U)^{-1} V_s\{\sum_{j\in s} d_j[R_j - h(x_j'\beta_U)]x_j\}I(\beta_U)^{-1}x_i \qquad \text{(A1.4)}$$

This expression treats the response indicators R_j as fixed. To account for the response mechanism also, we may write $\rho_{i0} = E_r(R_i \mid x_i)$ and

$$\text{var}(\hat{\rho}_i) = E_r[V_s(\hat{\rho}_i)] + V_r[E_s(\hat{\rho}_i)] \qquad \text{(A.1.5)}$$

In large samples, we may write $E_s(\hat{\rho}_i) \approx h(x_i'\beta_U)$. Assuming $\rho_{i0} = E_r(R_i \mid x_i)$, we may write $\beta_U = \beta_{U0} + O_p(N^{-0.5})$ and $V_r[E_s(\hat{\rho}_i)] = O(N^{-1})$. The first term in (A.1.5) is generally of $O(N^{-1})$ and so the second term may be treated as negligible if the sampling fraction n/N may be treated as negligible. In this case an expression for $\text{var}(\hat{\rho}_i)$ may be obtained by replacing β_U in (A1.4) by β_{U0}.

Annex 2. Derivation of Bias adjustment

We consider the bias of $\hat{S}^2_\rho$ defined below (5.5). We use the decomposition:

$$\hat{\rho}_i - \hat{\bar{\rho}}_U = (\hat{\rho}_i - \rho_i) + (\rho_i - \bar{\rho}_U) + (\bar{\rho}_U - \bar{\rho}_s) + (\bar{\rho}_s - \hat{\bar{\rho}}_U)$$

where $\bar{\rho}_s = N^{-1}\sum_s d_i \rho_i$ and use the approximation $E_r(\hat{\rho}_i) \approx \rho_i$ to obtain $E_r(\hat{\bar{\rho}}_U) \approx \bar{\rho}_s$ and:

$$\begin{aligned} E_r[(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2] &\approx V_r(\hat{\rho}_i) + (\rho_i - \bar{\rho}_U)^2 + (\bar{\rho}_s - \bar{\rho}_U)^2 + V_r(\hat{\bar{\rho}}_U) \\ &\quad -2Cov_r(\hat{\rho}_i, \hat{\bar{\rho}}_U) - 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_s - \bar{\rho}_U) \\ &= (\rho_i - \bar{\rho}_U)^2 + V_r(\hat{\rho}_i - \hat{\bar{\rho}}_U) + (\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_s - \bar{\rho}_U) \end{aligned}$$

It follows that

$$\begin{aligned} E_r(\hat{S}^2_\rho) \approx (N-1)^{-1}\{&\textstyle\sum_s d_i(\rho_i - \bar{\rho}_U)^2 + \sum_s d_i V_r(\hat{\rho}_i - \hat{\bar{\rho}}_U) \\ &+ \hat{N}_s(\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\bar{\rho}_s - \bar{\rho}_U)(N\bar{\rho}_s - \hat{N}_s\bar{\rho}_U)\} \end{aligned}$$

where $\hat{N}_s = \sum_s d_i$.

Taking expectation also with respect to the sampling design, we obtain:

$$E_s E_r(\hat{S}^2_\rho) \approx S^2_\rho + A_1 + A_2 \qquad \text{(A2.1)}$$

where $A_1 = E_s\{(N-1)^{-1}\sum_s d_i V_r(\hat{\rho}_i - \hat{\bar{\rho}}_U)\}$

$$A_2 = E\{(N-1)^{-1}[\hat{N}_s(\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\bar{\rho}_s - \bar{\rho}_U)(N\bar{\rho}_s - \hat{N}_s\bar{\rho}_U)]\}$$

Both A_1 and A_2 are terms of $O(1/n)$ and, following standard linearization arguments, we simplify these expressions by removing terms of lower order. First, A_1 is asymptotically equivalent to:

$$\lambda_1 = E_s\{N^{-1}\sum_s d_i V_r(\hat{\rho}_i)\}.$$

Using the results in Annex 1 and assuming the nonresponse model is true, we may write :

$$\lambda_1 = E_s\{N^{-1}\sum_s d_i \nabla h(x_i'\tilde{\beta})^2 x_i' \text{var}(\hat{\beta}) x_i\}.$$

