Volume 142 Issue 1 January 2012 ISSN 0378-3758

ELSEVIER

journal of statistical planning and inference

For Contents see page iii

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Statistical Planning and Inference 142 (2012) 201–211

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

journal of statistical planning and inference

Estimation of an indicator of the representativeness of survey response

Natalie Shlomo [a,*], Chris Skinner [a], Barry Schouten [b]

[a] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, United Kingdom
[b] Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands

ARTICLE INFO

Article history:
Received 28 October 2009
Received in revised form
27 June 2011
Accepted 5 July 2011
Available online 20 July 2011

Keywords:
Nonresponse
Quality
Representative
Response propensity
Sample survey

ABSTRACT

Nonresponse is a major source of estimation error in sample surveys. The response rate is widely used to measure survey quality associated with nonresponse, but is 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. We propose bias adjustments to the point estimator proposed by Schouten et al. (2009) and demonstrate the effectiveness of this adjustment in a simulation study where it is shown that the method is valid, especially for smaller sample sizes. We also propose linearization variance estimators which avoid the need for computer-intensive replication methods and show good coverage in the simulation study even when models are not fully specified. The use of the proposed procedures is also illustrated in an application to two business surveys at Statistics Netherlands.

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 et al., 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 squared bias component of mean squared error will typically dominate the variance component and hence it is desirable that the indicator reflect 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 referred to hereafter as SCB). The basic idea is that nonresponse bias depends critically on the contrast between the characteristics of respondents and nonrespondents.

[*] Corresponding author.
E-mail address: n.shlomo@soton.ac.uk (N. Shlomo).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.07.008

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 SCB, 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) and Schouten and Cobben (2007) in addition to the paper by SCB. The purpose of this paper is to consider in more detail some of the estimation issues associated with the R-indicator. The R-indicator proposed by SCB is subject to bias arising from the estimation of the response propensities. The bias is particularly problematic for small sample sizes, and a bias adjustment is developed. In addition, we develop linearization variance estimators as an alternative to the method of bootstrapping proposed in SCB. We evaluate these procedures in a simulation study and demonstrate the application of these procedures to a real business survey.

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 nonresponse 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 a real dataset 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, $r=\{i \in s;\ R_i=1\}$. Let $\mathbf{X}$ be a vector of auxiliary variables which may influence nonreponse, e.g. a vector consisting of age, gender, degree of urbanization of residence area and the observed status of the dwelling for a household survey or the type of business and the number of employees for a business survey.

We define the *response propensity* ρ_i as the conditional expectation (under a model) of R_i given the values of specified variables, such as the auxiliary variables above, and survey conditions (Little, 1986, 1988). If it is necessary to clarify that the conditioning is on a vector of variables $\mathbf{X}$, for example, then we write: $\rho_i=\rho_{\mathbf{X}}(\mathbf{x}_i)=E_r(R_i|\mathbf{X}_i=\mathbf{x}_i)$ to denote the conditional expectation of R_i given that the vector of variables $\mathbf{X}_i$ for unit i takes the value $\mathbf{x}_i$. Here $E_r(.)$ denotes expectation with respect to the 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, pp. 30–31) 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' (pp. 35–36, 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. In SCB response is defined 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) (Little and Rubin, 2002, p. 12) 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: $S_\rho=\sqrt{(N-1)^{-1}\sum_U(\rho_i-\overline{\rho}_U)^2}$, where $\overline{\rho}_U=\sum_U \rho_i/N$. In order to 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 $S_\rho \le \sqrt{\overline{\rho}_U(1-\overline{\rho}_U)} \le 0.5$. 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 R-indicator to nonresponse bias

The R-indicator 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, denoted E_s, and the response mechanism, denoted E_r as in Section 2. 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 in this case $\rho_i = \rho_{YX}(y_i, x_i)$ is conditional on both **Y** and **X** and the approximation is for large samples so that we take the first term only in the Taylor expansion. In addition, 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:

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

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, p. 92). An upper bound for the absolute bias can thus be expressed in terms of the R-indicator by

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

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

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

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. Second, we suppose that information is available on the values $\mathbf{x}_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$ of a vector $\mathbf{X}_i$ of auxiliary variables for all sample units, i.e. for both respondents and nonrespondents. We refer to this as *sample-based auxiliary information*. This is a key assumption and is natural if, for example, the variables making up $\mathbf{x}_i$ 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 $\mathbf{x}_i$, i.e. to set $\rho_i = \rho_X(\mathbf{x}_i) = E_r(R_i | \mathbf{X}_i = \mathbf{x}_i)$.

