RISQ

Representativity Indicators for Survey Quality

Statistical Properties of R-indicators

Work Package 3
Deliverable 2.1
Version 2[1]

Natalie Shlomo & Chris Skinner
University of Southampton, United Kingdom

Barry Schouten & Jelke Bethlehem
Centraal Bureau voor de Statistiek, Netherlands

Li-Chun Zhang
Statistisk Sentralbyrå, Norway

26 February 2009

SEVENTH FRAMEWORK PROGRAMME

The RISQ Project is financed by the 7th Framework Programme (FP7) of the European Union. Cooperation Programma, Socio-economic Sciences and the Humanities, Provision for Underlying Statistics

[1] Version 2 of deliverable 2.1 was made in order to address comments raised by Paul Biemer (RTI International and University of North Carolina) in his review of the first version (December 8, 2008).

Statistical Properties of R-indicators

1. Introduction

One of the most important factors affecting the quality of surveys of households or enterprises is nonresponse. The impact of nonresponse on survey quality is typically measured by the response rate. The response rate alone, however, is not sufficient as a quality indicator to capture the potential impact of nonresponse. The bias of estimates resulting from nonresponse also depends on the contrast between respondents and nonrespondents with respect to a target variable. The more they differ, the larger the bias will be. Good indicators that measure the degree to which the group of respondents of a survey still resembles the complete sample are currently lacking.

RISQ (Representativity Indicators for Survey Quality), a project funded from the 7th EU Framework Programme (FP7), was set up in order to fill the gap of indicators for nonresponse. The main objectives of the project are to elaborate and develop indicators for the representativity of survey response, to explore the characteristics of these indicators, and to show how to implement them in a practical data collection environment. With these indicators the project attempts to support the comparison of the quality of different surveys and to facilitate the efficient allocation of data collection resources. We call the indicators Representativity indicators or R-indicators.

This report is the first technical deliverable of RISQ and develops and discusses statistical theory for R-indicators. The paper concentrates on the statistical properties, i.e. bias and variance, of two potential R-indicators. Due to the missing data character of nonresponse, no statement about the representativeness of response is possible without information that is external to the survey. The availability and form of this auxiliary information strongly influences the use and interpretation of potential indicators. Furthermore, it is intuitively evident that the size of the survey sample is influential as well. In very small samples there is insufficient information contained in the response to make strong statements about the nature of the nonresponse. Both issues, auxiliary information and sample size, will not be discussed in this report, but are the topics of a forthcoming RISQ report. Here, we elaborate on the properties of R-indicators given a fixed, available set of auxiliary variables and given an unknown sample size.

One of the main objectives of the R-indicators we construct, is to enable the comparison of the representativeness of different surveys that contain the same sets of auxiliary variables. By doing so we disconnect the representativeness of the survey response from the survey topic(s). Or in other words the R-indicators view representativeness as a function of fully observed auxiliary information only. This way we can compare responses to surveys with different topics. It must be noted, however, that the lack of representativeness, as we propose to measure it, may be more harmful for some survey topics than for other survey topics.

Our construction of R-indicators is not designed to provide an adjustment for nonresponse nor to provide guidance in the selection of weighting variables. The R-indicators do, however, show the extent to which we need to rely on nonresponse adjustment methods. Nonresponse adjustment methods may remove part of the bias that is due to nonresponse under the assumption that nonrespondents resemble respondents with approximately the same values of the weighting variables. Lower values of R-indicators indicate that this assumption may be less likely.

In order to construct R-indicators for comparing different surveys, data sets were assembled from the participating countries in the RISQ research project: Belgium, The Netherlands, Norway, Slovenia and the UK. In this paper we illustrate the two R-indicators for a number of country data sets given auxiliary variables that are shared by all countries. The country data sets consist of both household and business surveys.

The report will focus on the following two R-indicators:

1. a measure based upon the variance of estimated response probabilities, as discussed in the papers by Cobben and Schouten (2005, 2007) which provided the basis of the original application, and as discussed in Schouten et al. (2008).

2. a related measure proposed by Särndal and Lundström (2008), in the context of selecting auxiliary variables for weighting adjustment.

We have selected the Cobben and Schouten (2005) indicator as it is directly based on a definition of representative response. They call response strongly representative when individual response probabilities are the same for all population units. Response is weakly representative with respect to some stratification variable when response propensities of the corresponding strata are equal. Cobben and Schouten propose the indicator as a natural measure for the deviation from weakly representative response. Also, their indicator conveniently appears in bounds for the nonresponse of survey estimators like the weighted response mean and the generalised regression estimator.

Särndal and Lundström (2008) propose a measure that enables the selection of variables in calibration estimators. Variables in weighting models ideally relate both to nonresponse and to the main survey items. Särndal and Lundström derive their measure from approximations to the bias of calibration estimators and argue that for many possible survey items a large value of the measure will correspond to a small bias of the calibration estimator for those items. This measure was selected because, as for the Cobben and Schouten measure, it can be computed using auxiliary information only and as it is based on a clear underlying goal, i.e. the minimisation of nonresponse bias.

The definition of both R-indicators will be presented in this report together with a discussion of their theoretical foundations and properties as well as a report on empirical results obtained from a simulation study and estimates from selected country data sets. We will focus on theoretical aspects of the R-indicators in the context of a single survey, where the aim is to estimate each R-indicator. For the practical motivation and potential uses of these indicators, see Cobben and Schouten (2005, 2007), Schouten et al. (2008) and Särndal and Lundström (2008).

The estimation of the R-indicators is very much dependent on the type of available auxiliary information. In this report, we shall assume that auxiliary information is available at the sample level. A quite different set of methods is available if the only auxiliary information available is in aggregated form at the population level – these methods will be discussed in a later RISQ deliverable.

In this report, we shall first formulate the theoretical framework in section 2. In particular, we shall discuss the notion of response propensities. A more detailed discussion of some of the issues arising in the definition of response propensities is given in Annex 1. The two R-indicators are defined formally at the population level in section 3. The estimation of these population-level R-indicators using sample data is then discussed in section 4. The theoretical properties of these point estimators are discussed in section 5. Since the estimators are subject to potentially non-negligible bias, we introduce bias-corrected estimators. We also consider the variances of the estimators and potential variance estimators and confidence intervals. Section 6 includes some discussion of the theoretical relationship between the R-indicators and nonresponse bias. A simulation study and results of that study are described in Section 7. Finally, illustrative findings from applications to the country data sets are given in Section 8.

2. Theoretical Framework

2.1 General notation and nature of available information

We suppose that a sample survey is undertaken, where a sample s is selected from a finite population U. The sizes of s and U are denoted n and N, respectively. The units in U are labelled$i = 1, 2, \ldots, N$. The sample is assumed to be drawn by a probability sampling design $p(.)$, where the sample$s$ 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. In some cases, we shall assume simple random sampling without replacement.

We suppose that the survey is subject to unit nonresponse. The set of responding units is denoted . Thus, we have $r \subset s \subset U$. We denote summation over the respondents, sample and population by Σ_r, Σ_s and Σ_U, respectively. We 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 shall suppose that the typical target of inference is a population mean $\bar{Y} = N^{-1} \sum_U y_i$ of a vector of survey variables, taking value y_i for unit i.

We suppose that the data available for estimation purposes consists first of the values $\{y_i = (y_{1,i}, y_{2,i}, \ldots, y_{L.i})^T; i \in r\}$ of the survey variables, observed only for respondents. Secondly, we suppose that information is available on the values of $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$, a vector of auxiliary variables. We shall usually suppose each $x_{k,i}$ is a binary indicator variable, where x_i represents one or more categorical variables, since this will be the case in the applications we consider, but our presentation allows for general $x_{k,i}$ values. We assume that values of x_i are observed for all respondents. For the majority of this document we shall also assume that x_i is known for all sample units, i.e. for both respondents and non-respondents. We refer to this as *sample-based auxiliary information*. This is a natural assumption if, for example, the variables making up x_i are available on a register. However, in many countries and survey settings the availability of auxiliary information on non-respondents may be very limited, e.g. because of the absence of a register. In such circumstances, *aggregate population-based auxiliary information* may be available. This might take the form of a (finite) population total and/or mean and/or covariance matrix of x_i. We shall refer briefly to the use of such information for estimation in this document. However, we shall postpone considering this possibility in detail until the Deliverable 2 of this workpackage (WP3). Of course, it is also possible that there exists some combination of sample-based and population-based auxiliary information, with the combination perhaps varying between the different variables constituting x_i. For simplicity, this document will focus just on the case of sample-based information on the whole of x_i.

2.2 Response propensities

We first assume for simplicity that nonresponse is what Rubin (1987) refers to as 'stable', that is that the response indicator variable R_i is defined for each population unit $i \in U$. We shall further assume that the sampling design and the nonresponse process are 'unconfounded' (Rubin, 1987) so that the probability of selecting $s \subset U$ 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.

We define the *response propensity* as a conditional expectation of the response indicator variable R_i given the values of specified variables and survey conditions (Little, 1986, 1988). In other words, the response propensity is the probability of response conditional on the specified variables and conditions. For example, we may write $\rho_{YX}(y_i, x_i) = E(R_i \mid y_i, x_i)$ as the response propensity, if the probability is conditional on y_i and x_i. Here, the subscript YX indicates the conditioning variables. We use the term 'response propensity' to indicate that the definition is specific to the conditioning variables and that we are not referring to any assumed 'true response probability' that exists for each unit, whatever the nature of the auxiliary variables. We might seek to interpret the probabilistic nature of the response propensity (i.e. the source of the expectation $E(.)$ in our definition) as being with respect to the *nonresponse process*. However, since the definition is conditional on arbitrary conditioning variables, it will implicitly also usually refer to some underlying *superpopulation model.* To illustrate this, let $\rho_Y(y_i) = E(R_i \mid y_i)$. Then, since $E(R_i \mid y_i) = E[E(R_i \mid y_i, x_i) \mid y_i]$, we may write:

$$\rho_Y(y_i) = E[\rho_{YX}(y_i, x_i) \mid y_i]. \tag{2.1}$$

Here, we are treating $\rho_{YX}(y_i, x_i)$ as a random variable, where the randomness derives from a superpopulation distribution for x_i and the expectation is taken across this distribution. Hence, if $\rho_{YX}(y_i, x_i)$ were interpreted as reflecting the response process then $\rho_Y(y_i)$ needs to be interpreted as reflecting a combination of this response process and the superpopulation model for x_i.

Note also that we implicitly assume that R_i depends only on values of survey or auxiliary variables for unit i and not for other units in the population.

We argue in Annex 1 (section A1.3) that an ideal definition of the response propensity would be the probability of response conditional on y_i, which in the general case would be a vector of all survey variables of interest. In this case, we would write $\rho_Y(y_i) = E(R_i \mid y_i)$. The attraction of this definition is that it would capture all aspects of the response process relevant to bias in estimation of population parameters defined in terms of y_i. However, under this definition, $\rho_Y(y_i)$ would in general not be directly estimable because, by assumption, y_i is missing for nonrespondents. An alternative definition, and the one we adopt, is to take the response propensity as $\rho_X(x_i) = E(R_i \mid x_i)$, where the vector of auxiliary variables is defined as in section 2.1. For simplicity, we shall usually write $\rho_i = \rho_X(x_i)$ and hence denote the response propensity just by ρ_i.