Turning to the term A_2, we may write

$$\hat{N}_s(\overline{\rho}_s - \overline{\rho}_U)^2 - 2(\overline{\rho}_s - \overline{\rho}_U)(N\overline{\rho}_s - \hat{N}_s\overline{\rho}_U) = \{\hat{N}_s - 2N\}(\overline{\rho}_s - \overline{\rho}_U)^2 + 2(\hat{N}_s - N)(\overline{\rho}_s - \overline{\rho}_U)\overline{\rho}_U$$

and, ignoring terms of lower order, A_2 is asymptotically equivalent to

$$\lambda_2 \approx -E_s\{(\overline{\rho}_s - \overline{\rho}_U)^2\} + 2\overline{\rho}_U E_s\{(N^{-1}\hat{N}_s - 1)(\overline{\rho}_s - \overline{\rho}_U)\}$$

$$= -\operatorname{var}_s(\overline{\rho}_s) + 2\overline{\rho}_U N^{-1}\operatorname{cov}_s(\hat{N}_s, \overline{\rho}_s).$$

Replacing A_1 and A_2 in (A2.1) by λ_1 and λ_2 respectively, we obtain the approximation:

$$E_p E_r(\hat{S}_\rho^2) \approx S_\rho^2 + \lambda_1 + \lambda_2.$$

Annex 3. Variance of Estimated R-indicator $\hat{R}(\boldsymbol{\rho})$ and Variance Estimation

From (5.5) and using linearization we have

$$\operatorname{var}[\hat{R}_\rho] \approx S_\rho^{-2}\operatorname{var}(\hat{S}_\rho^2) \quad \text{(A3.1.)}$$

To approximate $\operatorname{var}(\hat{S}_\rho^2)$ we shall decompose the distribution of $\hat{S}_\rho^2$ into the part induced by the sampling design for a fixed value of $\hat{\beta}$ and the part induced by the distribution of $\hat{\beta}$. We take the latter to be $\hat{\beta} \approx N(\beta, \Sigma)$, where:

$$\Sigma = J(\tilde{\beta})^{-1}\operatorname{var}\{\textstyle\sum_s d_i[R_i - h(x_i'\tilde{\beta})]x_i\}J(\tilde{\beta})^{-1} \quad \text{(A3.2)}$$

and $J(\beta) = E\{I(\beta)\}$ is the expected information rather than the observed information in (A1.3). These two choices of information are asymptotically equivalent (to first order) but the expected information has the advantage that Σ does not depend on s.

We write

$$\operatorname{var}(\hat{S}_\rho^2) = E_{\hat{\beta}}[\operatorname{var}_s(\hat{S}_\rho^2)] + \operatorname{var}_{\hat{\beta}}[E_s(\hat{S}_\rho^2)], \quad \text{(A3.3)}$$

where the subscript $\hat{\beta}$ denotes the distribution induced by $\hat{\beta} \approx N(\beta, \Sigma)$, which may be interpreted as arising from the response process. Following usual linearization arguments we obtain:

$$\text{var}_s(\hat{S}_\rho^2) \approx \text{var}_s[N^{-1}\sum_{i\in s} d_i(\rho_i - \overline{\rho}_U)^2]\Big|_{\beta=\hat{\beta}}$$

and, given the consistency of $\hat{\beta}$ for β (and for standard kinds of sampling designs), we have approximately:

$$E_{\hat{\beta}}[\text{var}_s(\hat{S}_\rho^2)] \approx \text{var}_s[N^{-1}\sum_{i\in s} d_i(\rho_i - \overline{\rho}_U)^2]. \qquad \text{(A3.4)}$$

Turning to the second component in (A3.3), we may write:

$$E_s(\hat{S}_\rho^2) \approx N^{-1}\sum_{i\in U}(\rho_i - \overline{\rho}_U)^2\Big|_{\beta=\hat{\beta}}.$$

As a linear approximation we have $\hat{\rho}_i \approx \rho_i + z_i'(\hat{\beta} - \tilde{\beta})$ where $z_i = \nabla h(x_i{}'\tilde{\beta})x_i$.