Nonresponse is *missing at random*, denoted MAR (Little and Rubin, 2002, p. 12), if R_i is conditionally independent of y_i given $\mathbf{x}_i$. In this case, we have $E_r(R_i | y_i, \mathbf{x}_i) = E_r(R_i | \mathbf{x}_i)$ and $\rho_i = \rho_{\mathbf{X}}(\mathbf{x}_i) = \rho_{Y\mathbf{X}}(y_i, \mathbf{x}_i)$ and so y_i may implicitly be included in the conditioning set. Hence the argument used to obtain the bias bound in Eq. (4.3) still applies if MAR holds. The bias bound and the R-indicator itself may, however, be too conservative. If MAR holds then $\rho_Y(y_i) = E_r[\rho_{\mathbf{X}}(\mathbf{x}_i) | y_i]$ and

$$\mathrm{var}(\rho_i) = \mathrm{var}[\rho_{\mathbf{X}}(\mathbf{x}_i)] = \mathrm{var}\{E[\rho_{\mathbf{X}}(\mathbf{x}_i) | y_i]\} + E\{\mathrm{var}[\rho_{\mathbf{X}}(\mathbf{x}_i) | y_i]\} = \mathrm{var}[\rho_Y(y_i)] + E\{\mathrm{var}[\rho_{\mathbf{X}}(\mathbf{x}_i) | y_i]\} \tag{5.1}$$

The first term on the right-hand side of Eq. (5.1) represents the variation of the conditional probabilities $\rho_Y(y_i)$, which we should ideally like to use in the R-indicator if we are only concerned about nonresponse bias for a target parameter defined in terms of y_i alone. The second term represents additional variation which is unrelated to nonresponse bias for such target parameters and may be viewed as redundant variability, i.e. noise, in the ρ_i for that purpose. Kreuter et al. (2010) present examples of auxiliary variables which are predictive of nonresponse but only weakly predictive of relevant y_i variables. In such cases, the second term in Eq. (5.1) may be relatively large and the R-indicator and its associated bias bound may be viewed as too conservative. It is therefore desirable to select only those auxiliary variables which are reasonably predictive of at least one of the y_i variables of key interest in the survey.

One special case occurs when nonresponse is missing completely at random (MCAR) so that it is independent of both $\mathbf{x}_i$ and y_i. In this case, both $\rho_{\mathbf{X}}(\mathbf{x}_i)$ and$\rho_Y(y_i)$ are constant so that both terms on the right-hand side of Eq. (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, $\rho_i = \rho_{\mathbf{X}}(\mathbf{x}_i)$ will represent a smoothed version of $\rho_{Y\mathbf{X}}(y_i, \mathbf{x}_i)$ 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 $\mathbf{x}_i$ given y_i then $\mathrm{var}[\rho_Y(y_i)]$ will necessarily be at least as large as $\mathrm{var}(\rho_i)$, i.e. $\mathrm{var}[\rho_{\mathbf{X}}(\mathbf{x}_i)]$ (following a parallel argument to the MAR case). It may be argued therefore that it is desirable to select the auxiliary variables constituting $\mathbf{x}_i$ in such a way that the MAR assumption holds as closely as possible. In any case, it must be emphasized that our definition of $\rho_i = \rho_{\mathbf{X}}(\mathbf{x}_i)$ relates to a specific choice of auxiliary variables $\mathbf{x}_i$. 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.