An important condition, in this case when the response propensity is defined as $\rho_X(x_i) = E(R_i \mid x_i)$, is whether nonresponse is *missing at random*, denoted MAR (Little and Rubin, 2002), that is whether nonresponse is conditionally independent of y_i given x_i. Under the MAR condition, we may write $E(R_i \mid y_i, x_i) = E(R_i \mid x_i)$ or, alternatively, $\rho_{YX}(y_i, x_i) = \rho_X(x_i)$. It follows from (2.1) that we may write:

$$\rho_Y(y_i) = E[\rho_X(x_i) \mid y_i] \tag{2.2}$$

If this is the case, it follows that $\rho_Y(y_i)$ can, in principle, be determined from $\rho_i = \rho_X(x_i)$ and so all aspects of the response process relevant to nonresponse bias are captured by the ρ_i. In fact, if MAR holds, the definition $\rho_i = \rho_X(x_i)$ might be viewed as conservative since we have:

$$\begin{aligned} \text{var}(\rho_i) = \text{var}[\rho_X(x_i)] &= \text{var}\{E[\rho_X(x_i) \mid y_i]\} + E\{\text{var}[\rho_X(x_i) \mid y_i]\} \\ &= \text{var}[\rho_Y(y_i)] + E\{\text{var}[\rho_X(x_i) \mid y_i]\} \end{aligned} \tag{2.3}$$

Note that again, we are treating ρ_i as random with respect to the superpopulation model for x_i. The first term on the right hand side of (2.3) represents the variation of the conditional probabilities $\rho_Y(y_i)$, which we should ideally like to use. The second term represents additional variation which is unrelated to nonresponse 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 (2.3) 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 NMAR then (2.3) 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 $\text{var}(\rho_i)$ will be at least as large as $\text{var}[\rho_Y(y_i)]$. 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 $\text{var}[\rho_Y(y_i)]$ will necessarily be at least as large as $\text{var}(\rho_i)$, i.e. $\text{var}[\rho_X(x_i)]$ (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 at the beginning of this section that we define the response propensity conditional on the survey conditions that apply when the data (described in section 2.1) 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 main 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.

See Annex 1 for a fuller discussion of some of the above points and the assumptions underlying our definition of response propensities.

2.3 Nonresponse models

In order to estimate R-indicators, we shall first estimate the response propensities, where these are defined as $\rho_i = E(R_i \mid x_i)$, as discussed in the previous section. To enable these probabilities to be estimated we shall make certain parametric modelling assumptions about how ρ_i depends on x_i. In this section, we first discuss alternative parametric models and then provide some supplementary discussion of non-parametric models and some implications of the complexity of the model for the variability of the ρ_i.

A general class of models representing the dependence of ρ_i on x_i may be expressed in the form:

$$\mathrm{g}(\rho_i) = x_i'\beta, \tag{2.4}$$

where g(.) is a specified link function, β is a vector of unknown parameters to be estimated, and x_i may involve the transformation of the original auxiliary variables (e.g. by including interaction terms) for the purpose of model specification. For simplicity, we assume equal inclusion probabilities for all population units. A standard choice of link function is the logit function, leading to the logistic regression model:

$$\log[\rho_i/(1-\rho_i)] = x_i'\beta \tag{2.5}$$

Another link function with similar behaviour to the logit is the probit function. We shall also consider the use of the identity link function, which gives the 'linear probability model':

$$\rho_i = x_i'\beta, \tag{2.6}$$

since this will offer particular simplifications in the case of population-based auxiliary information.

Särndal and Lundström (2008) consider the reciprocal link function, which gives:

$$\rho_i^{-1} = x_i'\lambda, \tag{2.7}$$

and they refer to ρ_i^{-1} as the *influence* and denote it ϕ_i. They assume that the vector x_i is defined in such a way that there exists a constant vector c such that $c'x_i = 1$ for all $i \in U$. This restriction will in most practical situations be met and is effectively equivalent to assuming that a constant intercept term is included in the auxiliary information.

Särndal and Lundström (2008) view (2.7) as a hypothetical model which will not hold in practice and they instead focus on a finite population approximation to this model. This approximation is obtained by first defining a value λ_U of λ which achieves the best fit of model (2.7) in the finite population. For mathematical convenience, they define the fit as the weighted sum of squared differences $\sum_U \rho_i(\rho_i^{-1} - x_i'\lambda)^2$ and this is minimised when:

$$\lambda_U = (\sum_U \rho_i x_i x_i')^{-1} \sum_U x_i , \qquad (2.8)$$

provided x_i is defined so that the inverted matrix in (2.8) is non-singular. This implies that a finite population approximation to ϕ_i is given by:

$$\phi_{Ui} = x_i'(\sum_U \rho_i x_i x_i')^{-1} \sum_U x_i . \qquad (2.9)$$

We refer to these quantities as the *approximate influences*.

All the above approaches employ global parametric models. We could also consider the dependence of ρ_i on x_i in more nonparametric or local way. The simplest case is when the variables in x_i are categorical and define a fixed number of classes. In this case, we take ρ_i to be constant within classes and define these constant values as the limits, as the population size increases, of the response rates within the classes. In more general cases, we might imagine a nonparametric model $\rho(x)$, which is a smooth function of x, where $\rho_i = \rho(x_i)$ may be interpreted as the limiting response rate for a small neighbourhood consisting of population units with values of x close to x_i in some sense. Such a representation reveals a further modelling issue. We have already emphasized in section 2.2. the strong dependence of ρ_i on the choice of the auxiliary variables constituting x_i. In cases where an auxiliary variable is continuous or detailed, for example age in years or location by spatial coordinates, there is an additional potential dependence of the definition of ρ_i on the degree of detail of the auxiliary variable in the model. In the nonparametric set-up this corresponds to the size of the classes or small neighbourhoods or equivalently on the smoothness of the function $\rho(.)$. The smaller the neighbourhoods, the less the degree of smoothing and the greater the potential variation of ρ_i. For example, if the auxiliary variable is location then values of ρ_i representing response rates in areas of 10,000 inhabitants are likely to be more variable than values of ρ_i representing response rates in areas of 100,000 inhabitants. This corresponds to the impact of degree of complexity in a parametric model. The more complex the model becomes, for example via the introduction of additional auxiliary variables, including polynomial terms in a continuous auxiliary variable or more interaction terms, the more variable are the corresponding ρ_i values likely to be.

2.4 The selection of auxiliary information

We make a brief intermezzo to address the selection of auxiliary variables x_i, as it is important to stress that the availability and the choice of x_i have a strong influence on the values of $\rho_i = \rho_X(x_i)$. As a consequence, also the values of R-indicators may depend on x_i .

First, it is evident that when no auxiliary information is available, it is impossible to make any statement about the representativeness of the survey response. Moreover, the representativeness of different surveys can only be compared in relation to auxiliary information which is available in every survey. With the same information we mean auxiliary variables that have the same definitions and categories.

Second, the nonresponse model itself may be of influence even when the auxiliary information is the same. Different models may lead to different estimates for response propensities and, hence, potentially to different values of R-indicators. For this reason, ideally the nonresponse models should be the same when R-indicators are computed.

When the auxiliary information and nonresponse models are the same for different surveys, then one may still wonder which auxiliary variables to include in the models and which not. One may either fix a model beforehand or employ a variable selection algorithm based on some significance level or stopping rule. In this report we fix the auxiliary variables beforehand and do not select models.

In a forthcoming RISQ paper we will elaborate extensively on the relation between auxiliary information, nonresponse models and measures for representativity.

3. Definition of R-Indicators at the population level

Let $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_N)'$ denote the vector of response propensities in the population. Following Schouten et al. (2009), the representativity of the response mechanism may be measured by the variation between the ρ_i and in particular by the standard deviation of the response propensities given by:

$$S(\boldsymbol{\rho}) = \sqrt{\frac{1}{N-1}\sum\nolimits_U (\rho_i - \bar{\rho}_U)^2}\ , \tag{3.1}$$

where $$\bar{\rho}_U = \sum\nolimits_U \rho_i / N\ . \tag{3.2}$$

It may be shown that:

$$S(\boldsymbol{\rho}) \leq \sqrt{\bar{\rho}_U(1-\bar{\rho}_U)} \leq \frac{1}{2}\ .$$

Hence, transforming $S(\boldsymbol{\rho})$ to:

$$R(\boldsymbol{\rho}) = 1 - 2S(\boldsymbol{\rho}) \tag{3.3}$$

ensures that $0 \leq R(\boldsymbol{\rho}) \leq 1$ and, as discussed by Schouten et al. (2009), $R(\boldsymbol{\rho})$ defines an R-indicator which takes values on the interval [0,1] with the value 1 indicating the most representative response, where the ρ_i display no variation, and the value 0 indicating the least representative response, where the ρ_i display maximum variation.

Note that the minimum value of $R(\boldsymbol{\rho})$ depends on the response rate. For $\bar{\rho}_U = 1/2$, the minimum value of $R(\boldsymbol{\rho})$ is 0. For $\bar{\rho}_U = 0$ or $\bar{\rho}_U = 1$, no variation in ρ_i is possible and the minimum value of $R(\boldsymbol{\rho})$ is 1. In general, the minimum value which $R(\boldsymbol{\rho})$ may take is given by $1 - 2\sqrt{\bar{\rho}_U(1-\bar{\rho}_U)}$.

Särndal and Lundström (2008) define the following R-indicator:

$$Q^2(\boldsymbol{\rho}) = [\sum\nolimits_U \rho_i]^{-1}[\sum\nolimits_U \rho_i(\phi_{Ui} - \bar{\phi}_{\rho U})^2] \tag{3.4}$$

where $\bar{\phi}_{\rho U}$ is the ρ_i- weighted mean of the ϕ_{Ui} given by

$$\bar{\phi}_{\rho U} = (\sum\nolimits_U \rho_i)^{-1}(\sum\nolimits_U \rho_i \phi_{Ui})\ . \tag{3.5}$$

This quantity is a weighted variance of the approximate influences. We may expect its magnitude to be inversely related to the magnitude of $R(\boldsymbol{\rho})$. Thus, in very rough terms, we expect $R(\boldsymbol{\rho})$ to decrease and $Q^2(\boldsymbol{\rho})$ to increase as the variability of the ρ_i increases.

It is important to emphasize again (see section 2.2) that the definitions of the R-indicators depend very much on the choice of auxiliary variables x_i. Furthermore, as discussed at the end of section 2, the definitions depend upon the smoothness of the modelled dependence of ρ_i on x_i. We may expect that the smoother the model, the less heterogeneous will be the ρ_i and hence, for example, the larger $R(\boldsymbol{\rho})$ will be.

4. Estimation

4.1 Estimation of population totals from sample and respondent data

In the following sections, we shall use the fact that, for a given variable z_i, the design-weighted sample total $\sum_s d_i z_i$ is a design-unbiased estimator of the population total $\sum_U z_i$ and the design-weighted respondent total $\sum_r d_i z_i$ is an unbiased estimator of the ρ_i - weighted population total $\sum_U \rho_i z_i$. By design-weights we mean the sample inclusion weights, i.e. the reciprocals of the sample selection probabilities.

4.2 Estimation of nonresponse models

The estimation of the models in section 2.3 depends on the nature of the auxiliary information (see section 2.1). Here, we assume sample-based auxiliary information. In this case the model in (2.4) can be estimated from the data on respondents and nonrespondents by maximum pseudo likelihood (Skinner, 1989) i.e. the parameter vector β in this model may be estimated by the value $\hat{\beta}$, which solves:

$$\sum_s d_i [R_i - g^{-1}(x_i'\beta)] x_i = 0 \tag{4.1}$$

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 linear probability model in (2.6) can be estimated in closed form by ordinary least squares or by weighted least squares, where the weights are the design weights.

For the reciprocal link function model in (2.7), Särndal and Lundström (2008) approximate the model by $\rho_i^{-1} \approx x_i'\lambda_U$, where λ_U is defined in (2.8) and estimate this approximate model by estimating λ_U from the sample data by:

$$\hat{\lambda}_U = (\sum_r d_i x_i x_i')^{-1} \sum_s d_i x_i \, . \tag{4.2}$$

Note that this estimation follows the strategy in section 4.1 and that it also assumes sample-based auxiliary information.

In the case of a nonparametric model with constant values of ρ_i within classes defined by x , there are various ways of determining the classes. We consider a classification tree method based on the CART algorithm. This algorithm is implemented in the SPSS computing package. CART is a method that builds classes which are homogenous with respect to the response rate by carrying out successive binary splits of the sample according to the values of x_i. The splitting is continued until a further split does not enhance the prediction of R_i or a stopping rule is met based on a minimum sample size within the classes. For each of the final classes determined by the algorithm, ρ_i for units i within that class is estimated by the number of respondents in the class divided by the sample size in the class.

4.3 Estimation of response propensities

For the generalized linear model in (2.4), the usual estimator of the response propensity ρ_i is:

$$\hat{\rho}_i = g^{-1}(x_i'\hat{\beta}), \tag{4.3}$$

where $\hat{\beta}$ is the estimator of β obtained as discussed in the previous section.

