R-indexes for the comparison of different fieldword strategies and data collection modes

Discussion paper 07002

Barry Schouten and Fannie Cobben

Acknowledgements to Dirk Heerwegh, Geert Loosveldt, Bob Groves, Jelke Bethlehem and Jeroen Pannekoek

Statistics Netherlands

Voorburg/Heerlen 2007

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
?	= nil or less than half of unit concerned
0 (0,0)	= less than half of unit concerned
?	= (between two figures) inclusive
blank	= not applicable
2005?2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/'06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/'04?2005/'06	= crop year, financial year, etc. 2003/'04 to 2005/'06 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg

Prepress
Statistics Netherlands - Facility Services

Information
E-mail: infoservice@cbs.nl
Via contact form: www.cbs.nl/infoservice

Where to order
E-mail: verkoop@cbs.nl

Internet
www.cbs.nl

ISSN: 1572-0314

6008307002 X-10

R-INDEXES FOR THE COMPARISON OF DIFFERENT FIELDWORK STRATEGIES AND DATA COLLECTION MODES

Summary: Many survey organisations focus on the response rate as being the quality indicator for the impact of non-response bias. As a consequence they implement a variety of measures to reduce non-response or to maintain response at some acceptable level. However, in general response rates are not good indicators of non-response bias. Response rates only relate to the accuracy of estimators and also limits the maximal impact of non-response under the worst case scenario. In that respect it is worthwhile to keep non-response as small as possible. It is not true in general that higher response rates imply smaller non-response bias. The literature gives various counterexamples.

We introduce a number of concepts and indicators to assess the similarity between the response and sample of a survey. Such quality indicators, which we call R-indexes, may serve as counterparts to survey response rates and are primarily directed at evaluating the non-response composition. These indexes may facilitate analysis of survey response over time, between various fieldwork strategies or data collection modes.

We apply the R-indexes to practical examples for illustrational purposes. However, the main objective of this paper is to outline directions for future research.

1. Introduction

It is a well-developed finding in the survey methodological literature that response rates by themselves are poor indicators of non-response bias, see e.g. Curtin, Presser and Singer (2000), Groves and Heeringa (2005), Groves, Presser and Dipko (2004), Keeter et al. (2000), Merkle and Edelman (2002), Schouten (2004). However, the field has not proposed alternative indicators of non-response that may be more useful indicators of survey quality.

Non-response has two main consequences for survey estimates. First, it reduces the sample size, i.e. it decreases the precision of the estimates. Second, it deteriorates the sampling design. The inclusion probabilities that were chosen in the design do not longer hold as the propensity to response is not known. As a consequence non-response may also introduce bias to the estimates. The decreased precision can be dealt with by an increased sample size. Without any auxiliary information about the sample units, not much can be done, however, about the bias that is introduced. One needs to assume that non-respondents are on average the same as respondents when it amounts to the key survey topics. In case auxiliary information can be linked from administrative data or in case good auxiliary population statistics are available, then the corresponding auxiliary variables can be used to calibrate the response to sample or population totals. The auxiliary information can also be employed to analyze and measure the impact of the non-response as we will propose.

The survey literature contains a vast and still growing amount of analyses into non-response error. For general overviews see Groves et al. (2002) and Stoop (2005). These analyses make use of auxiliary population or sample totals of demographic and socio-economic characteristics of households. Although response behaviour depends on the topic of the survey, a number of characteristics has been identified that relate to lower response rates. Age, type of household and degree of urbanisation usually have a composition in the response that is different from the original sample.

With the analysis of non-response came the concept of a continuum-of-resistance, see e.g. Fitzgerald and Fuller (1982) and Lin and Schaffer (1995). Households are thought to behave along two dimensions, ease-of-contact and ease-of-participation. Attached to those dimensions are individual contact and response probabilities, and when combined overall individual response probabilities. Clearly, these probabilities are unknown but can be modelled using the available auxiliary information. Associated with the continuum-of-resistance is the level of effort of the survey organisation. The more effort the survey researcher invests in contacting households and converting reluctant respondents, the higher the response rate. It seems that the level of effort has increased during the past decades in order to maintain acceptable response rates. The level of effort represents costs and can be balanced to response rates, see Kalsbeek et al. (1994). One may also attempt to differentiate the level of effort between households to get a balanced composition of the response, see Groves and Heeringa (2005), Biemer and Link (2006) and Van der Grijn, Schouten and Cobben (2006).

The question arises whether increased efforts, apart from a higher response rate, also help enhancing the response, or in other words lead to a response that is more 'representative' of the sample. This has been investigated by e.g. Lynn et al (2002) and Stoop (2005). However, we, first, have to ask ourselves the question how we can assess an enhancement of response. What do we mean by representative? It is this question that we will focus attention on in this paper.

We propose indicators, which we will call R-indexes ('R' for representativity), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as representative response.

However, in the literature there are many different interpretations of the concept representativity. See Kruskal and Mosteller (1979 a, b and c) for a thorough investigation of the statistical and non-statistical literature. Some authors explicitly define representativity. Hajèk (1981) links “representative” to the estimation of population parameters. Following Hajèk’s definition, calibration estimators are representative for the auxiliary variables that are calibrated. Bertino (2006) defines a so-called univariate representativeness index for continuous random variables. This index is a distribution free measure based on the Cramér – Von Mises statistic.

We disconnect the concept representativity from the estimation of population parameters as we like to make indicators independent of the survey topics under investigation. By making indicators independent we hope they can be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. Also, we want to define a measure that gives a multivariate perspective of the dissimilarity between sample and response.