5.1. Nonresponse models

In order to estimate the R-indicator, we first estimate the response propensities, $\rho_i = E(R_i | \mathbf{x}_i)$. To do this, we assume that ρ_i depends on $\mathbf{x}_i$ in a parametric way via

$$g(\rho_i) = \mathbf{x}_i' \boldsymbol{\beta}, \tag{5.2}$$

where $g(.)$ is a specified link function, $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathbf{x}_i$ 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 $\boldsymbol{\beta}$ by pseudo maximum likelihood (Skinner (1989, pp. 80–83)) i.e. $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}$, which solves

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

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

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

5.2. Estimation of R-indicator

As in SCB, 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 Eq. (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. We refer to the estimator in Eq. (5.5) as the SCB estimator.

6. Bias and confidence intervals

6.1. Bias and bias adjustment

We now consider the bias properties of the SCB estimator $\hat{R}_\rho$ defined in Eq. (5.5). We shall assume that the vector of auxiliary variables $\mathbf{x}_i$ 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.

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 maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is approximately unbiased for the 'census' parameter $\boldsymbol{\beta}_U$ which solves

$$\sum\nolimits_U [R_i - g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})] \mathbf{x}_i = 0 \tag{6.1}$$

(Skinner ,2003, p. 82). 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 $\rho_{iU} = g^{-1}(\mathbf{x}_i{}' \boldsymbol{\beta}_U)$ 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 $\rho_{iU0} = g^{-1}(\boldsymbol{x}_i{}' \boldsymbol{\beta}_{U0})$ and $\boldsymbol{\beta}_{U0}$ is the solution of

$$\sum\nolimits_U [\rho_{i0} - g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})] \mathbf{x}_i = 0, \tag{6.2}$$

where $\rho_{i0}=E_r(R_i|\mathbf{x}_i)$ is the true response propensity given $\mathbf{x}_i$ and we suppose that $g(\rho_{i0})$is not necessarily linear in $\mathbf{x}_i$, as in Eq. (5.2), i.e. the latter model may be misspecified. 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 Eq. (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 Eq. (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 $\hat{S}^2_\rho$ defined below (5.5).

We use the decomposition

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

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

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

It follows that:

$$E_r(\hat{S}^2_\rho)\approx(N-1)^{-1}\{\sum\nolimits_s d_i(\rho_i-\overline{\rho}_U)^2+\sum\nolimits_s d_iV_r(\hat{\rho}_i-\hat{\overline{\rho}}_U)+\hat{N}_s(\overline{\rho}_s-\overline{\rho}_U)^2-2(\overline{\rho}_s-\overline{\rho}_U)(N\overline{\rho}_s-\hat{N}_s\overline{\rho}_U)\}$$

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

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

$$E_sE_r(\hat{S}^2_\rho)\approx S^2_\rho+A_1+A_2 \tag{6.3}$$

where $A_1=E_s\{(N-1)^{-1}\sum_s d_iV_r(\hat{\rho}_i-\hat{\overline{\rho}}_U)\}$

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

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

$$\lambda_1=E_s\{N^{-1}\sum\nolimits_s d_iV_r(\hat{\rho}_i)\}.$$

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

$$\lambda_2\approx-E_s\{(\overline{\rho}_s-\overline{\rho}_U)^2\}+2\overline{\rho}_UE_s\{(N^{-1}\hat{N}_s-1)(\overline{\rho}_s-\overline{\rho}_U)\}=-\text{var}_s(\overline{\rho}_s)+2\overline{\rho}_UN^{-1}\text{cov}_s(\hat{N}_s,\overline{\rho}_s)$$

Replacing A_1 and A_2 in Eq. (6.3) by λ_1 and λ_2, respectively, we obtain $\lambda_1+\lambda_2$ as the approximate bias of $\hat{S}^2_\rho$. We thus propose as a bias-corrected SCB estimator of R_ρ

$$\tilde{R}_\rho=1-2\tilde{S}_\rho \tag{6.4}$$

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 the Annex. In the case of constant weights $d_i=N/n$ this gives

$$\hat{\lambda}_1=n^{-1}\sum_{i\in s}\nabla h(\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\mathbf{x}_i'[\sum_{j\in s}\nabla h(\mathbf{x}_j'\hat{\boldsymbol{\beta}})\mathbf{x}_j\mathbf{x}_j']^{-1}\mathbf{x}_i,$$