In the case of the linear probability model in (2.6), if β is estimated by (design-) weighted least squares, the implied estimator of ρ_i is given by:

$$\hat{\rho}_i^{OLS} = x_i'(\sum_s d_i x_i x_i')^{-1} \sum_s d_i x_i R_i, \tag{4.4}$$

which may also be expressed as:

$$\hat{\rho}_i^{OLS} = x_i'(\sum_s d_i x_i x_i')^{-1} \sum_r d_i x_i . \tag{4.5}$$

In the approach of Särndal and Lundström (2008) with the reciprocal link function, ϕ_i is estimated by:

$$\hat{\phi}_i = x_i'\hat{\lambda}_U, \tag{4.6}$$

where $\hat{\lambda}_U$ is defined in (4.2), so that:

$$\hat{\phi}_i = x_i'(\sum_r d_i x_i x_i')^{-1} \sum_s d_i x_i \tag{4.7}$$

and the resulting estimator of ρ_i is $\hat{\phi}_i^{-1}$.

For the logit or probit link function, the estimator $\hat{\rho}_i$ obtained from (4.3) must fall in the feasible interval [0,1]. This is not necessarily the case for either the estimator based on the linear probability model in (4.5) or the estimator $\hat{\phi}_i^{-1}$ of ρ_i based on (4.7).

All the estimators above assume that sample-based auxiliary information is available. In deliverable 2 for this Workpackage, we shall explore what is feasible in the case of population-based auxiliary information. In particular, we note that $\sum_s d_i x_i$ and $\sum_s d_i x_i x_i'$ are unbiased for $\sum_U x_i$ and $\sum_U x_i x_i'$, respectively and that in large samples we may expect that $\sum_s d_i x_i \approx \sum_U x_i$ and $\sum_s d_i x_i x_i' \approx \sum_U x_i x_i'$. It follows from (4.5) that we may approximate $\hat{\rho}_i^{OLS}$ by:

$$\tilde{\rho}_i^{OLS} = x_i'(\sum_U x_i x_i')^{-1} \sum_r d_i x_i \,, \qquad (4.8)$$

and from (4.7) that we may approximate $\hat{\phi}_i$ by:

$$\tilde{\phi}_i = x_i'(\sum_r d_i x_i x_i')^{-1} \sum_U x_i \,. \qquad (4.9)$$

Expressions (4.8) and (4.9) provide estimators of the response propensity for respondents when x_i is not available for individual nonrespondents but aggregate population-level information is available. The estimator in (4.8) requires knowledge of the population sums of squares and cross-products $\sum_U x_i x_i'$ of the elements of x_i. The estimator in (4.9) only requires knowledge of the population total of each of the elements of x_i.

4.4 Estimation of R-indicators

Let $\hat{\rho}_i$ be an estimator of the response probability ρ_i, as discussed in the previous section. Assuming that sample-based auxiliary information is available, $\hat{\rho}_i$ may be computed for each $i \in s$. An estimator of $\bar{\rho}_U$ is then given by

$$\hat{\bar{\rho}}_U = (\sum_s d_i \hat{\rho}_i)/N \,. \qquad (4.10)$$

Alternatively, we could replace N in the denominator by $\sum_s d_i$. We estimate the R-indicator $R(\boldsymbol{\rho})$ by:

$$\hat{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1} \sum_s d_i (\hat{\rho}_i - \hat{\bar{\rho}}_U)^2} \qquad (4.11)$$

Again, we could replace $N-1$ in this expression by $\sum_s d_i$. We shall consider bias-adjusted versions of $\hat{R}(\boldsymbol{\rho})$ in section 5.1

If $\hat{\rho}_i$ is only available for respondents ($i \in$), as in the case of aggregated population-level auxiliary information described at the end of the previous section, a possible estimator of $R(\boldsymbol{\rho})$ is:

$$\hat{R}_r(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum\nolimits_r d_i \hat{\rho}_i^{-1}(\hat{\rho}_i - \hat{\bar{\rho}}_r)^2}$$

where $\hat{\bar{\rho}}_r = (\sum\nolimits_r d_i)/N$. This corrects for nonresponse bias using $\hat{\rho}_i^{-1}$- weighting. The validity of this correction depends on the validity of the estimates $\hat{\rho}_i$.

We now turn to the estimation of $Q^2(\boldsymbol{\rho})$ in (3.4). Särndal and Lundström (2008) propose the following estimator:

$$q^2 = [\sum\nolimits_r d_i]^{-1}[\sum\nolimits_r d_i(\hat{\phi}_i - \bar{\phi}_r)^2], \tag{4.12}$$

where $\hat{\phi}_i$ is defined in (4.7) and $\bar{\phi}_r = (\sum\nolimits_r d_i\hat{\phi}_i)/(\sum\nolimits_r d_i)$. They note that in fact $\bar{\phi}$ can be reexpressed as $\bar{\phi}_r = (\sum\nolimits_s d_i)/(\sum\nolimits_r d_i)$.

The estimator in (4.12) is based only upon respondent data. However, $\hat{\phi}_i$ in (4.6) does depend on $\sum\nolimits_s d_i x_i$ which may not be available in the case of aggregated population level information. In such cases, we may replace $\hat{\phi}_i$ in (4.12) by $\tilde{\phi}_i$ from (4.9).

5. Bias and confidence intervals

5.1 Bias and bias adjustment

We may expect the estimator $\hat{R}(\boldsymbol{\rho})$ defined in (4.11) to be biased downwards for $R(\boldsymbol{\rho})$, defined in (3.3), because of the sampling variation in the estimated values $\hat{\rho}_i$. We approximate the bias as follows. We write

$$\hat{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\hat{\Delta}}\,,$$

where $\hat{\Delta} = \sum_{i\in s} d_i(\hat{\rho}_i - \hat{\rho}_U)^2$ and $\hat{\rho}_U$ is defined in (4.10). We derive in Appendix 4 the following approximation:

$$E(\hat{\Delta}) = \Delta + \lambda_1 + \lambda_2\,,$$

where $\Delta = \sum_U (\rho_i - \bar{\rho}_U)^2$,

$$\lambda_1 = E\{\textstyle\sum_s d_i V(\hat{\rho}_i \mid s)\}\,,$$

$$\lambda_2 = -N\,\mathrm{var}_p(\bar{\rho}_s) + 2\bar{\rho}_U\,\mathrm{cov}(\hat{N}_s, \bar{\rho}_s)\,,$$

$$\hat{N}_s = \textstyle\sum_s d_i \text{ and } \bar{\rho}_s = N^{-1}\sum_s d_i\rho_i\,.$$

An estimator of λ_1 is $\hat{\lambda}_1 = \sum_s d_i\hat{V}(\hat{\rho}_i \mid s)$, where $\hat{V}(\hat{\rho}_i \mid s)$ is the estimator of $V(\hat{\rho}_i \mid s)$ given in Annex 2. In the case of constant weights $d_i = N/n$ we have

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

where $\nabla h(x_i'\hat{\beta}) = \exp(x_i'\hat{\beta})/[1+\exp(x_i'\hat{\beta})]^2$. In the case of constant weights we also have $\lambda_2 = -N\,\mathrm{var}_p(\bar{\rho}_s)$. Under simple random sampling we may write $\lambda_2 = -(n^{-1} - N^{-1})\Delta$. It follows that a bias corrected estimator of Δ in the case of simple random sampling is:

$$\tilde{\Delta} = \hat{\Delta} - \hat{\lambda}_1 - \hat{\lambda}_2 = (1 + n^{-1} - N^{-1})\hat{\Delta} - (N/n)\textstyle\sum_s \nabla h(x_i'\hat{\beta})^2 x_i'[\sum_{j\in s}\nabla h(x_j'\hat{\beta})x_j x_j']^{-1}x_i\,. \quad (5.1)$$

A bias-corrected estimator of $R(\boldsymbol{\rho})$ in this case is given by:

$$\tilde{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\tilde{\Delta}} . \tag{5.2}$$

We may also expect the estimated R-indicator q^2 to be biased upwards for the same reason. We do not have as extensive an analysis of the bias of q^2. A simple bias correction is obtained by estimating the bias as:

$$\hat{B}(q^2) = [\sum_r d_i]^{-1}[\sum_r d_i \hat{V}(\hat{\phi}_i)] , \tag{5.3}$$

where $\hat{V}(\hat{\phi}_i)$ is an estimator of the variance of $\hat{\phi}_i$. An expression for such a variance estimator is given in Annex 5.

5.2 Standard errors and confidence intervals

A linearization variance estimator for $\hat{R}(\boldsymbol{\rho})$ is derived in Annex 4. It depends on two components. The first treats $\hat{\beta}$ as fixed and may be expressed as a linearization estimator in a fairly straightforward way. The second term allows for the variance of $\hat{\beta}$. The expression in Annex 4 assumes that a logistic regression model is fitted.

A linearization variance estimator for q^2 is derived in Annex 5. This estimator follows the approach of Shao and Steel (1999), where the variance is estimated with respect to the sampling design treating the response indicators as fixed (given our assumption that the sampling and response processes are unfounded). Then, provided the sampling fraction is small, this variance estimator should be approximately unbiased for estimating the variance of q^2 with respect to both the sampling design and the response process.

An alternative approach would be to use a replication variance estimation method, such as the bootstrap or jackknife (Wolter, 2007; Shao and Tu, 1995). This would generally involve recomputing the estimated R-indicator as $\hat{R}_b$ for B replicate samples $b = 1, 2, ..., B$ and then forming the appropriate variance estimator. For example, the bootstrap estimator for the standard error of the R-indicator is:

$$s_R^{BT} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{R}_b^{BT} - \hat{\bar{R}}^{BT}\right)^2}$$

where $\hat{\bar{R}}^{BT} = \frac{1}{B}\sum_{b=1}^{B}\hat{R}_b^{BT}$.

Confidence intervals for an R-indicator can be constructed from a variance estimator $v[\hat{R}]$ in the usual way by assuming $\hat{R}$ is normally distributed. Thus, a $100(1-\alpha)\%$ confidence interval is given by $\hat{R} \pm \xi_{1-\alpha/2} v[\hat{R}]^{0.5}$, where $\xi_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

In the case of the bootstrap, it is also possible to obtain a $100\,(1-\alpha)\,\%$ confidence interval estimates directly, by ordering the estimates $\hat{R}$ for the different replicates and defining the confidence interval in terms of the $\alpha/2$ and $1-\alpha/2$ quantiles.

The bootstrap or jackknife may also be used to bias-correct $\hat{R}$. For the bootstrap, $2\hat{R}-\hat{\bar{R}}^{BT}$ is a bias-corrected estimate of the R-indicator (e.g. Efron and Tibshirani, 1993, sect 10.6).

In this paper we do not provide confidence intervals, but merely restrict ourselves to bias-adjusted estimates. This is mostly for practical reasons. In the forthcoming paper we will address confidence intervals extensively.

6. Relation of R-indicators to nonresponse bias

Suppose that y_i denotes a vector of survey variables of interest and consider estimation of the population mean $\bar{Y} = \sum_{i \in U} y_i / N$. A standard design-weighted estimator of θ_h which does not weight for nonresponse is:

$$\bar{y}_{dr} = \sum_{i \in s} d_i R_i y_i / \sum_{i \in s} d_i R_i$$

where d_i is the design weight. We evaluate the bias of $\bar{y}_d$ as an estimator of $\bar{Y}$ by taking its expectation with respect to the random sampling mechanism, denoted E_s, and with respect to the conditional distribution of R_i given y_i, denoted $E(R_i \mid y_i)$. We allow here for a general MNAR mechanism and write $\rho_Y(y) = E(R_i \mid y_i = y)$ (see discussion in section 2.2. We have:

$$EE_s(\bar{y}_{dr}) = EE_s\left(\sum_{i \in s} d_i R_i y_i / \sum_{i \in s} d_i R_i \right) \approx \sum_{i \in U} \rho_Y(y_i) y_i / \sum_{i \in U} \rho_Y(y_i), \qquad (6.1)$$

where the approximation is for large samples and we have used the assumption that the sampling and response mechanisms are unconfounded. Hence the bias of $\bar{y}_d$ depends on nonresponse only via $\rho_Y(y)$. It follows from (6.1) that

$$\begin{aligned} Bias(\bar{y}_{dr}) &\approx \sum_{i \in U} \rho_Y(y_i)[y_i - \bar{Y}] / \sum_{i \in U} \rho_Y(y_i) \\ &= corr_{\rho y} S(\mathbf{\rho}) S_y / \bar{\rho}_U, \qquad (6.2) \end{aligned}$$

where $corr_{\rho y} = (N-1)^{-1} \sum_{i \in U} [\rho_Y(y_i) - \bar{\rho}_U][y_i - \bar{Y}_h] / S_\rho S_y$, $S_y^2 = (N-1)^{-1} \sum_{i \in U} (y_i - \bar{Y})^2$ and $S(\mathbf{\rho})$ and $\bar{\rho}_U$ are defined in (3.1) and (3.2).