Hence

$$\sum_{i\in U}(\rho_i - \overline{\rho}_U)^2\Big|_{\beta=\hat{\beta}} \approx \sum_{i\in U}(\rho_i - \overline{\rho}_U)^2 + 2\sum_{i\in U}(\rho_i - \overline{\rho}_U)(z_i - \overline{z}_U)'(\hat{\beta} - \tilde{\beta})$$
$$+\sum_{i\in U}(z_i - \overline{z}_U)'(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \tilde{\beta})'(z_i - \overline{z}_U)$$

where $\overline{z}_U = N^{-1}\sum_U z_i$.

In large samples, we assume that $\hat{\beta}$ is normally distributed so that $(\hat{\beta} - \tilde{\beta})$ is uncorrelated with $(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \tilde{\beta})'$. Hence, we have

$$\text{var}_{\hat{\beta}}[E_s(\hat{S}_\rho^2)] \approx 4A'\Sigma A + \text{var}_{\hat{\beta}}\{tr[B(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \tilde{\beta})']\}, \qquad \text{(A3.5)}$$

where $A = N^{-1}\sum_{i\in U}(\rho_i - \overline{\rho}_U)(z_i - \overline{z}_U)$, $B = N^{-1}\sum_{i\in U}(z_i - \overline{z}_U)(z_i - \overline{z}_U)'$ and Σ is defined in (A3.2). The second term involves the fourth moments of $\hat{\beta}$ which may also be expressed in terms of Σ since $\hat{\beta}$ is assumed normally distributed.

The variance of $\hat{S}^2_\rho$ may be estimated by the sum of the estimated components of (A3.3). The first of these appears in (A3.4) and may be estimated by a standard design-based estimator of $\text{var}_s[\sum_{i\in s} d_i(\rho_i - \bar{\rho}_U)^2]$, where this is treated as the variance of a linear statistic $\text{var}_s[\sum_{i\in s} u_i]$ and u_i is replaced by $d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$ in the expression for the variance estimator. The second component of the variance appears in (A3.5). To estimate this term requires estimating A, B and Σ. First, z_i may be estimated by $\hat{z}_i = \nabla h(x_i'\hat{\beta})x_i$. Then A may be estimated by $\hat{A} = N^{-1}\sum_{i\in s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)(\hat{z}_i - \hat{\bar{z}}_U)$, B may be estimated by $\hat{B} = N^{-1}\sum_{i\in s} d_i(\hat{z}_i - \hat{\bar{z}}_U)(\hat{z}_i - \hat{\bar{z}}_U)'$, where $\hat{\bar{z}}_U = N^{-1}\sum_s d_i \hat{z}_i$, and Σ may be estimated by a standard estimator of the covariance matrix of $\hat{\beta}$.

Finally, the variance of $\hat{R}_\rho$ may be estimated by plugging the estimated variance of $\hat{S}^2_\rho$ into (A3.1) and replacing S^2_ρ by $\hat{S}^2_\rho$.

References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251 – 260.

Cobben, F. and Schouten, B. (2007). An empirical validation of R-indicators. Discussion paper, CBS, Voorburg.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R.M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72, 167-189.

Groves, R.M., Brick, J.M., Couper, M., Kalsbeek, W. Harris-Kojetin, F. Kreuter, Pennell, B., Raghunathan, T., Schouten, B., Smith, T., Tourangeau, R., Bowers, A., Jans, M., Kennedy, C., Levenstein, R., Olson, K., Peytcheva, E., Ziniel, S. and Wagner, J. (2008). Issues facing the field: alternative practical measures of representativeness of survey respondent pools. *Survey Practice*, October 2008 http://surveypractice.org/

Heerwegh, D., Abts, K. and Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 3-10.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A. and Rubin,D.B. (2002). *Statistical Analysis with Missing Data.* 2nd Ed. Hoboken, NJ.: Wiley.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Särndal, C-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester: Wiley.