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

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 $\lambda_2=-V_s(\overline{\rho}_s)$. Under simple random sampling, we may write $\lambda_2=-(n^{-1}-N^{-1})S^2_\rho$. It follows that a bias corrected estimator of S^2_ρ in the case of simple random sampling is

$$\tilde{S}^2_\rho=\hat{S}^2_\rho-\hat{\lambda}_1-\hat{\lambda}_2=(1+n^{-1}-N^{-1})\hat{S}^2_\rho-n^{-1}\sum_{i\in s}\nabla h(\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\mathbf{x}_i'[\sum_{j\in s}\nabla h(\mathbf{x}_j'\hat{\boldsymbol{\beta}})\mathbf{x}_j\mathbf{x}_j']^{-1}\mathbf{x}_i. \tag{6.5}$$

6.2. Standard errors and confidence intervals

A linearization variance estimator for the SCB estimator $\hat{R}_\rho$ is now derived in terms of a variance estimator $\nu(\hat{S}^2_\rho)$ of $\hat{S}^2_\rho$, assuming that a logistic regression model is fitted and holds. From Eq. (5.5) and using linearization we have

$$\mathrm{var}[\hat{R}_\rho] \approx S_\rho^{-2}\mathrm{var}(\hat{S}^2_\rho). \tag{6.6}$$

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

$$\boldsymbol{\Sigma} = \mathbf{J}(\boldsymbol{\beta})^{-1}\mathrm{var}\{\textstyle\sum_s d_i[R_i - h(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i\}\mathbf{J}(\boldsymbol{\beta})^{-1}, \tag{6.7}$$

and $\boldsymbol{J}(\boldsymbol{\beta}) = E\{\boldsymbol{I}(\boldsymbol{\beta})\}$ is the expected information rather than the observed information in Eq. (6.7). These two choices of information are asymptotically equivalent (to first order) but the expected information has the advantage that $\boldsymbol{\Sigma}$ does not depend on s.

We write

$$\mathrm{var}(\hat{S}^2_\rho) = E_{\hat{\boldsymbol{\beta}}}[\mathrm{var}_s(\hat{S}^2_\rho)] + \mathrm{var}_{\hat{\boldsymbol{\beta}}}[E_s(\hat{S}^2_\rho)], \tag{6.8}$$

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

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

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

$$E_{\hat{\boldsymbol{\beta}}}[\mathrm{var}_s(\hat{S}^2_\rho)] \approx \mathrm{var}_s\left[N^{-1}\sum_{i\in s} d_i(\rho_i - \overline{\rho}_U)^2\right]. \tag{6.9}$$

Turning to the second component in Eq. (6.8), we may write

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

As a linear approximation we have $\hat{\rho}_i \approx \rho_i + \mathbf{z}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ where $\mathbf{z}_i = \nabla h(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i$. Hence

$$\sum_{i\in U}(\rho_i - \overline{\rho}_U)^2\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \approx \sum_{i\in U}(\rho_i - \overline{\rho}_U)^2 + 2\sum_{i\in U}(\rho_i - \overline{\rho}_U)(\mathbf{z}_i - \overline{\mathbf{z}}_U)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sum_{i\in U}(\mathbf{z}_i - \overline{\mathbf{z}}_U)'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{z}_i - \overline{\mathbf{z}}_U)$$

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

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

$$\mathrm{var}_{\hat{\boldsymbol{\beta}}}[E_s(\hat{S}^2_\rho)] \approx 4\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} + \mathrm{var}_{\hat{\boldsymbol{\beta}}}\{tr[\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\}, \tag{6.10}$$

where $\mathbf{A} = N^{-1}\sum_{i\in U}(\rho_i - \overline{\rho}_U)(\mathbf{z}_i - \overline{\mathbf{z}}_U)$, $\mathbf{B} = N^{-1}\sum_{i\in U}(\mathbf{z}_i \text{-} \overline{\mathbf{z}}_U)(\mathbf{z}_i - \overline{\mathbf{z}}_U)'$and $\boldsymbol{\Sigma}$ is defined in Eq. (6.7). The second term involves the fourth moments of $\hat{\boldsymbol{\beta}}$ which may also be expressed in terms of $\boldsymbol{\Sigma}$ since $\hat{\boldsymbol{\beta}}$ is assumed normally distributed.