Expression (6.2) is also obtained in e.g. Bethlehem (1988) and Särndal and Lundström (2005). Using (3.3) and the fact that $corr_{\rho y} \le 1$, an upper bound for the bias is given by

$$| Bias(\bar{y}_{dr}) | \le S(\rho) S_y / \bar{\rho}_U = \frac{(1 - R(\rho)) S_y}{2 \bar{\rho}_U} \qquad (6.3)$$

The upper bound depends upon the survey item y. If y is binary, the maximum possible value for S_y is 0.5. Hence, in his case we also have the following bound:

$$Bias(\overline{y}_{dr}) \leq \frac{1 - R(\boldsymbol{\rho})}{4\overline{\rho}_U}.$$

7. Simulation studies of the properties of the estimated R-indicators

7.1 Design of simulation studies

In this section, we carry out simulation studies to assess the sampling properties of the two *R-indicators*:

- $\hat{R}(\boldsymbol{\rho})$, defined in (4.11);
- q^2, defined in (4.12).

The simulation studies are based on samples drawn from Census data. The sample designs are similar to some standard household and individual surveys carried out at National Statistics Institutes. The Census data is based on the 1995 20% Israel Census Sample containing 753,711 individuals aged 15 and over in 322,411 households. We used 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 nonresponse 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 consider both a survey with equal inclusion probabilities and a survey with different inclusion probabilities within strata.

For each type of survey, we carried out a two-step design to define response probabilities in the Census data. 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 Census from these probabilities. In the second step, we fitted a logistic regression model, as in (2.5), to these Census data and thus determined a ‘true’ response propensity for each unit as predicted by this model fitted to the population. The dependent variable of the model is the response indicator and the independent variables of the model the explanatory variables used in the first step. This two-step design ensures that we have a known model generating the response propensities 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 *R-indicators* were then estimated from the sample.

7.2 Results

Response probabilities are estimated for each sample drawn from the population. The smaller the sample size, the more difficult it is to obtain the correct model. For example, assuming that we know the correct logistic regression model that was used to generate the 'true' response probabilities in the population, applying this model to samples at different sampling rates results in higher variance for the coefficients as the samples get smaller. Figures 1 to 3 present histograms of the intercept coefficient under the correct logistic regression model for 1000 samples drawn according to the sampling rates: 1:50, 1:100 and 1:200. The true value is -1.926.

Figure 1: Histogram of Estimated Intercept for 'True' Logistic Regression Model (1000 samples drawn at 1:50). 'True' Intercept=-1.926

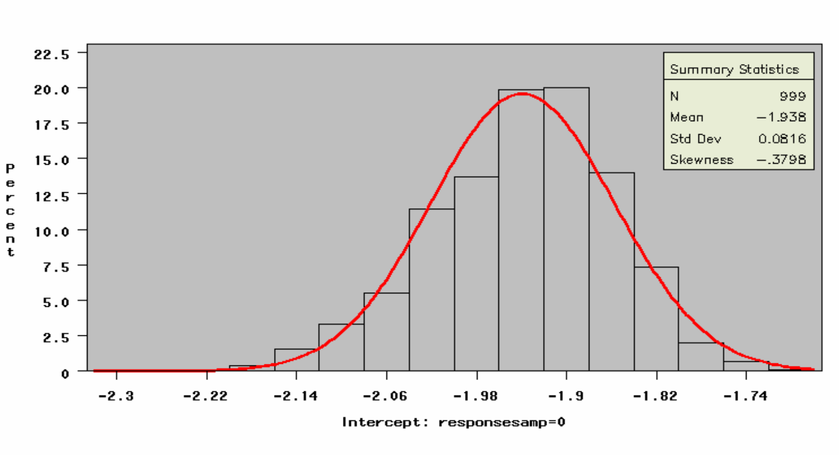


Figure 2: Histogram of Estimated Intercept for 'True' Logistic Regression Model (1000 samples drawn at 1:100) 'True' Intercept=-1.926

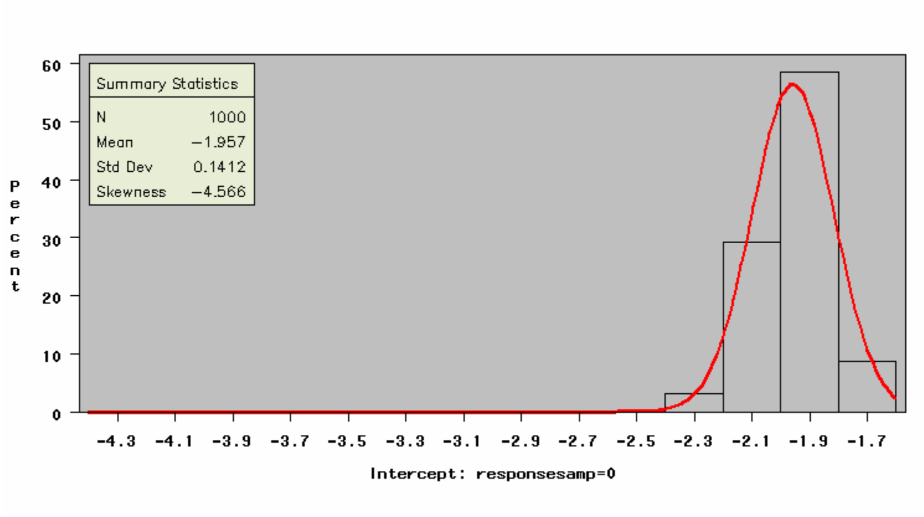
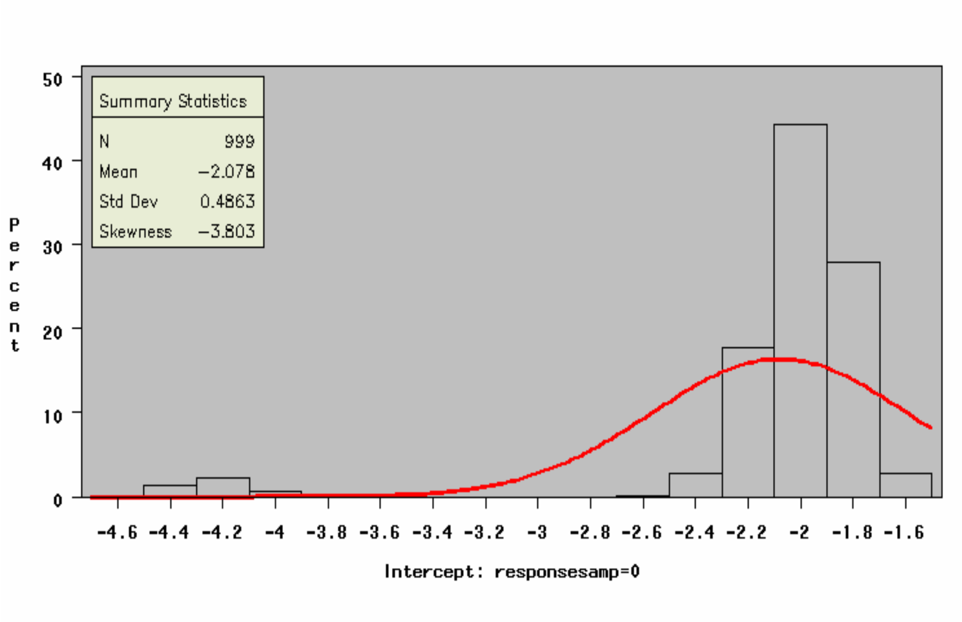


Figure 3: Histogram of Estimated Intercept for 'True' Logistic Regression Model (1000 samples drawn at 1:200) 'True' Intercept=-1.926



The figures show that the smaller the sample size, the more difficult it is to obtain the 'true' model in the sample. Figure 3 in particular shows the skewed distribution that is obtained for the intercept of the logistic model.

Throughout these simulations we examine the sampling properties of the *R-indicators* as well as the impact of model misspecification on their properties. Because smaller sample sizes generally lead to the selection of a less complex model, we shall consider that misspecification is represented by a simpler model.

In Table 1, we examine samples drawn for a Household Survey at different sampling rates, estimate response probabilities for each sample and calculate the measure $\hat{R}(\boldsymbol{\rho})$ (defined in (4.11) and its bias corrected version $\tilde{R}(\boldsymbol{\rho})$ (defined in (5.2).

We present results both when the true model is fitted and when a less complex model is fitted. In Table 2, we present results for the Individual Survey.

Table 1: Household Survey - Simulation Means of $\hat{R}(\boldsymbol{\rho})$ and its bias-corrected version, $\tilde{R}(\boldsymbol{\rho})$ for 500 Samples 'True' R-Indicator = 0.8780

Sampling Fraction	'True' Logistic Model (Number of Persons, Locality Type, Child Indicator)		Less Complex Logistic Model (Number of Persons)	
	$\hat{R}(\boldsymbol{\rho})$	$\tilde{R}(\boldsymbol{\rho})$	$\hat{R}(\boldsymbol{\rho})$	$\tilde{R}(\boldsymbol{\rho})$
1:200 (n=1,612)	0.8713	0.8831	0.8788	0.8868
1:100 (n=3,224)	0.8724	0.8781	0.8792	0.8831
1:50 (n=6,448)	0.8751	0.8779	0.8812	0.8831

Table 2: Individual Survey - Simulation Means of $\hat{R}(\boldsymbol{\rho})$ and its bias-corrected version, $\tilde{R}(\boldsymbol{\rho})$ for 500 Samples 'True' R-Indicator = 0.8767

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)	
	$\hat{R}(\boldsymbol{\rho})$	$\tilde{R}(\boldsymbol{\rho})$	$\hat{R}(\boldsymbol{\rho})$	$\tilde{R}(\boldsymbol{\rho})$
1:200 (n=3,769)	0.8537	0.8775	0.8944	0.9079
1:100 (n=7,537)	0.8652	0.8776	0.9009	0.9079
1:50 (n=15,074)	0.8705	0.8768	0.9028	0.9063

Tables 1 and 2 show that the estimator $\hat{R}(\boldsymbol{\rho})$ performs well in terms of explaining the bias. If the specified model is correct, there is some downward bias and this tends to increase as the sample size increases. 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. The degree of underestimation is, however, small in Tables 1 and 2. We observe that the bias correction is, however, very effective when the true model holds. The bias correction decreases with the increase in sample sizes and hence we obtain a stabilizing of $\tilde{R}(\boldsymbol{\rho})$. Using a less complex logistic model to estimate response probabilities results in a ‘smoothing’ of the probabilities and hence an overestimation in $\hat{R}(\boldsymbol{\rho})$ and the bias correction can exacerbate the overestimation.

In Table 3, we present results of an Individual Survey that has a survey design based on differential inclusion probabilities within strata. The aim is to see the impact of a more complex survey design on $\hat{R}(\boldsymbol{\rho})$ and $\tilde{R}(\boldsymbol{\rho})$. The logistic regression model used for estimating response probabilities in the samples is the ‘true’ model. The sample was stratified by three household sizes (1 person, 2 persons and 3 and over persons) and within each strata a different inclusion probability was defined. We observe that the estimators continue to be approximately unbiased.