Schouten, B. and Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes. Discussion paper 07002, CBS Voorburg.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In Skinner, C.J., Holt, D. and Smith, T.M.F. eds. *Analysis of Complex Surveys*, Chichester: Wiley.

Skinner, C.J. (2003). Introduction to Part B. In Chambers, R.L. and Skinner, C.J. *Analysis of Survey Data*, Chichester: Wiley.

Wagner, J.R. (2008). Adaptive survey design to reduce nonresponse bias. *PhD dissertation*, University of Michigan, USA.

Table 1: Household Survey - Simulation Means of $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ (across 500 simulated samples)

Sampling Fraction (sample size)	'True' Logistic Model (Number of Persons, Locality Type, Child Indicator) $R_\rho =$ 0.8780		Less Complex Logistic Model (Number of Persons) $R_{\rho U0} = 0.8842$	
	$\hat{R}_\rho$	$\tilde{R}_\rho$	$\hat{R}_\rho$	$\tilde{R}_\rho$
1:200 (n=1,612)	0.8700	0.8813	0.8755	0.8830
1:100 (n=3,224)	0.8735	0.8786	0.8801	0.8834
1:50 (n=6,448)	0.8749	0.8765	0.8807	0.8814

Table 2: Individual Survey - Simulation Means of $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ (across 500 simulated samples)

Sampling Fraction	'True' Logistic Model (Number of Persons, Sex, Age Groups, Income Groups, Locality Type, Child Indicator) $R_\rho = 0.8767$		Less Complex Logistic Model (Number of Persons, Sex and Age Groups) $R_{\rho U0} = 0.9023$	
	$\hat{R}_\rho$	$\tilde{R}_\rho$	$\hat{R}_\rho$	$\tilde{R}_\rho$
1:200 (n=3,769)	0.8587	0.8809	0.8941	0.9073
1:100 (n=7,537)	0.8686	0.8796	0.9008	0.9072
1:50 (n=15,074)	0.8748	0.8795	0.9029	0.9054

Table 3: Household Survey - Simulation mean of linearization estimator of variance of $\hat{R}_\rho$ *and simulation variance (across 500 simulated samples)* (10^{-3})

Sampling Fraction	'True' Logistic Model (Number of Persons, Locality Type, Child Indicator)		Less Complex Logistic Model (Number of Persons)	
	Simulation mean of linearization estimator	Simulation Variance	Simulation mean of linearization estimator	Simulation Variance
1:200 (n=1,612)	0.40	0.43	0.40	0.45
1:100 (n=3,224)	0.20	0.19	0.20	0.20
1:50 (n=6,448)	0.10	0.10	0.10	0.11

Table 4: Individual Survey - Simulation mean of linearization estimator of variance of $\hat{R}_\rho$ *and simulation variance (across 500 simulated samples)* (10^{-3})

Sampling Fraction	'True' Logistic Model (Number of Persons, Sex, Age Groups, Income Groups, Locality Type, Child Indicator)		Less Complex Logistic Model (Number of Persons, Sex and Age Groups)	
	Simulation mean of linearization estimator	Simulation Variance	Simulation mean of linearization estimator	Simulation Variance
1:200 (n=3,769)	0.21	0.23	0.19	0.19
1:100 (n=7,537)	0.10	0.11	0.09	0.11
1:50 (n=15,074)	0.05	0.05	0.04	0.05

Table 5: Description of 2007 Dutch Business Surveys

STS retail 2007	STS industry 2007
n=93,799	n=64,413
Response=49.5% (15days)	Response=48.8% (15days)
Response=78.0% (30days)	Response=78.7% (30days)
Response=85.8% (45days)	Response=85.7% (45days)
Response=88.2% (60days)	Response=88.3% (60days)
All businesses retail	All businesses industry
Stratified design on size class and business type	Stratified design on size class and business type
unequal design weights	unequal design weights
Fieldwork 90 days	Fieldwork 90 days
Paper + web	Paper + web

Table 6: Bias-adjusted R-indicators, 95% Confidence Intervals and Standardized Maximal Bias for Dutch Business Surveys using Small and Full Sets of Auxiliary Variables