The variance of $\hat{S}^2_\rho$ may be estimated by the sum of the estimated components of Eq. (6.8). The first of these appears in Eq. (6.9) and may be estimated by a standard design-based estimator of $\mathrm{var}_s[\sum_{i\in s} d_i(\rho_i - \overline{\rho}_U)^2]$, where this is treated as the variance of a linear statistic $\mathrm{var}_s[\sum_{i\in s} u_i]$ and u_i is replaced by $d_i(\hat{\rho}_i - \hat{\overline{\rho}}_U)^2$ in the expression for the variance estimator. The second component of the variance appears in Eq. (6.10). To estimate this term requires estimating $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\Sigma}$. First, $\mathbf{z}_i$ may be estimated by $\hat{\mathbf{z}}_i = \nabla h(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i$. Then $\mathbf{A}$ may be estimated by $\hat{\mathbf{A}} = N^{-1}\sum_{i\in s} d_i(\hat{\rho}_i - \hat{\overline{\rho}}_U)(\hat{\mathbf{z}}_i - \hat{\overline{\mathbf{z}}}_U)$, $\mathbf{B}$ may be estimated by $\hat{\mathbf{B}} = N^{-1}\sum_{i\in s} d_i(\hat{\mathbf{z}}_i - \hat{\overline{\mathbf{z}}}_U)(\hat{\mathbf{z}}_i - \hat{\overline{\mathbf{z}}}_U)'$, where $\hat{\overline{\mathbf{z}}}_U = N^{-1}\sum_s d_i\hat{\mathbf{z}}_i$, and $\boldsymbol{\Sigma}$ may be estimated by a standard estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

Finally, the variance of the SCB estimator $\hat{R}_\rho$ may be estimated by plugging the estimated variance of $\hat{S}^2_\rho$ into Eq. (6.6) and replacing S^2_ρ by $\hat{S}^2_\rho$.

A confidence interval for R_ρ with level $1-\alpha$ is given by $1 - 2\sqrt{\tilde{S}^2_\rho \pm z_{\alpha/2}\nu(\hat{S}^2_\rho)^{0.5}}$.

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—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 nonresponse within the household. In addition, we assume that every household has an equal probability to be included in the sample.
- Individual survey—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 population (census) file. In the first step, we determined probabilities of response based on explanatory variables that typically lead to differential nonresponse 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).

Five hundred 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 the SCB estimator $\hat{R}_\rho$, defined in Eq. (5.5), and its proposed bias corrected version $\tilde{R}_\rho$, defined in Eq. (6.4), obtained from the 500 repeated samples drawn for the 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. Also included in the tables is the percentage of Relative Bias calculated from the simulation study as: $100\{[\sum_{B=1}^{500}(\hat{R}_{\rho B}-R_\rho)/R_\rho]/500\}$ where $\hat{R}_{\rho B}$ is the value of $\hat{R}_\rho$ computed for the Bth simulation sample and similarly for $\tilde{R}_\rho$.

The results for the 'true' model provide evidence of downward bias in the SCB estimator, 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

Table 1
Household survey—simulation means of $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ and their percent relative bias (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			
	SCB$\hat{R}_\rho$		Proposed$\tilde{R}_\rho$		SCB$\hat{R}_\rho$		Proposed$\tilde{R}_\rho$	
	Mean	Relative bias (%)	Mean	Relative bias (%)	Mean	Relative bias (%)	Mean	Relative bias (%)
1:200 (n=1612)	0.8700	−0.91	0.8813	0.38	0.8755	−0.98	0.8830	−0.14
1:100 (n=3224)	0.8735	−0.51	0.8786	0.07	0.8801	−0.46	0.8834	−0.09
1:50 (n=6448)	0.8749	−0.35	0.8765	−0.17	0.8807	−0.40	0.8814	−0.32

Table 2
Individual survey—simulation means of $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ and their percent relative bias (across 500 simulated samples).