The R-indexes that we propose are either directly or indirectly based on estimated response probabilities. They differ in the way the response probabilities are estimated and employed. The estimation of response probabilities implies that R-indexes themselves are random variables, and, consequently, have a precision and possibly a bias. The sample size of a survey, therefore, plays an important role in the assessment of the R-indexes as we will see.

In order to be able to use R-indexes as tools for monitoring and comparing survey quality in the future, they need to have the features of a measure. That is we want the R-indexes to be interpretable, measurable and normalizable and also to satisfy the mathematical properties of a measure. Especially, the interpretation and normalization are no straightforward features. With this paper we have two objectives:

1. Identify a number of promising R-indexes
2. Outline the main issues for future research

In section 2, we start with a discussion of the concept representative response. Next, in section 3, we pose a number of R-indexes and so-called marginal R-indexes. Section 4 is devoted to the features of R-indexes. Finally, section 5 contains a discussion and plans for future research. We refer also to Cobben and Schouten (2005), Heerwegh and Loosveldt (2006) and Schouten and Cobben (2006).

2. The concept of representative response

We, first, discuss what it means that a survey respondent pool is representative of the sample. Next, we make the concept representativity mathematically rigorous by giving a definition.

2.1 What does representative mean?

Literature warns us not to single-mindedly focus on response rates as an indicator of survey quality, e.g. Groves (1989) or Biemer and Lyberg (2003). This can easily be illustrated by an example from the 1998 Dutch survey POLS (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English).

Table 2.1.1 contains the one and two month POLS survey estimates for the proportion of the Dutch population that receives a form of social allowance and the proportion that has at least one parent that was born outside the Netherlands. Both variables are taken from administrative data and are artificially treated as survey questions. The sample proportions are also given in table 2.1.1. After one month the response rate was 47%, while after the full period of interview of two months the rate was 60%. In the 1998 POLS the first month was CAPI (Computer Assisted Personal Interview). Non-respondents after the first month were allocated to CATI (Computer Assisted Telephone Interview) in case they had a registered, land-line phone. Otherwise, they were allocated once more to CAPI. Hence, the second month of interview gave another 12% of response. However, from table 2.1.1 we can see that after the second month the survey estimates have a larger bias than after the first month.

Table 2.1.1: Response means in POLS for the first month of interview and the full period of interview of two months.

Variable	*After 1 month*	*After 2 months*	*Sample*
Receiving social allowance	10,5%	10,4%	12,1%
Non-native	12,9%	12,5%	15,0%
Response rate	47,2%	59,7%	100%

From the example it seems clear that the increased effort led to a less representative response with respect to both auxiliary variables. But what do we mean by representative in general?

It turns out that the term representative is often used with hesitation in the statistical literature. Kruskal and Mosteller (1979 a, b and c) show that it is a garbage can for a lot of different interpretations. They make an extensive inventory of the use of the word in the literature and identify nine interpretations. A number of interpretations is omnipresent in the statistical literature. The interpretations that Kruskal and Mosteller named ‘absence of selective forces’, ‘miniature of the population’, and ‘typical or ideal cases’ relate to probability sampling, quota sampling and purposive sampling. In the next section we will propose a definition that corresponds to the ‘absence of selective forces’ interpretation. First we will motivate why we make this choice.

The concept of representative response is also closely related to the missing-data-mechanisms Missing-Completely-at-Random (MCAR), Missing-at-Random (MAR) and Not-Missing-at-Random (NMAR) that are often referred to in the literature, see Little and Rubin (2002). A missing-data-mechanism is MCAR in case the probability of response does not depend on the survey topic of interest. The mechanism is MAR if the response probability depends on observed data only, which is, hence, a weaker assumption than MCAR. If the probability depends on missing data also, than the mechanism is said to be NMAR. These mechanisms in fact find their origin in model-based statistical theory. Somewhat loosely translated, with respect to a survey topic, MCAR means that respondents are on average the same as non-respondents, MAR means that within known subpopulations respondents are on average the same as non-respondents, and NMAR is all but MAR. The addition of the survey topic is essential. Within one questionnaire some survey items can be MCAR, while other items are MAR or NMAR. Furthermore, the MAR assumption for one item holds for a particular stratification of the population. A different item may need a different stratification.

Conforming to the missing-data-mechanisms, Hajèk (1981) lets representativity be a feature of a sampling design and with respect to a variable, rather than letting representativity be a feature of the sampling method alone. He calls the couple sampling method and estimator representative with respect to some variable X if the estimator applied to X is equal to the population parameter almost surely. This definition can easily be extended to surveys with non-response and it implies that a calibration estimator is representative in conjunction with any sampling method with respect to the calibration variables. One may relax Hajèk's definition by replacing the almost sure equality by equality in expectation.

From the perspective that we wish to monitor and compare the response to different surveys in topic or time, it is not interesting to define a representative response as dependent on the survey topics itself nor as dependent on the estimator used. We focus on the quality of data collection and not on the estimation. We, therefore, compare the response composition to that of the sample. Clearly, the survey topics influence the probability that households participate in the survey, but the influence cannot be measured or tested and, hence, from our perspective this influence cannot be the input for an assessment of response quality. We propose to judge the composition of response by pre-defined sets of variables that are observed external to the survey and can be employed for each survey under investigation. We want the respondent selection to be as close as possible to a 'simple random sample of the survey sample', i.e. with as little relation as possible between response and characteristics that distinguish units from each other. The latter can be interpreted as that selective forces are absent in the selection of respondents or as MCAR with respect to all possible survey variables.

2.2 Definition of a representative response subset

Let $i = 1,2,3,\ldots,N$ be the unit labels for the population. By s_i we denote the 0-1-sample indicator, i.e. in case unit i is sampled it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit i. If unit i is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is n. Finally, π_i denotes the first-order inclusion probability of unit i.