Table 3: Individual Survey – Simulation Means of $\hat{R}(\boldsymbol{\rho})$ and $\tilde{R}(\boldsymbol{\rho})$ for 500 Samples with Differential Inclusion Probabilities -
‘True’ R-Indicator = 0.8767

	$\hat{R}(\boldsymbol{\rho})$	$\tilde{R}(\boldsymbol{\rho})$
Inclusion Probabilities (1:200 for 1 Person Households, 1:100 for 2 Person Households and 1:50 for 3 and over Person Households (n=10,966)	0.8670	0.8776

Besides using a logistic regression model to estimate response probabilities, we also used a non-parametric classification tree based on the CART algorithm. The variables used to carry out splits for the Household Survey were: number of persons, extended type of locality, child indicator, region, sex and age group of the head of household. For each terminal node, a response probability is estimated by the number of respondents in the node divided by the sample size in the node. This procedure is based on the saturated model.

We implemented CART classification tree algorithm on the Household Survey using two methods: the first based on one tree that was used for all samples which was calculated according to the ‘true’ response indicator in the population; and the second based on calculating a tree for each separate sample based on the sample response indicator. In each terminal node of the tree, response probabilities were estimated and R-indicators calculated. The results are presented in Table 4.

Table 4: Household Survey – Simulation Means of $\hat{R}(\boldsymbol{\rho})$ for 500 Samples Based on a Classification Tree (CART) - 'True' R-Indicator = 0.8780

	1:200	1:100	1:50
One tree for all Samples	0.8245	0.8572	0.8767
Different tree for each sample	0.7828	0.8146	0.8406

From Table 4, as the sample sizes increase, $\hat{R}(\boldsymbol{\rho})$ increases, i.e. the variance of the response probabilities decrease denoting a 'smoothing' of the response probabilities. The increase in $\hat{R}(\boldsymbol{\rho})$ is not as severe when using one tree for all possible sample sizes based on the 'true' population response propensities. Further work would need to apply a bias correction to this saturated model.

In Tables 5 and 6 we examine the properties of q^2 based on the variance of the response influences. For this indicator, we expect low values to reflect good quality and small nonresponse bias. We compare the full set of explanatory variables in the model used in this simulation to a less complex model as before.

Table 5: Household Survey - Simulation Means of q^2 for 500 Samples

Sampling Fraction	Full Model (Number of Persons, Locality Type, Child Indicator) 'True' R-indicator=0.0087	Less Complex Model (Number of Persons) 'True' R-Indicator=0.0082
	q^2	q^2
1:200 (n=1,612)	0.0103	0.0091
1:100 (n=3,224)	0.0096	0.0087
1:50 (n=6,448)	0.0089	0.0084

Table 6: Individual Survey - Simulation Means of q^2 for 500 Samples

Sampling Fraction	Full Model (Number of Persons, Locality Type, Child Indicator, Income Group, Sex and Age Group) 'True' R-indicator=0.0072	Less Complex Model (Number of Persons, Sex and Age Group) 'True' R-Indicator=0.0057
	q^2	q^2
1:200 (n=3,769)	0.0083	0.0066
1:100 (n=7,537)	0.0071	0.0061
1:50 (n=15,074)	0.0065	0.0057

Results from Tables 5 and 6 show the decrease in q^2 as the sample sizes increase. Further work would be to apply a bias correction for q^2. In addition, Workpackage WP4 will include calculation of confidence intervals for the *R-indicators*.

8. Country studies

One of the main objectives for constructing *R-indicators* is to enable the comparison of the representativeness of different surveys for given sets of auxiliary variables. In order to do so data sets were assembled from the participating countries in the RISQ research project: Belgium, The Netherlands, Norway, Slovenia and the UK. In this section the two R-indicators are estimated for these data sets given auxiliary variables that are shared by all countries.

Below is a short description of the data sets used. Note that more detailed information about the data sets can be found in RISQ Deliverable 1.

The Dutch Health Survey 2005

The Dutch Health Survey is a continuous survey of individuals with questions about health, life style and use of medical care. It consists of three questionnaires; a CAPI base module, a CAPI topical module about health and a supplementary paper questionnaire. The number of cases in the file is 15,411. The response rate was 67.3%.

Dutch Consumer Satisfaction survey 2005

The Consumer Confidence Survey is a continuous survey of households with questions about general economic development, and the financial situation of the household. The survey is meant to provide insight into short term economic development, and early indicators of differences in consumer trends. The number of cases in the file is 17,908. The response rate was 66.9%.

Dutch Short Term Statistic on Industry 2007

The Dutch Short Term Statistics on Industry is a monthly survey for Eurostat. It measures turnover for businesses in The Netherlands. The number of cases in the file is 64,413. The response rate was 92.5%

Dutch Short Term Statistic on Retail 2007

The Dutch Short Term Statistics on Retail is a monthly survey for Eurostat. It measures turnover for businesses in The Netherlands. The number of cases in the file is 93,799. The response rate was 92.3%.

UK 2001 Census Link File

The UK 2001 Census Link Study contains the response outcome of six major UK government household surveys linked to 2001 UK census data on a range of household and individual characteristics, interviewer observations about the household and extensive information about the interviewer and area information. The number of cases varies between surveys. For this report, we provide the *R-indicators* for the Labour Force Survey from May-June 2001, including all households that had a successful link with the Census data. The number of households in the dataset is 7,830 and the response rate about 80%.

Norwegian European Social Survey 2006

ESS is a biennial multi-country survey of individuals covering over 30 nations. It is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. The data set only contains the survey data of Norway. The number of cases in the file is 2,673. The response rate was 65.5%.

Norwegian Survey of Level of Living 2004

The survey of living conditions has two main purposes. One is to throw light on the main aspects of the living conditions in general and for various groups of people. Another purpose is to monitor development in living conditions, both level and distribution. Over a three-year period the cross-sectional survey of living conditions will cover all main areas of the living conditions. The survey topics change during a three-year cycle. Housing conditions, participation in organisations, leisure activities, offences and fear of crime were topics in 2004. It is a survey of individuals. The number of cases in the file is 4,837. The response rate was 69.1%.

Belgium European Social Survey 2006

As described for the Norwegian dataset, the ESS is an EU harmonized social survey. The data set contains the survey data of Belgium. The number of cases in the file is 2,927. The response rate was 61.4%.

Slovenian Labour Force Survey 2007

The Slovenian Labour Force Survey is an EU harmonized rotating panel survey conducted continuously through the year. The data contains employment related characteristics and demographic characteristics of all individuals 15 years or older living in selected households. The number of households varies between 7,010 and 7,160 households which is around 16,900 responding individuals. The response rate is around 80%.

Slovenian Survey on usage of information-communication technologies (ICT) in enterprises 2007

The Slovenian survey is an EU harmonized annual survey on the usage of ICT and provides information on whether the enterprises use computers, the internet, electronic commerce and other ICTs. The number of cases in the file is 1,998. The response rate is 87.6%.

We considered the following choices of auxiliary variables:

- Small fixed set. Selected variables are gender, age, interaction with degree of urbanization and region for household surveys, and size of business and business type for business surveys. In modelling response influences or response probabilities, the set is fixed and all variables included.

- Large fixed set. This is the small fixed set extended with a selection with relevant variables. The selection may be different for each country, however, all variables are included in models.
- Best fit using full set. All variables are candidate for inclusion in models. Only those variables are included that are significant according to pre-scribed level.

Table 7 presents initial results on some country datasets for the small fixed set of variables only. As can be seen from the table:

- there is consistency between the two indicators $\tilde{R}(\boldsymbol{\rho})$ and q^2 where a high $\tilde{R}(\boldsymbol{\rho})$ reflects in a low q^2,
- there is a correspondence between the response rates (and sample sizes) to the values of the indicators.

There are three business surveys in Table 7. For these business surveys, higher response rates generally produced a higher $\tilde{R}(\boldsymbol{\rho})$. The Netherlands Short Term Statistic on Retail is clearly showing less representatitivity than the Short Term Statistic on IB Industry in spite of having approximately the same response rate.

Table 7 does not contain confidence intervals. The estimates for the standard errors have not yet been implemented in software. In subsequent RISQ papers we will include approximate confidence intervals. Cobben and Schouten (2008) approximated 95% intervals using a naïve bootstrap estimator. They find standard errors of 2% for sample sizes close to 2000 and of 0.5% for sample sizes close to 30000. This would imply, for instance, that the Slovenian LFS values for $\tilde{R}(\boldsymbol{\rho})$ are not significantly different.

Note that for table 7 we fixed models beforehand. We did not select auxiliary variables but included all variables even when they gave no significant contribution. In such a setting model diagnostics are not relevant as long as numerical approximations of the estimates have converged. The R-indicator by itself is a measure for the lack of association between response and the selected auxiliary variables.

For the social surveys (households and individuals), Figures 4 and 5 provide a scatter-plot of each of the indicators $\tilde{R}(\boldsymbol{\rho})$ and q^2 against the response rates. As can be seen in the figures, the patterns are similar for both indicators, i.e. higher response rates reflect in higher $\tilde{R}(\boldsymbol{\rho})$ and lower q^2. The variability between the indicators, for example for surveys between 65% to 70% response rates, demonstrate that the response rate alone is not a sufficient quality indicator and that they should be combined with *R-Indicator*s to assess the bias that might incur from nonresponse. In Workpackage WP4 we will assess the *R-indicators* under different choices of auxiliary variables for all of the country datasets and how they can be used in practice.

Table 7: R-Indicators for Small Fixed Set of Variables for Country Datasets

	Sample Size	Response Rate	$\tilde{R}(\boldsymbol{\rho})$	q^2
Norway:				
European Social Survey 2006 (Individuals)	2,673	65.6%	0.762	0.044
Survey on Level of Living 2004 (Individuals)	4,837	69.1%	0.872	0.027
Slovania:				
LFS q3/2007 (Individuals)	2,219	70.1%	0.854	0.034
LFS q4/2007 (Individuals)	2,215	69.3%	0.807	0.057
LFS q1/2008 (Individuals)	2,247	68.2%	0.897	0.025
LFS q4/2007-q1/2008 (Households)	3,710	87.7%	0.951	0.002
ICT Survey 2007 (Enterprises)	1,998	87.6%	0.854	0.011
Netherlands:				
Short Term Statistic on IB Industry 2007 (Enterprises)	64,413	92.5%	0.933	-
Short Term Statistic on Retail 2007 (Enterprises)	93,799	92.3%	0.879	-
Health Survey 2005 (Individuals)	15,411	67.3%	0.832	0.029
Consumer Satisfaction Survey 2005 (Households)	17,908	66.9%	0.833	0.039
Belgium:				
European Social Survey 2006 (Individuals)	2,927	61.4%	0.807	0.074
UK:				
LFS May-June 2001 (Households)	7,830	80.5%	0.928	0.004

Figure 4: Scatterplot of $\tilde{R}(\boldsymbol{\rho})$ vs. Response Rates for Social Surveys in Table 7