Sampling Fraction (sample size)	'True' Logistic Model (Number of Persons, Sex, Age Groups, Income Groups, Locality Type, Child Indicator) R_ρ=0.8767				Less Complex Logistic Model (Number of Persons, Sex and Age Groups) $R_{\rho U0}$=0.9023			
	SCB$\hat{R}_\rho$		Proposed $\tilde{R}_\rho$		SCB $\hat{R}_\rho$		Proposed $\tilde{R}_\rho$	
	Mean	Relative bias (%)	Mean	Relative bias (%)	Mean	Relative bias (%)	Mean	Relative bias (%)
1:200 (n=3769)	0.8587	−2.05	0.8809	0.48	0.8941	−0.91	0.9073	0.55
1:100 (n=7537)	0.8686	−0.92	0.8796	0.33	0.9008	−0.17	0.9072	0.54
1:50 (n=15,074)	0.8748	−0.22	0.8795	0.32	0.9029	0.07	0.9054	0.34

Table 3
Household survey—simulation mean of linearization estimator of variance of $\hat{R}_\rho$ with coverage rate and, simulation variance (across 500 simulated samples) (10^{-3}).

Sampling Fraction (sample size)	'True' Logistic Model (Number of Persons, Locality Type, Child Indicator)			Less Complex Logistic Model (Number of Persons)		
	Simulation Mean of Linearization Estimator	Coverage Rate (%)	Simulation Variance	Simulation Mean of Linearization Estimator	Coverage Rate (%)	Simulation Variance
1:200 (n=1612)	0.40	94.1	0.43	0.40	93.6	0.45
1:100 (n=3224)	0.20	95.6	0.19	0.20	95.8	0.20
1:50 (n=6448)	0.10	94.4	0.10	0.10	92.6	0.11

Table 4
Individual survey—simulation mean of linearization estimator of variance of $\hat{R}_\rho$ with coverage rate and simulation variance (across 500 simulated samples) (10^{-3}).

Sampling Fraction (sample size)	'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	Coverage Rate (%)	Simulation Variance	Simulation Mean of Linearization Estimator	Coverage Rate (%)	Simulation Variance
1:200 (n=3769)	0.21	94.4	0.23	0.19	94.8	0.19
1:100 (n=7537)	0.10	93.4	0.11	0.09	92.4	0.11
1:50 (n=15,074)	0.05	93.8	0.05	0.04	92.1	0.05

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 the SCB estimator 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 the SCB estimator.

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 complete 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. The instability of the bias correction for the less complex set of auxiliary variables is likely caused by the misspecification of the model. Since the bias correction depends on the correct specification of the logistic model, it may not perform quite so well in these cases.

The Relative Root Mean Square Errors (RRMSE) were also calculated from the simulation study as: $100\{R_\rho^{-1}[\sqrt{\sum_{B=1}^{500}(\hat{R}_{\rho B}-R_\rho)^2/500}]\}$ but not presented in the tables. There was no systematic evidence of the bias adjustment leading to larger RRMSEs. For the Household Survey, the RRMSE was stable across the SCB and proposed estimators for both types of models. For the Individual Survey, there was more variability in the RRMSEs due to the fluctuating bias corrections across the types of models. This result is reflected by the Relative Bias that is presented in the tables.

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 the SCB estimator for the household and individual surveys, respectively. The tables also include the Coverage Rate defined as the percentage of times that the true R_ρ is

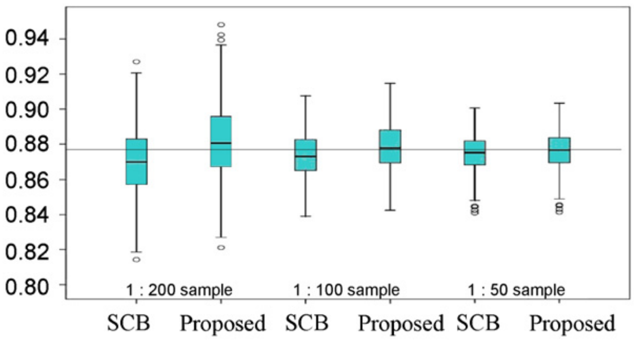

Fig. 1. Household survey box plots for SCB $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ for 500 simulated samples with 1:200, 1:100 and 1:50 sampling fractions—'True' R-Indicator=0.8780.