The key to our definitions lies in the individual response propensities. Let ρ_i be the probability that unit i responds in case it is sampled.

The interpretation of a response propensity is not straightforward by itself. We follow a model-assisted approach, i.e. the only randomness is in the sample and response indicators. A response probability is a feature of a labelled and identifiable unit, a biased coin that the unit carries in a pocket so to say, and is, therefore, inseparable from that unit. With a little effort, however, all concepts can be translated to a model-based context.

First, we give a strong definition.

Definition (strong): *A response subset is representative with respect to the sample if the response propensities ρ_i are the same for all units in the population*

$$\rho_i = P[r_i = 1 \mid s_i = 1] = \rho, \quad \forall i, \tag{1}$$

and if the response of a unit is independent of the response of all other units.

This definition is similar to simple random sampling without replacement, except that the size of the response is not fixed but random. If a missing-data-mechanism would satisfy this definition then the mechanism would correspond to Missing-Completely-at-Random (MCAR) with respect to all possible survey questions. Although the definition is appealing the validity of it can never be tested in practice. We have no replicates of the response of one single unit. We, therefore, also construct a weaker definition that can be tested in practice.

Definition (weak): *A response subset is representative for a categorical variable X with H categories if the average response propensity over the categories is constant*

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad for \quad h = 1, 2, \ldots, H, \tag{2}$$

where N_h is the population size of category h, ρ_{hk} is the response propensity of unit k in class h and summation is over all units in this category.

The weak definition corresponds to a missing-data-mechanism that is MCAR with respect to X as MCAR states that we cannot distinguish respondents from non-respondents based on knowledge of X.

3. R-indexes and marginal R-indexes

In the previous section we defined strong and weak representative response. Both definitions make use of individual response probabilities that are unknown in practice. First, we start with the situation where these probabilities are known. Next, from there on we base the same R-indexes on estimated response propensities. Since we are not only interested in overall indicators of representativity, we also discuss so-called marginal R-indexes. However, we will only give the basic ideas behind marginal R-indexes and leave this topic to future papers. The section ends with illustration of the R-indexes to survey datasets.

3.1 R-indexes in case the individual response propensities are known

We, first, consider the hypothetical situation where the individual response propensities are known. Clearly, in that case we can even test the strong definition and we simply want to measure the amount of variation in the response propensities; the more variation the less representative in the strong sense. Let $\tilde{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)'$ be a vector of response propensities, let $\mathbf{1} = (1,1,\ldots,1)'$ be the N-vector of ones, and let $\rho_0 = \mathbf{1} \times \bar{\rho}$ be the vector consisting of the average population propensity.

Any distance function $d(\tilde{\rho}_1, \tilde{\rho}_2)$ in $[0,1]^N$ would suffice in order to measure the deviation from a strong representative response by calculation of $d(\tilde{\rho}, \rho_0)$. Note that the height of the overall response does not play a role.

A distance function or metric $d(\tilde{\rho}_1, \tilde{\rho}_2)$ must satisfy three properties:

1. $d(\tilde{\rho}_1, \tilde{\rho}_2) > 0$ and $d(\tilde{\rho}_1, \tilde{\rho}_2) = 0$ if and only if $\tilde{\rho}_1 = \tilde{\rho}_2$,
2. $d(\tilde{\rho}_1, \tilde{\rho}_2) = d(\tilde{\rho}_2, \tilde{\rho}_1)$,
3. $d(\tilde{\rho}_1, \tilde{\rho}_2) \le d(\tilde{\rho}_1, \tilde{\rho}_3) + d(\tilde{\rho}_3, \tilde{\rho}_2)$ or triangle inequality.

The Euclidean distance is a straightforward distance function. When applied to a distance between $\tilde{\rho}$ and ρ_0, this measure is proportional to the standard deviation of the response probabilities

$$S(\tilde{\rho}) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\rho_i - \bar{\rho})^2}\,. \tag{3}$$

It is not difficult to show that

$$S(\tilde{\rho}) \le \sqrt{\bar{\rho}(1-\bar{\rho})} \le \frac{1}{2}. \tag{4}$$

The first inequality in (4) follows by letting $\bar{\rho}N$ of the propensities be equal to 1 and all other $(1-\bar{\rho})N$ propensities be equal to 0. This gives a maximum variation of the propensities when fixing the average propensity. The second equality in (4) follows from taking $\bar{\rho} = 1/4$.

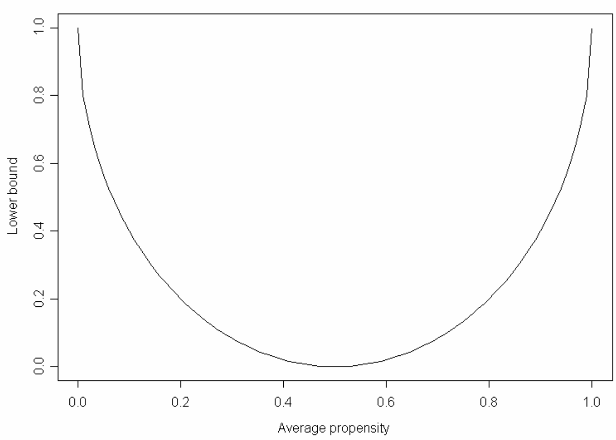
We want the R-index to take values on the interval [0,1] with the value 1 being strong representativity and the value 0 being the maximum deviation from strong representativity. A candidate R-index is

Alternative 1*: R-index based on standard deviation of response propensities*

$$R_1(\tilde{\rho}) = 1 - 2S(\tilde{\rho}). \quad (5)$$

Note that the minimum value of (5) depends on the response rate, see figure 3.1.1. For $\bar{\rho} = 0{,}5$ it has a minimum value of 0. For $\bar{\rho} = 0$ and $\bar{\rho} = 1$, clearly, no variation is possible and the minimum value is 1. Paradoxically, the lower bound increases in case the response rate decreases from 0,5 to 0. For a low response rate there is less room for individual response propensities to have a large variation.