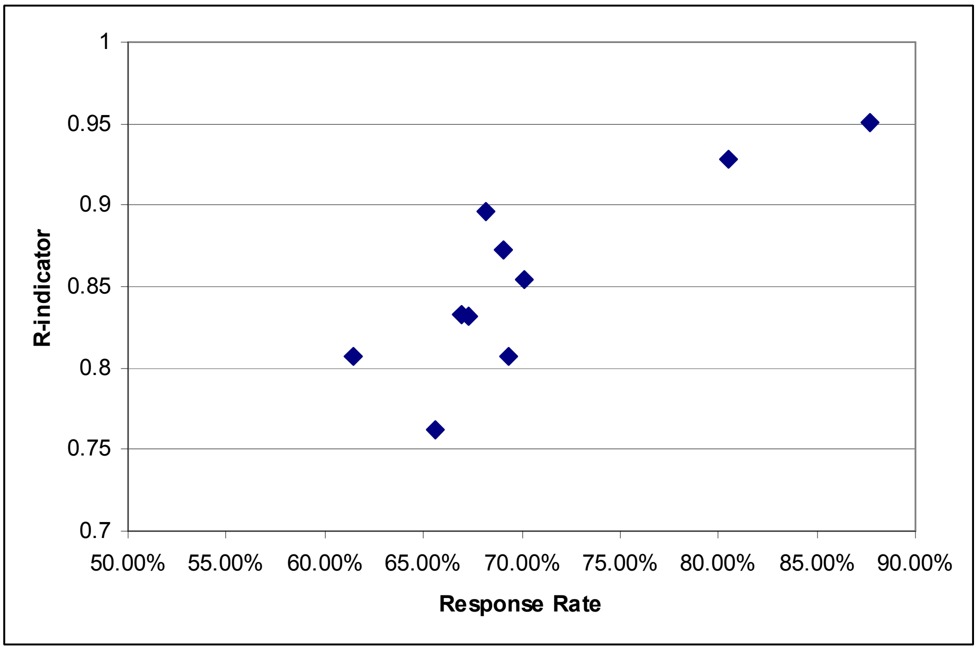
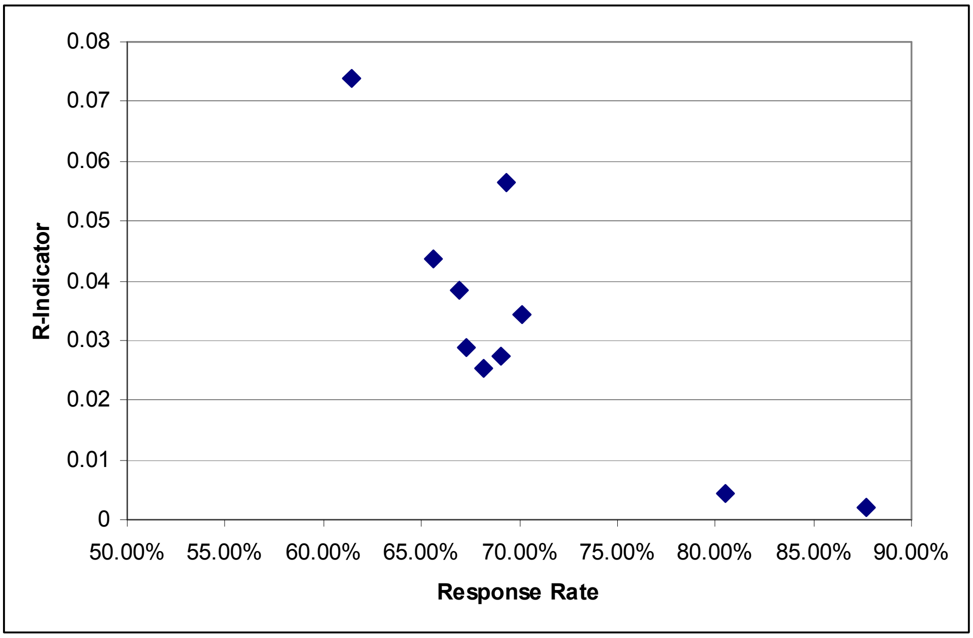


Figure 5: Scatterplot of q^2 vs. Response Rates for Social Surveys in Table 7



Acknowledgements: We thank the members of the RISQ project: Katja Rutar from Statistični Urad Republike Slovenije, Geert Loosveldt and Koen Beullens from Katholieke Universiteit, Leuven, Øyvin Kleven from Statistisk Sentralbyrå, Norway, Mattijn Morren and Fannie Cobben from Centraal Bureau voor de Statistiek, Netherlands, Ana Marujo and Gabi Durrant from the University of Southampton, UK for their valuable input and for calculating the *R-Indicators* on their country datasets.

Furthermore, we thank Paul Biemer (RTI International and University of North Carolina) for reviewing the first version of this paper. The current (second) version was made following his comments (Biemer 2009). Other comments will be addressed in subsequent RISQ deliverables.

Annex 1. More detailed discussion of definition of response probabilities

In this Annex, we provide a more detailed discussion of the definition of the response propensities introduced in section 2.2. We first outline, in section A1.1, two approaches to setting up statistical frameworks within which nonresponse can be postulated to arise in a stochastic way. This section describes sufficient conditions for it to be feasible to postulate the existence of response probabilities. Such conditions are not, however, sufficient to ensure that the definition of response probability is unique. In section A1.2, we discuss whether uniqueness can be achieved conceptually via consideration of the survey process. Our broad conclusion is negative. In section A1.3, we discuss how uniqueness can be achieved instead via some considerations of nonresponse bias. We argue, on the basis of bias considerations, that it is reasonable to define response probabilities as conditional probabilities of response given specified variables, which may either be auxiliary or survey variables. Given this specific definition, we refer to these conditional probabilities as *response propensities*. Given the discussion in section A1.2, we prefer to avoid assuming that unique 'true' response probabilities have an existence, separate from the specification of any variables upon which the probability is conditioned.

A1.1 Basic Approaches to Defining Response Propensities

The representation of nonresponse as a stochastic outcome has been postulated in the survey methodology literature on a number of occasions. For example, Platek et al. (1977) and Cassell et al. (1983) each provide formal introductions to the idea that each unit i in the population has a response probability ρ_i (when it is sampled) and different units respond independently. The definition of ρ_i can proceed in at least two ways.

Two-phase approach

Cassel et al. (1983) define ρ_i via the 'two-phase' model for nonresponse in which the response process is viewed as a second phase of sampling, where r is selected from s by a mechanism $q(.|s)$, so that the probability that the set of respondents consists of r conditional on s is given by $q(r|s)$. Oh and Scheuren (1983) refer to this as the 'quasi-randomization' approach to unit nonreesponse. If follows that the response propensity for a unit $i \in s$ is given by the inclusion probability of i with respect to $q(.|s)$, which may be denoted $\rho_{i|s}$ to emphasise that this probability may be conditional on s. If it can be assumed that this probability does not depend upon which sample was selected, then one can write $\rho_{i|s} = \rho_i$ for all $s \subset U$, where ρ_i is defined for all $i \in U$.

Cassel et al. (1983) also suppose that units respond independently of each other so that:

$$q(r \mid s) = \prod_{i \in r} \rho_i \prod_{i \in s-r} (1-\rho_i) .$$

This assumption can easily be extended, however, for example to the case where nonresponse is clustered, such as when all members of a household either respond or do not respond.

‘Census’ nonresponse approach

An alternative approach to defining ρ_i involves first postulating the existence of the response indicator variable R_i for each unit i in the population so that $r = \{i \in s \mid R_i = 1\}$, i.e. $R_i = 1$ if unit i responds (when it is sampled) and $R_i = 0$ if not, where $i \in U$. This is called the ‘stable response’ assumption by Rubin (1987, p.30). Fay (1991) (see also Shao and Steel, 1999) conceives of R_i as the outcome of a census, which would have been obtained from extending the survey to all the population. It is further assumed that the sampling mechanism is ‘unconfounded’ with the R_i (Rubin, 1987, p.36) so that the probability of selecting $s \subset U$ remains $p(s)$, whatever the values of the $R_i, i \in U$. Conversely, the R_i are the same whatever sample s is selected (of course, this cannot be checked since only one sample is observed). Given the existence of the response indicator variable R_i, if it is also assumed that these are random then the response propensity may be defined directly as $E(R_i) = \rho_i$.

The key assumption in either approach is that nonresponse does not depend on the ‘configuration’ of the sample. To give an example where this assumption might fail, consider a multi-stage sampling design which, for large households, only selects a fixed number (say three) individuals at random within the household to control costs. Whether one member of the household responds might depend upon which other members of the household are also selected and so $\rho_{i|s}$ might vary for different choices of s. In this case R_i is not well-defined since whether unit i responds depends not just on the condition that unit i is sampled, but also on which other units are sampled. Indeed the ‘census’ interpretation of Fay (1991) becomes problematic because an individual’s nonresponse behaviour when all members of the household are selected might differ from the individual’s behaviour for the actual sampling design.

Such an example seems unusual, however. In many surveys, sampled units will not know which other units have been asked to take part in the survey. And even in household surveys, it is usual either to select one person or all eligible persons from the household. Thus, the assumption that nonresponse does not depend on the configuration of the sample seems a fairly uncontentious assumption compared to other assumptions that may need to be made about the pattern of nonresponse.

A1.2 Can uniqueness be achieved conceptually via consideration of the survey process?

The previous section described sufficient conditions for it to be feasible to define a response probability. Even if such conditions apply, there still remains the question of whether further conditions are needed for ρ_i to be uniquely defined. In this section we address this question via consideration of the survey process.

It is clear that ρ_i (and indeed R_i) will depend upon the survey strategy employed and that its interpretation must therefore take account of this dependence. For example, if nonresponse includes non-contact then ρ_i will depend upon the number of contact attempts. Dalenius (1983) argues, however, that this dependence undermines the usefulness of the concept of response probability. He writes: 'it appears utterly unrealistic to postulate fixed "response probabilities" which are independent of the varying circumstances under which an effort is made to elicit a response. Whether an individual selected for a survey will respond or not may in many circumstances be determined by factors external to the individual". As illustration, he refers to the possible dependence of nonresponse upon the characteristics of the interviewer.

This dependence on circumstances may be unproblematic in relation to broad aspects of the survey strategy, such as the number of contact attempts. It seems more challenging, however, as the circumstances become more detailed. For example, it does not seem difficult to imagine that the probability of both non-contact and refusal may depend upon time of day or day of the week, as recognized in the case of non-contact by Politz and Simmons (1949). But, as Dalenius (1983) suggests, this implies that it may be more realistic to postulate a series of different values of ρ_i for an individual according to the survey circumstances than to postulate a unique value. And as the circumstances become more detailed, the greater the number of values of ρ_i that might be anticipated. Indeed, if one considers the potential dependence of ρ_i on the interviewer and one conceives of the interviewer as drawn from an effectively infinite population of possible interviewers, then one might imagine a distribution of possible values of ρ_i.

The question of whether ρ_i is uniquely defined also arises in debates (see e.g. Lessler and Kalsbeek, 1992, Ch. 7) about the choice between the 'random model' of nonresponse in which the indicators R_i are random with expectation ρ_i and the 'fixed model' where the indicators are treated as fixed, so that the population divides into one stratum of responders and one stratum of non-responders (e.g. Cochran, 1977, sect 13.2). The notion of a response probability clearly refers to the random model, but without further consideration of how the model is to be used, it does not seem possible to say that the random model is correct and the fixed model is false. Not only does the fixed model perfectly fit the nonresponse outcomes (i.e. the values of R_i for $i \in s$) but it might be argued, as above, that it arises as a special case of the random model (with $\rho_i = R_i$) as a consequence of conditioning on more and more detailed aspects of the survey circumstances.

One might seek to object to the fixed model on the grounds that it is inconceivable that a unit would respond in exactly the same way on repeated occasions and this argument might be extended to seek to define ρ_i as some kind of 'long run' proportion of times in which unit i would respond. However, this line of argument does not seem very promising. As noted earlier, response behaviour is likely to be very time-dependent and not adequately represented by a sequence of Bernoulli trials. And, if one has to define response indicators R_{it} for different times t then the problem simply multiplies. Moreover, in most practical survey contexts, it would appear very difficult to see how empirical evidence on the 'randomness' of responses could be obtained.

If one does not reject the fixed model, then one might argue that there exists a range of models which are equally valid in the sense that they all fit the data just as well (assuming that the data just consist of the values of R_i for $i \in s$) , but which vary according to their interpretation in terms of the degree to which they condition on the survey circumstances experienced by unit i. Thus the fixed model represents the most extreme degree of conditioning. A model which assigns a fixed response probability (matching the overall response rate) conditions least. The problem with this argument for the R-index is that it will vary considerably across this range of models, in fact across the whole range of its possible values.

In summary, the nature of the survey process does not appear to provide conceptual grounds upon which one can argue that a 'true' ρ_i is uniquely defined. In other words, without making further assumptions, such as about the relation between ρ_i and other variables, or setting further specific requirements for the R-index, the ρ_i do not appear to be identified (in an inferential sense) and so further assumptions are needed if they are to be estimated.

A1.3 Achieving Uniqueness via consideration of nonresponse bias

Now consider how a unique definition of ρ_i may be achieved by consideration of nonresponse bias. Suppose that y_i denotes a vector of survey variables of interest and we are concerned about the possible bias induced by nonresponse in the estimation of population parameters defined in terms of the $y_i, i \in U$. We argue: (1) that it suffices to define ρ_i as $E(R_i \mid y_i)$, i.e. the conditional probability of response given y_i, for this purpose and (2) $E(R_i \mid y_i)$ can be viewed as uniquely defined (subject to some caveats).