		Small Set				Full Set			
Survey		15d	30d	45d	60d	15d	30d	45d	60d
Industry	R	92.1%	93.3%	94.0%	94.2%	90.5%	91.8%	93.1%	93.3%
	CI	91.3-92.8	92.7-94.0	93.5-94.4	93.8-94.6	89.7-91.3	91.3-92.2	92.6-93.5	92.8-93.8
	B	16.2%	8.5%	7.0%	6.6%	19.5%	10.4%	8.1%	7.6%
Retail	R	96.1%	94.6%	94.0%	94.1%	88.1%	87.9%	88.3%	89.0%
	CI	95.4-96.7	94.0-95.2	93.5-94.5	93.6-94.6	87.3-88.8	87.3-88.6	87.6-88.9	88.3-89.6
	B	7.9%	6.9%	7.0%	6.7%	24.0%	15.5%	13.6%	12.5%

Figure 1: Household Survey Box plots for $\hat{R}_\rho$ *and its Bias-Corrected Version,* $\tilde{R}_\rho$ *for 500 simulated samples with 1:200, 1:100 and 1:50 sampling fractions - 'True' R-Indicator = 0.8780*

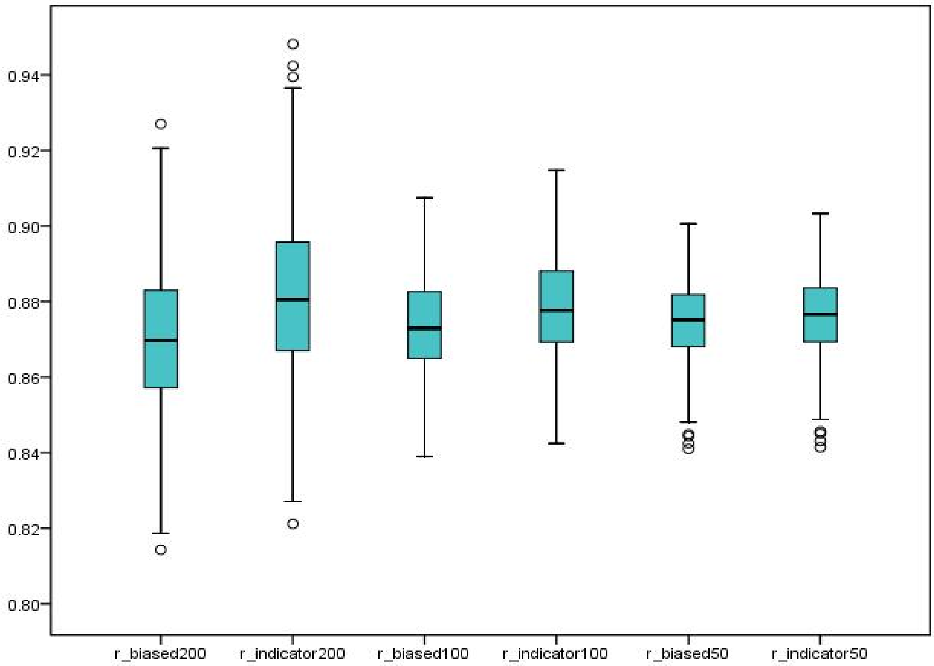

Figure 2: Individual Survey Box plots for $\hat{R}_{\rho}$ *and its Bias-Corrected Version,* $\tilde{R}_{\rho}$ *for 500 simulated samples with 1:200, 1:100 and 1:50 sampling fractions - 'True' R-Indicator = 0.8767*