Figure 3.1.1: Minimum value of R-index (5) as a function of the average response propensity.



Instead of the standard deviation one may use the variance of the response propensities. This was an R-index originally proposed by Cobben and Schouten (2005).

Alternative 2*: R-index based on variance of response propensities*

$$R_2(\tilde{\rho}) = 1 - 4S^2(\tilde{\rho}). \quad (6)$$

We will show that R_1 has a close relation to the well-known χ^2-statistic that is often used to test independence and goodness-of-fit. Suppose that the response propensities are only different for classes h defined by a categorical variable X. Let $\bar{\rho}_h$ and f_h be, respectively, the response propensity and the population function of class h, i.e.

$$f_h = \frac{N_h}{N}, \quad for \quad h = 1,2,\ldots,H. \tag{7}$$

Hence, for all i with $X_i = h$ the response propensity is $\rho_i = \overline{\rho}_h$.

Since the variance of the response propensities is the sum of the 'between' and 'within' variances over classes h, and the within variances are assumed to be zero it holds that

$$S^2(\tilde{\rho}) = \frac{1}{N-1}\sum_{h=1}^{H} N_h(\overline{\rho}_h - \overline{\rho})^2 = \frac{N}{N-1}\sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2 \approx \sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2. \tag{8}$$

The χ^2-statistic measures the distance between real and expected proportions. However, only for fixed marginal distributions f_h and $\overline{\rho}$ it is a true distance function in the mathematical sense. We can apply the χ^2-statistic to X in order to 'measure' the distance between the true response behaviour and the response behaviour that is expected in case response is independent of X. In other words we measure the deviation from weak representativity with respect to X.

We can rewrite the χ^2-statistic to get

$$\begin{aligned}
\chi^2 &= \sum_{h=1}^{H} \frac{(N_h\overline{\rho}_h - N_h\overline{\rho})^2}{N_h\overline{\rho}} + \sum_{h=1}^{H} \frac{(N_h(1-\overline{\rho}_h) - N_h(1-\overline{\rho}))^2}{N_h(1-\overline{\rho})} \\
&= \sum_{h=1}^{H} \frac{Nf_h(\overline{\rho}_h - \overline{\rho})^2}{\overline{\rho}} + \sum_{h=1}^{H} \frac{Nf_h(\overline{\rho}_h - \overline{\rho})^2}{(1-\overline{\rho})} \\
&= \frac{N}{\overline{\rho}(1-\overline{\rho})} \sum_{h=1}^{H} f_h(\overline{\rho}_h - \overline{\rho})^2 \\
&= \frac{N-1}{\overline{\rho}(1-\overline{\rho})} S^2(\tilde{\rho}).
\end{aligned} \tag{9}$$

An association measure, see e.g. Agresti (2002), that transforms the χ^2-statistic to the [0,1] interval is Cramèr's V

$$V = \sqrt{\frac{\chi^2}{N(\min\{C,R\}-1)}}, \tag{10}$$

where C and R are, respectively, the number of columns and rows in the underlying contingency table. Cramèr's V attains a value 0 if observed proportions exactly match expected proportions and it's maximum is 1. In our case the denominator equals N, since the response indicator has only two categories, response and non-response. As a consequence, (10) changes into

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\overline{\rho}(1-\overline{\rho})}} S(\tilde{\rho}). \tag{11}$$

From (11) we can see that for large N Cramèr's V is approximately equal to the standard deviation of the response propensities standardized by the maximal standard deviation $\sqrt{\overline{\rho}(1-\overline{\rho})}$ for a fixed average response propensity $\overline{\rho}$. It becomes clear from a simple example that (11) is not a distance function unless we fix $\overline{\rho}$. For

- $f_1 = f_2 = 0{,}5$, $\overline{\rho}_1 = 0{,}9$ and $\overline{\rho}_2 = 1$,
- $f_1 = f_2 = 0{,}5$, $\overline{\rho}_1 = 0{,}45$ and $\overline{\rho}_2 = 0{,}55$,

the standard deviation (3) is the same, but (11) is bigger in the first case. Since we do not want an R-index to depend on the average response probability, we work with either (5) or (6).

3.2 R-indexes in practice

In the previous section we assumed that we know the individual response propensities. Of course, in practice these propensities are unknown. Furthermore, in a survey we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R_1 and R_2. An obvious way to do this, is to use response-based estimators for the individual response propensities and the average response propensity.

We let $\hat{\rho}_i$ denote an estimator for ρ_i that uses all or a subset of the available auxiliary variables. Methods that support such estimation are for instance logistic or probit regression models (Agresti 2002) and CHAID classification trees (Kass 1980). By $\hat{\overline{\rho}}$ we denote the weighted sample average of the estimated response propensities, i.e.

$$\hat{\overline{\rho}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_i \frac{s_i}{\pi_i}, \tag{12}$$

where we use the inclusion weights.

We replace R_1 and R_2 by the estimators $\hat{R}_1$ and $\hat{R}_2$, where π_i denotes the inclusion probability of unit i:

Alternative 1: *The R-index based on the estimated standard deviation of the response propensities*

$$\hat{R}_1(\tilde{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\overline{\rho}})^2}\ . \tag{13}$$

Alternative 2: *The R-index based on the estimated variation of the response propensities*

$$\hat{R}_2(\tilde{\rho}) = 1 - 4\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\overline{\rho}})^2\ . \tag{14}$$

Note that in (13) and (14) there are in fact two estimation steps based on different probability mechanisms. The response propensities themselves are estimated and the variation in the propensities is estimated. We return to the consequences of the two estimation steps in section 4.