To make argument (1), suppose that we are interested in parameters which may be expressed as smooth functions of totals of the form $\sum_U h(y_i)$, where $h(.)$ is an arbitrary function. This class of parameters includes most of those considered in official statistics. Typical estimators of such parameters consist of the corresponding smooth function of estimated totals of the form $\sum_s w_i R_i h(y_i)$, where w_i is a survey weight. And the bias of such an estimator can generally be expressed as a function of terms $\sum_s w_i E(R_i \mid y_i) h(y_i)$ (i.e. the expectation of $\sum_s w_i R_i h(y_i)$ with respect to the response process conditional on y_i and s, assuming w_i is fixed under this distribution). Hence the nonresponse bias will depend upon nonresponse only via the conditional expectation $E(R_i \mid y_i)$, i.e. the conditional probability of response given y_i. Any stochastic variation in R_i which is not dependent upon y_i might contribute to variance but to not to bias. This is also illustrated in Section 6. Thus, by defining a *response propensity* ρ_i as $E(R_i \mid y_i)$, we may ensure that the definition includes any component of a response probability relevant to nonresponse bias for estimates based upon y_i. To emphasise the dependence on y_i, we write $\rho_Y(y) = E(R_i \mid y_i = y)$.

Let us now turn to argument (2) and consider whether $\rho_Y(y) = E(R_i \mid y_i = y)$ is uniquely defined. We can, in principle, imagine that y_i might be observed for both respondents and nonrespondents and conceive of the estimation of $\rho_Y(y)$ from such data. The simplest case is when the variables in y_i are categorical and define a fixed number of 'classes' in the population. In this case, for a large enough sample size, it should be possible (subject to a suitable sampling design and estimation method) to be able to estimate the population response rate in each class (i.e. the mean value of R_i among population units in this class) to any given precision. As the population size increases, we might define $\rho_Y(y)$ as the limit of the response rate, where y is the value of y_i for units in that class. In this case, $\rho_Y(y)$ is uniquely defined as, at least in principle, an estimable quantity.

This argument may be extended to cases when the y_i do not define a fixed number of classes, e.g. when one or more variables in y_i is continuous. In this case we might define $\rho_Y(y)$ as the limit of the population response rate, as the population size increases, within a fixed 'small neighbourhood' of y. For example, if y_i is the (continuous) age of individual i then $\rho_Y(y)$ might be defined as the response rate of individuals in the population within one year of age of y. Again, we might define ρ_i as $\rho_Y(y_i)$ and again this quantity should be estimable. As discussed at the end of section 2.3, there is a certain arbitrariness in the specification of what 'small' means in a small neighbourhood. Hence the uniqueness of ρ_i, if it is defined as $E(R_i \mid y_i)$, is subject to this degree of smoothness of the model being given.

A1.4 Defining Response Propensities which are Estimable

We argued in the previous section that $\rho_Y(y) = E(R_i \mid y_i = y)$ provides a means of defining a response probability uniquely in a way which is relevant to considerations of nonresponse bias. A basic problem with this definition, however, is that $\rho_Y(y)$ is generally only estimable in a direct way if y_i is observable for both respondents and nonrespondents. And, we are precisely interested in the situation when this assumption does not hold, i.e. when y_i is missing for nonrespondents. We propose instead to consider a vector x_i of auxiliary variables, which are observed for both respondents and nonrespondents (see section 2.1). In this case, $\rho_X(x) = E(R_i \mid x_i = x)$ is directly estimable. Moreover, a critical condition is whether

$$E(R_i \mid x_i, y_i) = E(R_i \mid x_i). \qquad \text{(A1.1)}$$

If condition (A1.1) holds, nonresponse is said to be missing at random (MAR). Otherwise, nonresponse is said to be not missing at random (NMAR). If (A1.1) holds then we may write

$$\rho_Y(y_i) = E(R_i \mid y_i) = E[E(R_i \mid x_i) \mid y_i] = E[\rho_X(x_i) \mid y_i] \qquad \text{(A1.2)}$$

It follows that $\rho_Y(y)$ becomes estimable under the MAR condition.

Annex 2. Variance of $\hat{\rho}_i$ for logistic regression model

The estimating equations in (4.1) may be expressed as:

$$\sum\nolimits_s d_i[R_i - h(x_i{'}\beta)]x_i = 0 \tag{A2.1}$$

where $h(\eta) = \exp(\eta)/[1+\exp(\eta)]$.

Let $\hat{\beta}$ solve (A2.1). Then in large samples we may approximate the distribution of $\hat{\beta}$ (c.f. Skinner, 1989) by:

$$\hat{\beta} \approx \tilde{\beta} + I(\tilde{\beta})^{-1}\sum\nolimits_s d_i[R_i - h(x_i'\tilde{\beta})x_i]$$

(A2.2)

where $\tilde{\beta} = p\lim(\hat{\beta})$ and $I(\beta) = \sum\nolimits_s d_i \nabla h(x_i{'}\beta)x_i x_i{'}$ is the information matrix and $\nabla h(\eta) = \partial h(\eta)/\partial\eta = h(\eta)[1-h(\eta)]$. In particular, we have

$$\mathrm{var}(\hat{\beta}) \approx I(\tilde{\beta})^{-1}\,\mathrm{var}\{\sum\nolimits_s d_i[R_i - h(x_i'\tilde{\beta})]x_i\} I(\tilde{\beta})^{-1} \tag{A2.3}$$

and, since $\hat{\rho}_i = h(x_i{'}\hat{\beta})$ from (4.3), we have

$$\mathrm{var}(\hat{\rho}_i \mid s) \approx \nabla h(x_i'\tilde{\beta})^2 x_i' \,\mathrm{var}(\hat{\beta})x_i = \nabla h(x_i'\tilde{\beta})^2 x_i' I(\tilde{\beta})^{-1}\,\mathrm{var}\{\sum_{j\in s} d_j[R_j - h(x_j'\tilde{\beta})]x_j \mid s\} I(\tilde{\beta})^{-1}x_i \tag{A2.4}$$

The above large sample argument may be applied within different inferential frameworks.

First, it may be applied in a purely design-based framework, where distributions are based only on the sampling design and $\tilde{\beta}$ is the limiting value of $\hat{\beta}$ with respect to the design, as both the size of the sample and the finite population increase. In this case, the population values of R_i are treated as fixed.

Second, this argument may be applied in a model-based framework, where the distributions are based only on the nonresponse model for a sequence of fixed samples with increasing size. In the latter case, $\tilde{\beta}$ is the limiting value of $\hat{\beta}$ with respect to the nonresponse model as the sample size increases. In particular, if the nonresponse model in (2.4) is assumed to be true then $\tilde{\beta}$ is the true value of β. If nonresponse is independent between different units we may write

$$\text{var}\{\textstyle\sum_s d_i[R_i - h(x_i'\tilde{\beta})]x_i \mid s\} = \textstyle\sum_s d_i^2 \nabla h(x_i'\beta)x_i x_i' \qquad \text{(A2.5)}$$

where the expression $\text{var}\{.\mid s\}$ is used to emphasise that this variance is with respect to the nonresponse process and is conditional on the choice of sample s. If the survey weights are constant so that $d_i = d$, the right hand side of (A2.5) is equal to $dI(\beta)$ and it follows from (A2.4) that we may write:

$$\text{var}(\hat{\rho}_i \mid s) \approx \nabla h(x_i'\beta)^2 x_i' [\sum_{j \in s} \nabla h(x_j'\beta) x_j x_j']^{-1} x_i \,. \qquad \text{(A2.6)}$$

Annex 3. Derivation of bias adjustment

Let $\hat{\Delta} = \sum_{i \in s} d_i (\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$ where $\hat{\bar{\rho}}_U$ is defined in (4.10). Write:

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

where $\bar{\rho}_s = N^{-1} \sum_s d_i \rho_i$ and $\bar{\rho}_U$ is defined in (3.2). Hence we have

$$\begin{aligned}(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2 = {} & (\hat{\rho}_i - \rho_i)^2 + (\rho_i - \bar{\rho}_U)^2 + (\bar{\rho}_U - \bar{\rho}_s)^2 + (\bar{\rho}_s - \hat{\bar{\rho}}_U)^2 \\ & +2(\hat{\rho}_i - \rho_i)(\rho_i - \bar{\rho}_U) + 2(\hat{\rho}_i - \rho_i)(\bar{\rho}_U - \bar{\rho}_s) + 2(\hat{\rho}_i - \rho_i)(\bar{\rho}_s - \hat{\bar{\rho}}_U) \\ & + 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_U - \bar{\rho}_s) + 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_s - \hat{\bar{\rho}}_U) + 2(\bar{\rho}_U - \bar{\rho}_s)(\bar{\rho}_s - \hat{\bar{\rho}}_U).\end{aligned}$$

We assume the estimator $\hat{\rho}_i$ is such that $E(\hat{\rho}_i \mid s) = \rho_i$, where $E(. \mid s)$ denotes expectation with respect to the response mechanism (holding the sample s fixed) . It follows that $E(\hat{\bar{\rho}}_U \mid s) = \bar{\rho}_s$ and further that:

$$\begin{aligned}E((\hat{\rho}_i - \hat{\bar{\rho}}_U)^2 \mid s, y_s, x_s) = {} & V(\hat{\rho}_i \mid s) + (\rho_i - \bar{\rho}_U)^2 + (\bar{\rho}_s - \bar{\rho}_U)^2 + V(\hat{\bar{\rho}}_U \mid s) \\ & -2Cov(\hat{\rho}_i, \hat{\bar{\rho}}_U \mid s) - 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_s - \bar{\rho}_U) \\ = {} & (\rho_i - \bar{\rho}_U)^2 + V(\hat{\rho}_i - \hat{\bar{\rho}}_U \mid s) + (\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\rho_i - \bar{\rho}_U)(\bar{\rho}_s - \bar{\rho}_U)\end{aligned}$$

It follows that

$$\begin{aligned}E(\hat{\Delta} \mid s) = {} & \sum_s d_i (\rho_i - \bar{\rho}_U)^2 + \sum_s d_i V(\hat{\rho}_i - \hat{\bar{\rho}}_U \mid s) \\ & + \hat{N}_s (\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\bar{\rho}_s - \bar{\rho}_U)(N\bar{\rho}_s - \hat{N}_s \bar{\rho}_U)\end{aligned} \qquad \text{(A3.1)}$$

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

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

$$E(\hat{\Delta}) = \Delta + \lambda_1 + \lambda_2 \qquad \text{(A3.2)}$$

where $\Delta = \sum_U (\rho_i - \bar{\rho}_U)^2$

$$\lambda_1 = E\{\sum_s d_i V(\hat{\rho}_i - \hat{\bar{\rho}}_U \mid s)\} \qquad \text{(A3.3)}$$

$$\lambda_2 = E\{\hat{N}_s (\bar{\rho}_s - \bar{\rho}_U)^2 - 2(\bar{\rho}_s - \bar{\rho}_U)(N\bar{\rho}_s - \hat{N}_s \bar{\rho}_U)\}$$

where λ_1 and λ_2 represent the bias in the estimation of Δ by $\hat{\Delta}$. We may consider approximating these terms using asymptotic arguments. If we treat d_i as of asymptotic order N/n, we see that both λ_1 and λ_2 are also of order N/n. We now consider simplifying these expressions by dropping lower order terms.