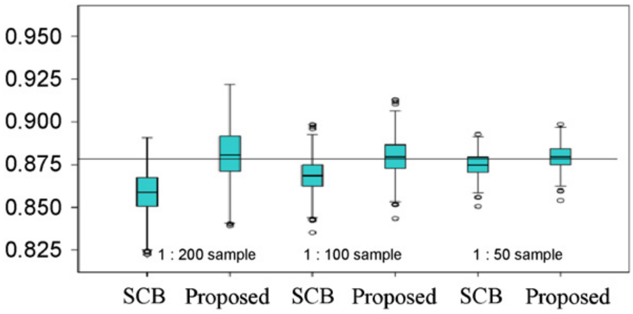

Fig. 2. Individual survey box plots for SCB $\hat{R}_\rho$ and its bias-corrected version, $\tilde{R}_\rho$ for 500 simulated samples with 1:200, 1:100 and 1:50 sampling fractions—'True' R-Indicator=0.8767.

included in the confidence interval calculated by the linearization variance estimator $100\{[\sum_{B=1}^{500} I(R_\rho \in \hat{R}_{\rho B} \pm 2\sqrt{\text{var}_B(\hat{R}_{\rho B})})]/500\}$ where $\text{var}_B(\hat{R}_{\rho B})$ is the estimated linearization variance for the Bth simulation sample and I is the indicator function.

The linearization variance estimator is seen to be approximately unbiased across the range of conditions represented in these tables under the different sample sizes with good coverage as seen by the coverage rate in the tables.

Figs. 1 and 2 present box plots comparing the SCB estimator and its proposed bias adjusted version for the Household and Individual Survey simulation, respectively, when fitting the 'true' logistic regression model. The gains from the bias adjustment are evident.

8. Application to real surveys

We demonstrate the use of 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.

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 complete set of auxiliary variables consisting of (i) business size class (based on number of employees), (ii) business sub-type and (iii) VAT 2006 as collected by the Tax Board and a reduced set consisting of just (i) and (ii). Table 6 provides the results of the unadjusted and bias adjusted R-indicators, 95% confidence intervals and the standardized maximal bias (obtained by plugging estimated response propensities into Eq. (4.4)) after 15, 30, 45 and 60 days of fieldwork for each of the business surveys. Because of the large sample size, the bias adjustment had a small impact.

The samples for the business surveys are large and hence the confidence intervals are reduced 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 reduced. 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 reduced 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

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 unequal design weights	Stratified design on size class and business type unequal design weights
Fieldwork 90 days	Fieldwork 90 days
Paper + web	Paper + web

Table 6
Unadjusted (top) and bias-adjusted (bottom) R-indicators, 95% confidence intervals and standardized maximal bias (formula 4.4) for Dutch business surveys using reduced and complete sets of auxiliary variables.

Survey		Reduced set				Complete set			
		15 d	30 d	45 d	60 d	15 d	30 d	45 d	60 d
Industry	*R* (%)	91.9	93.2	93.9	94.1	90.2	91.6	92.9	93.2
		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* (%)	95.9	94.5	93.9	94.0	87.9	87.8	88.2	88.9
		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

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 as well as social surveys. The bias adjustment is particularly effective for small sample sizes. In addition, the variance estimation provides good coverage and avoids the need for computer-intensive resampling methods.

The indicator is defined with respect to a set of auxiliary variables. An R-indicator cannot be viewed separately from the auxiliary vector $\mathbf{X}$ that was used to define it. As such, indicator values should always be reported with reference to the auxiliary variables. Consequently, when comparing multiple surveys within one survey institute, over survey institutes or even over countries, one needs to fix the set of auxiliary variables used for each of the surveys. There are two conflicting aims that need to be balanced when selecting auxiliary variables in the comparison of different surveys. On the one hand, it is desirable to choose auxiliary variables that are maximally correlated with the variables of analytic interest in each survey. On the other hand, the choice is constrained to the set of auxiliary variables that is available for each of the surveys. The wider the scope of the comparison, the more restrictive the availability of variables will be. Within one survey institute one is likely to use one sampling frame, have access to the same register data and collect similar paradata for surveys. Multiple countries, however, may have completely different traditions and legislation, which will limit the set of auxiliary variables that is shared. More discussion on the selection of auxiliary variables is in Schouten et al. (submitted for publication).