We add a third alternative R-index that relates to measures for the proportional reduction of error, see e.g. Goodman and Kruskal (1979). In linear regression a measure that is often used is the proportion of variance explained by the model, the $\Re^2$-statistic. Analogous to this statistic so-called pseudo $\Re^2$-statistics have been developed for categorical dependent variables. Many of them are based on the likelihood function. We assume that a logit or probit model is used for the estimation of response propensities and that inclusion probabilities are equal for all units. The case of unequal inclusion probabilities can be dealt with, but complicates the analysis.

Let L_0 be the likelihood of the 'empty' regression model with only an intercept and L_1 be the likelihood function of the regression model that is used to estimate response propensities. One of the measures that is often used, is Nagelkerke's pseudo $\Re^2$. It is defined as

$$\Re^2 = \left(\frac{1 - \left(\frac{L_0}{L_1} \right)^{2/n}}{1 - (L_0)^{2/n}} \right)^2 . \tag{15}$$

From the perspective of representativity, we want a regression model to be unable to reduce the prediction error of the individual response propensity. This is a paradox, as in analysis one aims at the opposite. However, we want the response propensities to be equal, and, hence, want to find a 'constant' error. R-index $\hat{R}_3$, therefore, has the following form

Alternative 3: *The R-index based on the proportional reduction of error*

$$\hat{R}_3(\tilde{\rho}) = 1 - \Re^2 \tag{16}$$

R-index (16) is somewhat different from the other two R-indexes in that it is not directly related to a distance function for the response propensities.

The likelihoods in (15) can be written as functions of estimated response propensities

$$L_0 = \hat{\bar{\rho}}^{n\hat{\bar{\rho}}} (1 - \hat{\bar{\rho}})^{n(1-\hat{\bar{\rho}})}, \tag{17}$$

and

$$L_1 = \prod_{h=1}^{H} \hat{\bar{\rho}}_h^{\,n_h \hat{\bar{\rho}}_h} (1 - \hat{\bar{\rho}}_h)^{n_h (1-\hat{\bar{\rho}}_h)}, \qquad (18)$$

with $\hat{\bar{\rho}}_h$ the average estimated response propensity in class h of variable X.

3.3 Marginal R-indexes

The three indicators defined in section 3.2 give overall views. We are, however, especially interested in the groups in the population that relate to the perceived dissimilarity between the response subset and the sample. We, therefore, need indicators that measure dissimilarity of single auxiliary variables conditional on the other auxiliary variables. We call such indicators marginal R-indexes.

We let X be an auxiliary variable, possibly but not necessarily an element of the set of auxiliary variables used to estimate response propensities. We restrict ourselves to categorical X and let X have H classes.

We propose two marginal R-indexes:

Alternative 1: *Marginal R-index MR_1 for X is based on estimated deviations from the average response propensity of the subclasses defined by X*

$$MR_1(h, \tilde{\rho}) = f_h \left(\frac{\hat{\bar{\rho}}_h}{\hat{\bar{\rho}}} - 1 \right) \; for \; h = 1,2,3,\ldots,H. \qquad (19)$$

The indicator (19) is proposed by Heerwegh and Loosveldt (2006) and corresponds directly to the non-response bias of the proportion of the population falling in class h, denoted by f_h. It does not account for the dependence between auxiliary variables. It can, however, be computed for any auxiliary variable.

The second marginal R-index that we propose attempts to account for the dependence between auxiliary variables, i.e. it adjusts for dissimilarity caused by auxiliary variables other than the variable X under investigation.

Alternative 2: *The marginal R-index MR_2 for X is the estimated centralized regression parameter for the subclasses defined by X in logistic regression*

$$MR_2(h, \tilde{\rho}) = centralized \; regression \; parameters \; for \; category \; h. \qquad (20)$$

Marginal R-index MR_2 makes use of so-called centralized regression parameters. In logistic regression the first or last category is usually taken as the reference category. This choice is arbitrary. However, from the viewpoint of statistical inference it is good practice to take a large category as the category of reference. In our case we do not want to choose a reference category as we want to give a picture of all categories simultaneously. Hence, we want the regression parameters to be free of a reference category. In other words we want to centralize the parameters so that on average they are zero. In the appendix we describe in detail how this transformation can be done. Here we only give an example.

Table 3.3.1: An example of the transformation of regression parameters to centralized regression parameters for a logistic regression model containing an intercept and the main effects of three covariates age, region and ethnic background

Variable and category		*Regression parameter*	*Centralized parameter*
Age			
	0-34 years	0	0,05
	35-54 years	-0,02	0,03
	> 55 years	-0,14	-0,01
Ethic background			
	Native	0	0,42
	Moroccan	-0,82	-0,40
	Turkish	-0,73	-0,32
	Surinam	-0,28	0,14
	Dutch Antilles	-0,34	0,08
	Other non-western	-0,67	-0,25
	Other western	-0,10	0,32
Region			
	Groningen	0	0,21
	Friesland	0,15	0,36
	Drenthe	0,11	0,32
	Overijssel	0,22	0,43
	Flevoland	-0,01	0,20
	Gelderland	0,04	0,25
	Utrecht	-0,26	-0,05
	Noord-Holland	-0,17	0,04
	Zuid-Holland	-0,19	0,02
	Zeeland	0,11	0,32
	Noord-Brabant	0,02	0,23
	Limburg	0,15	0,36
	Amsterdam	-1,04	-0,83
	Rotterdam	-0,72	-0,51
	Den Haag	-0,81	-0,60
	Utrecht	-0,96	-0,75
Intercept		0,59	-0,10