Starting with λ_1, note that $V(\hat{\bar{\rho}}_U \mid s)$ only generates a term of order N/n^2. Hence, we shall drop this term and write approximately:

$$\lambda_1 = E\{\textstyle\sum_s d_i V(\hat{\rho}_i \mid s)\}$$

Using the results in Annex 2 and assuming the nonresponse model is true, we may write :

$$\lambda_1 = E\{\textstyle\sum_s d_i \nabla h(x_i{}'\beta)^2 x_i{}'\mathrm{var}(\hat{\beta} \mid s) x_i\}\,.$$

Turning to the term λ_2, we may write

$$N\bar{\rho}_s - \hat{N}_s\bar{\rho}_U = N(\bar{\rho}_s - \bar{\rho}_U) - (\hat{N}_s - N)\bar{\rho}_U$$

Hence

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

and, ignoring terms of order less than N/n, we may write

$$\begin{aligned}\lambda_2 &= -NE\{(\bar{\rho}_s - \bar{\rho}_U)^2\} + 2\bar{\rho}_U E\{(\hat{N}_s - N)(\bar{\rho}_s - \bar{\rho}_U)\} \\ &= -N\,\mathrm{var}_p(\bar{\rho}_s) + 2\bar{\rho}_U\,\mathrm{cov}(\hat{N}_s, \bar{\rho}_s)\,,\end{aligned}$$

where the subscript p refers to the sampling design.

Annex 4. Variance of estimated R-indicator $\hat{R}(\boldsymbol{\rho})$ and variance estimation

As in section 5.1 we write

$$\hat{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\hat{\Delta}}\,,$$

where $\hat{\Delta} = \sum_{i \in s} d_i (\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$. As a linear approximation we have

$$\operatorname{var}[\hat{R}(\boldsymbol{\rho})] \approx N^{-1} E(\hat{\Delta})^{-1} \operatorname{var}(\hat{\Delta}) \qquad \text{(A4.1.)}$$

To approximate $\operatorname{var}(\hat{\Delta})$ we shall decompose the distribution of $\hat{\Delta}$ 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:

$$\Sigma = J(\tilde{\beta})^{-1} \operatorname{var}\{\textstyle\sum_s d_i [R_i - h(x_i' \tilde{\beta})] x_i\} J(\tilde{\beta})^{-1} \qquad \text{(A4.2)}$$

and $J(\beta) = E\{I(\beta)\}$ is the expected information rather than the observed information in (A2.3). 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

$$\operatorname{var}(\hat{\Delta}) = E_{\hat{\beta}}[\operatorname{var}_p(\hat{\Delta})] + \operatorname{var}_{\hat{\beta}}[E_p(\hat{\Delta})] \qquad \text{(A4.3)}$$

where the subscript p refers to the distribution induced by the sampling design and 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:

$$\operatorname{var}_p(\hat{\Delta}) \approx \operatorname{var}_p[\sum_{i \in s} d_i (\rho_i - \bar{\rho}_U)^2]\Big|_{\beta = \hat{\beta}}.$$

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

$$E_{\hat{\beta}}[\operatorname{var}_p(\hat{\Delta})] \approx \operatorname{var}_p[\sum_{i \in s} d_i (\rho_i - \bar{\rho}_U)^2]. \qquad \text{(A4.4)}$$

Turning to the second component in (A4.3), we may write:

$$E_p(\hat{\Delta}) \approx \sum_{i\in U}(\rho_i - \bar{\rho}_U)^2\Big|_{\beta=\hat{\beta}}$$

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

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

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

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

$$\mathrm{var}_{\hat{\beta}}[E_p(\hat{\Delta})] \approx 4A'\Sigma A + \mathrm{var}_{\hat{\beta}}\{tr[B(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \tilde{\beta})']\} \qquad \text{(A4.5)}$$

The sond term in (A4.5) can be further evaluated to

$$\begin{aligned}\mathrm{var}_{\hat{\beta}}\{tr[B(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \tilde{\beta})']\} &= \mathrm{var}\{\sum_i\sum_j B_{ij}(\hat{\beta} - \tilde{\beta})_i(\hat{\beta} - \tilde{\beta})_j\} \\ &= \mathrm{cov}[\sum_i\sum_j B_{ij}(\hat{\beta} - \tilde{\beta})_i(\hat{\beta} - \tilde{\beta})_j, \sum_k\sum_l B_{kl}(\hat{\beta} - \tilde{\beta})_k(\hat{\beta} - \tilde{\beta})_l] \\ &= \sum_i\sum_j\sum_k\sum_l B_{ij}B_{kl}[\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}] \\ &= 2tr[B\Sigma B\Sigma] \qquad \text{(A4.6)}\end{aligned}$$

where $A = \sum_{i\in U}(\rho_i - \bar{\rho}_U)(z_i - \bar{z}_U)$, $B = \sum_{i\in U}(z_i - \bar{z}_U)(z_i - \bar{z}_U)'$and Σ is defined in (A4.2). 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 $\hat{\Delta}$ may be estimated by the sum of the estimated components of (A4.3). The first of these appears in (A4.4) and may be estimated by a standard design-based estimator of $\mathrm{var}_p[\sum_{i\in s} d_i(\rho_i - \bar{\rho}_U)^2]$, where this is treated as the variance of a linear statistic $\text{var}_p[\sum_{i \in s} u_i]$ 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 (A4.5) and (A4.6). 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 $\hat{A} = \sum_{i \in s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)(\hat{z}_i - \hat{\bar{z}}_U)$, B may be estimated by $\hat{B} = \sum_{i \in s} d_i(\hat{z}_i - \hat{\bar{z}}_U)(\hat{z}_i - \hat{\bar{z}}_U)'$, 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 matrix of $\hat{R}(\boldsymbol{\rho})$ may be estimated by plugging the estimated variance of $\hat{\Delta}$ into (A4.1) and replacing $E(\hat{\Delta})$ by $\hat{\Delta}$.

Annex 5. Variance of estimated R-indicator q^2 and variance estimation

The estimated R-indicator q^2 was defined in (4.12). Särndal and Lundström (2008) show that it may be expressed alternatively as:

$$q^2 = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}),$$

where $\bar{m}_{r;d} = \sum_s d_i / \sum_r d_i$, $\bar{m}_{s;d} = \sum_s d_i \hat{\phi}_i / \sum_s d_i$ and $\hat{\phi}_i$ is defined in (4.7). It follows that we can write $q^2 = g(u)$, where $u = (u_1, u_2, u_3, u_4)$, $u_k = \sum_s d_i u_{ki}$, $k = 1,2,3,4$, $u_{1i} = 1$, $u_{2i} = R_i$, $u_{3i} = x_i$, $u_{4i} = R_i x_i x_i'$ and

$$g(u) = u_2^{-1} u_3 ' u_4^{-1} u_3 - u_1^2 u_2^{-2}.$$

Note that we abuse notation slightly by stacking two scalars, a vector and a matrix into u. We then linearize $g(u)$ to obtain:

$$g(u) \approx g(\mu) + \nabla_g [u - \mu]$$

where $\mu = E_s(u)$, $\nabla_g = \partial g(u) / \partial u$, evaluated at $u = \mu$ and E_s denotes expectation with respect to the sampling design. Hence $\mu_1 = N$, $\mu_2 = \sum_U R_i$, $\mu_3 = \sum_U x_i$, $\mu_4 = \sum_U R_i x_i x_i '$. We then obtain:

$$\begin{aligned} \nabla_g [u - \mu] = & -2\mu_1 \mu_2^{-2}(u_1 - \mu_1) \\ & +(-\mu_2^{-2}\mu_3 ' \mu_4^{-1}\mu_3 + 2\mu_1^2 \mu_2^{-3})(u_2 - \mu_2) \\ & +(2\mu_2^{-1}\mu_3 ' \mu_4^{-1})(u_3 - \mu_3) \\ & -\mu_2^{-1}\mu_3 ' \mu_4^{-1}(u_4 - \mu_4)\mu_4^{-1}\mu_3 \end{aligned} \qquad \text{(A5.1)}$$

Thus, we can approximate $\mathrm{var}_s(q^2)$ by $\mathrm{var}_s(\sum_s d_i z_i)$, where

$$\begin{aligned} z_i = & -2\mu_1 \mu_2^{-2} u_{1i} + (-\mu_2^{-2}\mu_3 ' \mu_4^{-1}\mu_3 + 2\mu_1^2 \mu_2^{-3}) u_{2i} \\ & +2\mu_2^{-1}\mu_3 ' \mu_4^{-1} u_{3i} - \mu_2^{-1}\mu_3 ' \mu_4^{-1} u_{4i} \mu_4^{-1}\mu_3 \end{aligned}$$

We then estimate $\mathrm{var}_s(q^2)$ by a conventional estimator of the variance of this linear statistic, with the values of μ replaced by u. Provided the sampling fraction is small this will also provide a suitable estimator of the variance of q^2 with respect to both the sampling design and the response process (Shao and Steel, 1999)

Finally, we provide an expression for an estimator $\hat{V}(\hat{\phi}_i)$ of the variance of $\hat{\phi}_i$ for use in (5.3). We write $\hat{\phi}_i = u_3 ' u_4^{-1} x_i$ and following a similar argument to above we approximate the variance of $\hat{\phi}_i$ by the variance of $(u_3 ' \mu_4^{-1} - \mu_3 ' \mu_4^{-1} u_4 \mu_4^{-1}) x_i$ which may be expressed as $(\sum_s d_j z_j) x_i$, where

$$z_i = x_i ' \mu_4^{-1} - R_i \mu_3 ' \mu_4^{-1} x_i x_i ' \mu_4^{-1}$$

We then set $\hat{V}(\hat{\phi}_i) = x_i ' \hat{V}(\sum_s d_i z_i) x_i$, where $\hat{V}(\sum_s d_i z_i)$ is a standard design-based estimator of the variance of the linear statistic $\sum_s d_i z_i$, where we plug in u_3 and u_4 for μ_3 and μ_4 respectively. Note that this results in z_i being replaced by $\hat{z}_i = x_i '(\sum_r d_i x_i x_i ')^{-1}(1 - R_i \hat{\phi}_i)$ in the variance estimator.

References

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

Biemer, P. (2009), Review of the paper Statistical properties of R-indicators, version 1, Note, RTI International and University of North Carolina, USA.

Cassel, C.M., Särndal, C-E. and Wretman, J.H. (1983) Some uses of statistical models in connection with the nonresponse problem. In Madow, W.G. and Olkin, I. eds. Incomplete Data in Sample Surveys, Vol. 3, Proceedings of a Symposium, New York: Academic Press,143-160.

Cobben, F. and Schouten, B. (2005), Bias measures for evaluating and facilitating flexible fieldwork strategies, Paper presented at 16th International Workshop on Household Survey Nonresponse, August 28-31, Tällberg, Sweden.

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

Cochan, W.G. (1977) Sampling Techniques, 3rd Ed, New York: Wiley.

Dalenius, T. (1983) Some reflections on the problem of missing data. In Madow, W.G. and Olkin, I. eds. Incomplete Data in Sample Surveys, Vol. 3, Proceedings of a Symposium, New York: Academic Press, 411-413.

Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. New York: Chapman and Hall.

Fay, R.E. (1991) A design-based perspective on missing data variance. In Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.

Lessler, J.T. and Kalsbeek, W.D. (1992) Nonsampling Errors in Surveys. New York: Wiley.

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 Buseiness and Economic Statistics, 6, 287-301.

Little, R.J.A. and Rubin,D.B. (2002) Statistical Analysis with Missing Data. Hoboken, NJ.: Wiley., Oh, H.L. and Scheuren, F.J. (1983) Weighting adjustment for unit nonresponse. In Madow, W.G. and Olkin, I. eds. Incomplete Data in Sample Surveys, Vol. 2, Proceedings of a Symposium, New York: Academic Press, 143-184.

Platek, R., Singh, M.P. and Tremblay, V. (1977) Adjustments for nonresponse in surveys. Survey Methodology, 3, 1-24

Politz, A.N. and Simmons, W.R. (1949) An attempt to get the "not at homes" into the sample without callbacks. Journal of the American Statistical Association, 44, 9-31.

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, John Wiley & Sons, Chichester, England.

Särndal, C-E and Lundström, S. (2008) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator, Journal of Official Statistics, 24, 167-191.

Schouten, B., Cobben, F. and Bethlehem, J. (2008) Indicators for the Representativeness of Survey Response. Survey Methodology (to appear)

Shao, J. and Steel, P. (1999) Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. Journal of the American Statistical Association, 94, 254-265.

Shao, J. and Tu,D. (1995) The Jackknife and the Bootstrap. New York: Springer.

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.

Wolter, K.M. (2007) Introduction to Variance Estimation. 2nd Ed. New York: Springer.