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 A. 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

$$\sum\nolimits_s d_i[R_i - h(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i = 0. \tag{A1}$$

Let $\hat{\boldsymbol{\beta}}$ solve (A1). Then in large samples we may approximate the distribution of $\hat{\boldsymbol{\beta}}$ with respect to the sampling design (c.f. Skinner, 1989, pp. 80–83,) by the distribution of

$$\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_U + \mathbf{I}(\boldsymbol{\beta}_U)^{-1}\sum\nolimits_s d_i[R_i - h(\mathbf{x}_i'\boldsymbol{\beta}_U)\mathbf{x}_i], \tag{A2}$$

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

$$V_s(\hat{\boldsymbol{\beta}}) \approx \mathbf{I}(\boldsymbol{\beta}_U)^{-1} V_s\left\{\sum\nolimits_s d_i[R_i - h(\mathbf{x}'{}_i\boldsymbol{\beta}_U)]\mathbf{x}_i\right\}\mathbf{I}(\boldsymbol{\beta}_U)^{-1} \tag{A3}$$

and, since $\hat{\rho}_i = h(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ from Eq. (5.4), we have

$$V_s(\hat{\rho}_i) \approx \nabla h(\mathbf{x}_i'\boldsymbol{\beta}_U)^2 \mathbf{x}_i' V_s(\hat{\boldsymbol{\beta}})\mathbf{x}_i = \nabla h(\mathbf{x}_i'\boldsymbol{\beta}_U)^2 \mathbf{x}_i' \mathbf{I}(\boldsymbol{\beta}_U)^{-1} V_s\{\sum_{j\in s} d_j[R_j - h(\mathbf{x}_j'\boldsymbol{\beta}_U)]\mathbf{x}_j\}\mathbf{I}(\boldsymbol{\beta}_U)^{-1}\mathbf{x}_i \tag{A4}$$

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

$$\mathrm{var}(\hat{\rho}_i) = E_r[V_s(\hat{\rho}_i)] + V_r[E_s(\hat{\rho}_i)] \tag{A5}$$

In large samples, we may write $E_s(\hat{\rho}_i) \approx h(\mathbf{x}_i'\boldsymbol{\beta}_U)$. Assuming $\rho_{i0} = E_r(R_i|\mathbf{x}_i)$, we may write $\boldsymbol{\beta}_U = \boldsymbol{\beta}_{U0} + O_p(N^{-0.5})$ and $V_r[E_s(\hat{\rho}_i)] = O(N^{-1})$. The first term in Eq. (A5) 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 $\mathrm{var}(\hat{\rho}_i)$ may be obtained by replacing $\boldsymbol{\beta}_U$ in Eq. (A4) by $\boldsymbol{\beta}_{U0}$.

References

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

Cobben, F., 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., 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., 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., Loosveldt, G., 2007. Minimizing survey refusal and noncontact rates: do our efforts pay off? Survey Research Methods 1, 3–10.

Kreuter, F., Olsen, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casa-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., Raghunathan, T.E., 2010. Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. Journal of the Royal Statistical Society, Series A 173, 389–407.

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., Rubin, D.B., 2002. Statistical Analysis with Missing Data, 2nd Ed. Wiley, Hoboken, NJ.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Särndal, C.-E., Lundström, S., 2005. Estimation in Surveys with Nonresponse. Wiley, Chichester.

Schouten, B., 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., Bethlehem, J., 2009. Indicators for the representativeness of survey response. Survey Methodology 35, 101–113.

Schouten, B., Shlomo, N., Skinner, C.J.. Indicators for monitoring and improving representativeness of response. Journal of Official Statistics, submitted for publication. 〈http://eprints.soton.ac.uk/158353/〉.

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

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

Wagner, J.R., 2008. Adaptive survey design to reduce nonresponse bias. Ph.D. Dissertation. University of Michigan, USA.