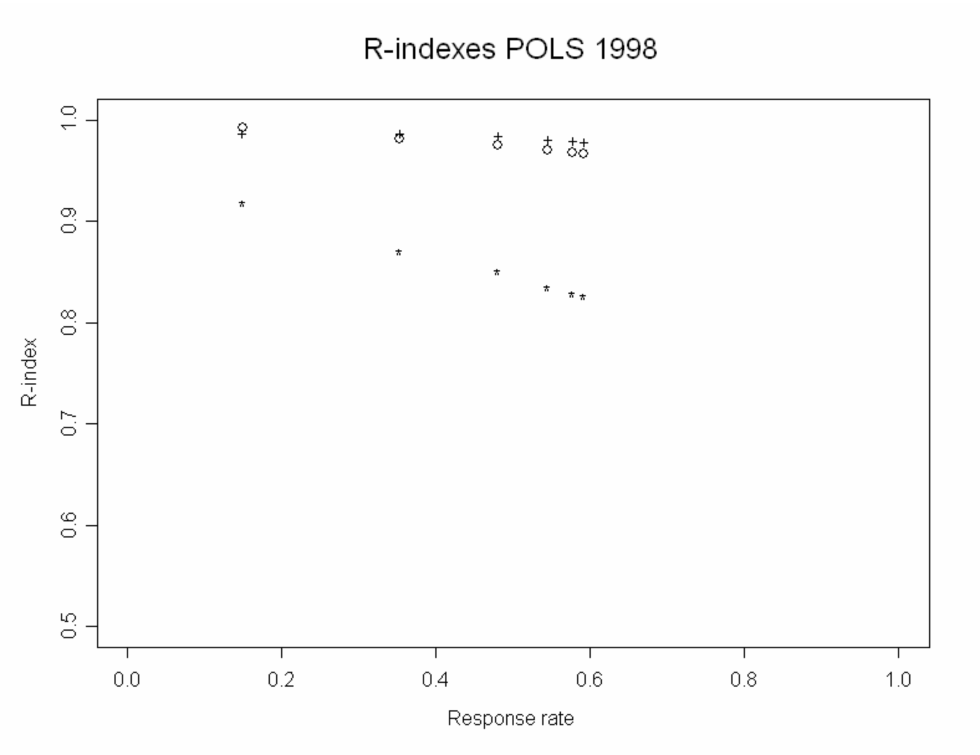
Suppose that response propensities are estimated using a logistic regression model containing an intercept and three covariates: age in three categories, region in 16 categories and ethnic background in 7 categories. We will use the same covariates in section 3.4. Table 3.3.1 shows the regression parameters and the centralized regression parameters.

For indicators (20) it is implicitly assumed that X is used as an explanatory variable in the estimation of the response propensities. If it is not, than we assume that X does not give a significant contribution to the regression model and, hence, does not cause a selective response and $MR_2(h, \widetilde{\rho}) = 0$ for all categories.

3.4 Example

We applied the proposed R-indexes and marginal R-indexes to the survey data from the 1998 POLS that we described in section 2.1. Recall that the survey was a combination of face-to-face and telephone interviewing in which the first month was CAPI only. The sample size was close to 40.000 and the response rate was approximately 60%. We linked the fieldwork administration to the sample and deduced for each contact attempt whether it resulted in a response. This way we can monitor the traces of the R-indexes during the fieldwork period.

Figure 3.4.1: R-indexes for first six contact attempts in POLS 1998. $\hat{R}_1$, $\hat{R}_2$ *and* $\hat{R}_3$ *are depicted as '*', 'O' and '+'.*



For the estimation of response rates we used a logistic regression model with region, ethnic background and age as independent variables. Region was a classification with 16 categories, the 12 provinces and the four largest cities Amsterdam, Rotterdam, The Hague and Utrecht as separate categories. Ethnic background has seven categories: native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western non-native and other western non-native. The classification is based on the country of birth of the parents of the selected person. The variable age has three categories: 0 – 34 years, 35 – 54 years, and 55 years and older.

In figure 3.4.1 R_1, R_2 and R_3 are plotted against the response rate for the first six contact attempts in POLS. The leftmost values correspond to the respondent pool after one attempt was made. For each additional attempt the response rate increases, but all three indicators show a drop in representativity. This result confirms findings in Schouten (2004). The paths of R-indexes R_2 and R_3 are close to each other, but they cross each other after the first contact attempt.

Figures 3.4.2 - 3.4.5 depict the marginal R-indexes MR_1 and MR_2 for the variables age and ethnic background. Marginal R-index MR_1 does not account for cross-effects between variables, marginal R-index MR_2 does account for these effects. Again the marginal R-indexes are plotted for the first six contact attempts.

Figures 3.4.2 and 3.4.3 both show that at the first attempt persons of 55 years and older are overrepresented, while the other two age groups are underrepresented. This is not a surprising result as elderly people are easier to contact. From the first to the fourth attempt both marginal R-indexes become gradually smaller and they almost disappear at the fourth attempt. The last two attempts show growing marginal R-indexes.