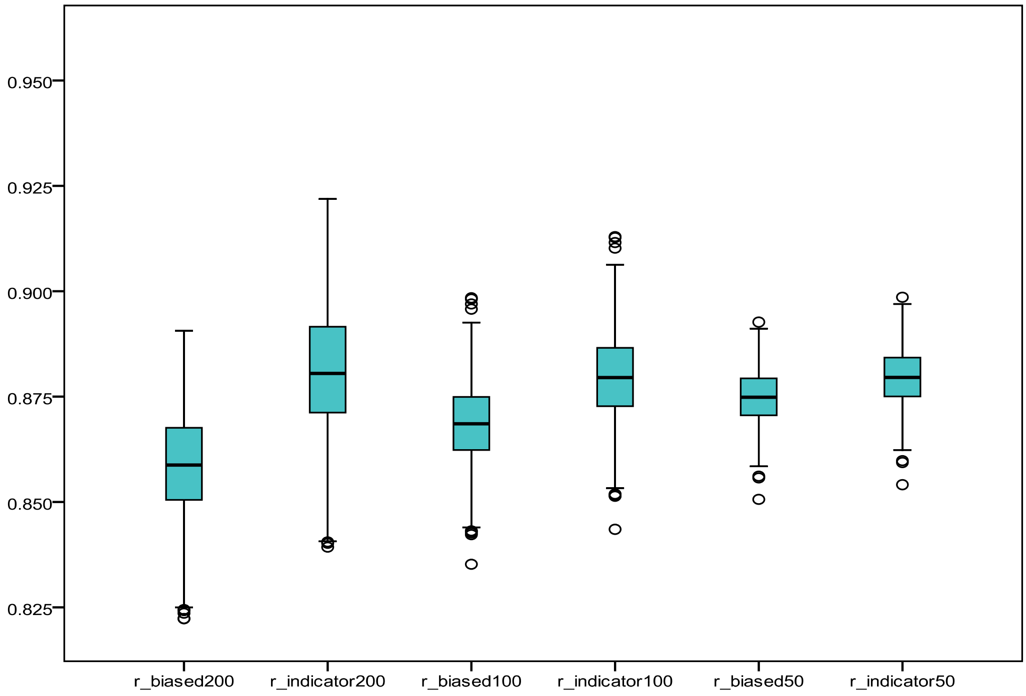

Figure 3: Plot of Response Rates against Bias Adjusted R-indicators at 15, 30, 45 and 60 Days of Fieldwork for the 2007 Dutch STS Industry survey

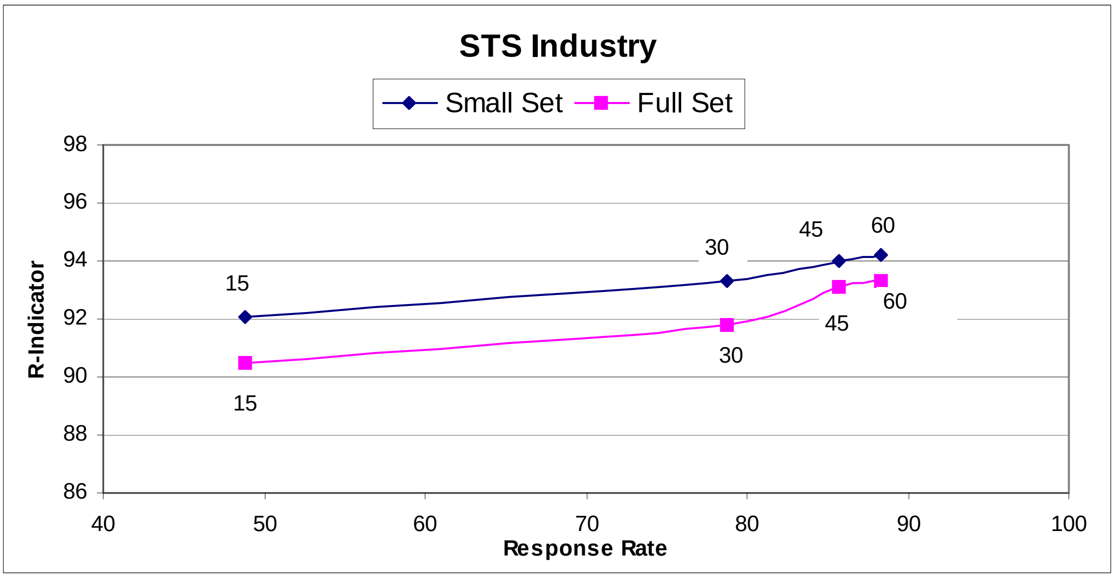

Figure 4: Plot of Response Rates against Bias Adjusted R-indicators at 15, 30, 45 and 60 Days of Fieldwork for the 2007 Dutch STS Retail survey