Figure 3.4.2: Marginal R-index MR_1 for age in three categories (<34, 35-54, >54) and for the first six contact attempts

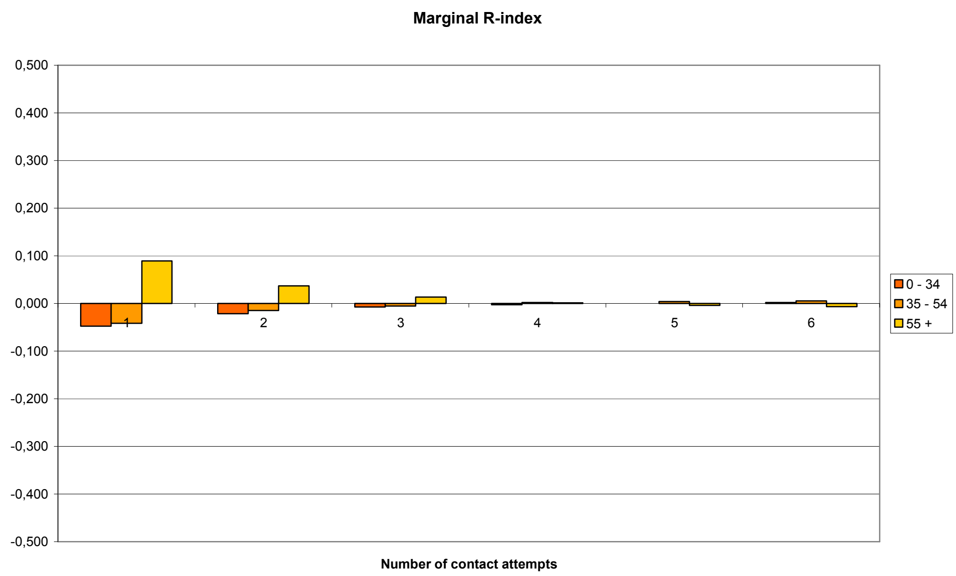


Figure 3.4.3: Marginal R-index MR_2 for age in three categories (<34, 35-54, >54) and for the first six contact attempts.

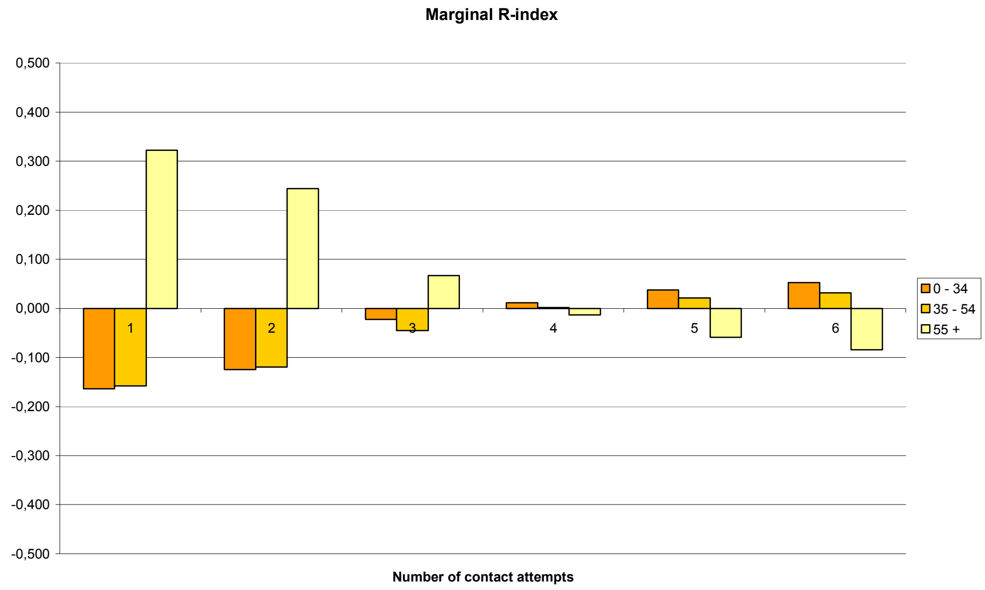


Figure 3.4.4: Marginal R-index MR_1 for ethnic background (native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western, other western) and for the first six contact attempts.

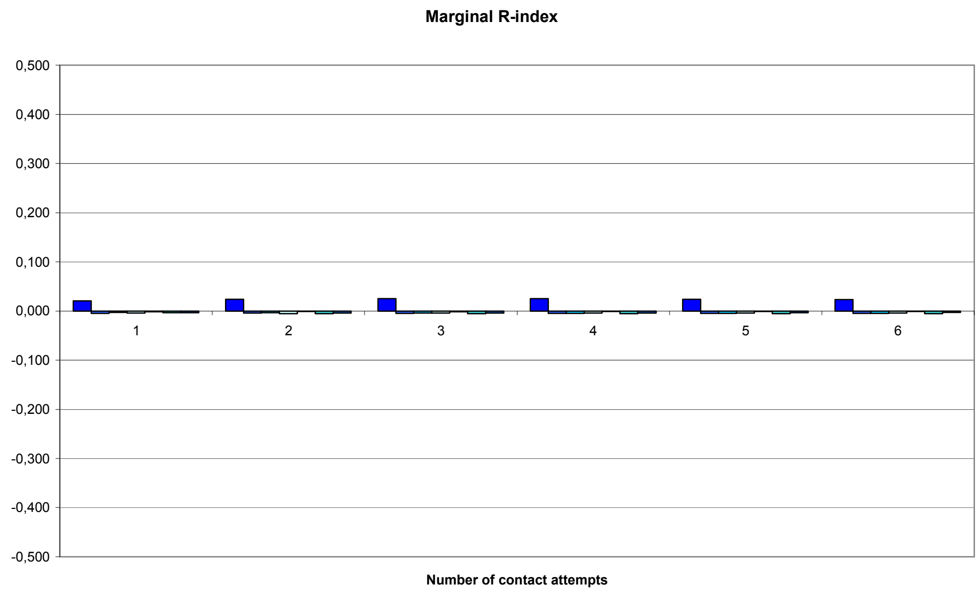
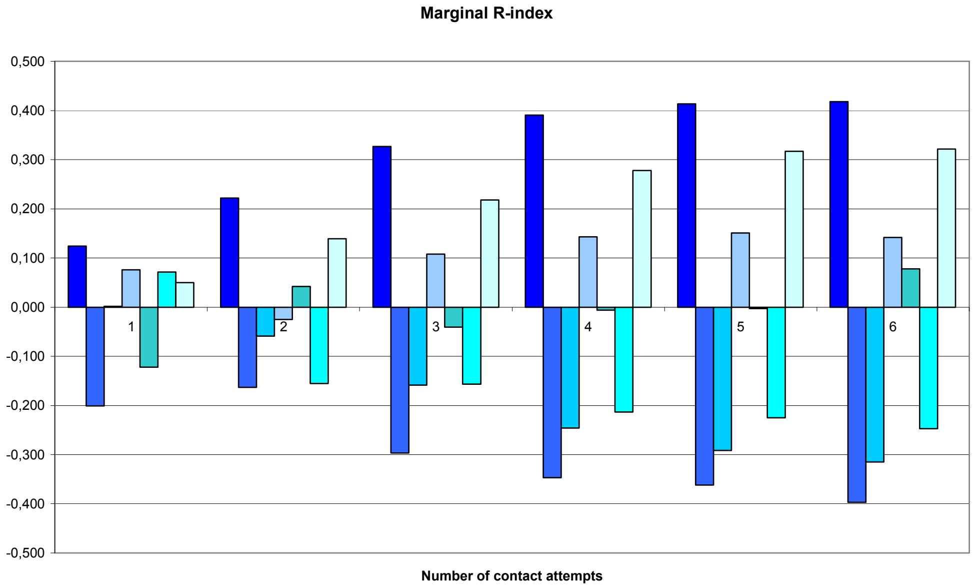


Figure 3.4.5: Marginal R-index MR_2 for ethnic background (native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western, other western) and for the first six contact attempts.



The picture for ethnic background is different. Marginal R-index MR_1 does not show large changes when the number of contact attempts is increased and is rather constant. However, marginal R-index MR_2 leads to the conclusion that, when corrected for cross-effects, the marginal R-indexes become more diffuse as the number of contact attempts grows. Also both marginal R-indexes give different views to persons that are over or underrepresented. MR_1 indicates that all non-native groups are underrepresented, while the MR_2 shows that only Moroccans, Turkish and other non-western non-natives are underrepresented. These observations can be explained by the correction for cross-effects and by the relative small sizes of all non-native ethnic groups. As we are interested in the marginal impact of auxiliary variables, we prefer the use of MR_2.

4. Features of R-indexes

In the previous section we identified a number of candidate indicators for representativity either overall or conditional on other variables. As literature shows there is a vast number of association measures, e.g. Goodman and Kruskal (1979). Association measures have a strong relation to R-indexes. Essentially, R-indexes attempt to measure in a multivariate setting the lack of association. In this section we discuss the desired features of R-indexes.

4.1 Features in general

We want R-indexes to be based on a distance function or metric in the mathematical sense. The triangle inequality property of a distance function allows for a partial ordering of the variation in response propensities which enables interpretation. A distance function can easily be derived from any mathematical norm. In section 2 we chose to use the Euclidean norm as this norm is commonly used. The Euclidean norm led us to an R-index that uses the standard deviation of response propensities. Other norms like the supremum norm would lead us to alternative distance functions. In section 4.3 we will show, however, that the Euclidean norm based R-indexes have interesting normalization features.

The third R-index that we proposed in section 2 employs the likelihood functions of regression models and as far as we can see it is not directly related to a mathematical distance function for response propensities themselves. Nonetheless, it is interesting as it is based on an indicator for the proportion of error reduction.

We must also make a subtle distinction between R-indexes and distance functions. Distance functions are symmetric while an R-index measures a deviation with respect to a specific point, namely the situation where all response propensities are equal. If we change the vector of individual propensities, then this point is in most cases shifted. However, if we fix the average response propensity then the distance function facilitates interpretation.

Apart from a relation to a distance function, we want to be able to measure, interpret and normalize the R-indexes. In section 2 we already derived estimators for ‘population’ R-indexes that are not measurable in case response propensities are unknown and all we have is the response to a survey. Hence, we made R-indexes measurable by switching to estimators. The other two features are discussed separately in the next two sections.

4.2 Interpretation

The second feature of R-indexes is the ease with which we can interpret their values and the concept they are measuring. We moved to estimators for R-indexes that are based on the samples of surveys and on estimators of individual response probabilities. Both have far-reaching consequences for the interpretation and comparison of R-indexes.

Since the R-indexes are estimators themselves, they are also random variables. This means that they depend on the sample, i.e. they are potentially biased and have a certain accuracy. But what are they estimating?

Let us, first, assume that the sample size is arbitrarily large so that accuracy does not play a role and that the selection of a model for response propensities is no issue. In other words, we are able to fit any model for any fixed set of auxiliary variables.

There is a strong relation between the R-indexes and the availability and use of auxiliary variables. In section 2 we defined strong and weak representativity. Even in the case where we are able to fit any model, we are not able to estimate response propensities beyond the 'resolution' of the available auxiliary variables. Hence, we can only draw conclusions about weak representativity with respect to the set of auxiliary variables. This implies that whenever an R-index is used, it is necessary to complement its value by the set of covariates that served as a grid to estimate individual response propensities. If R-indexes are used for comparative purposes, then those sets must be the same. We must add that it is not necessary that all auxiliary variables are indeed used for the estimation of propensities, since they may not add any explanatory power to the model. However, the same sets should be available. The R-indexes then measure a deviation from weak representativity.

Clearly, in practice the sample size is not arbitrarily large. The sample size affects estimation steps, the estimation of response propensities and the estimation of the R-index using a sample. However, the consequences are only serious for the estimation of the response propensities.

If we would know the individual response propensities, then the sample-based estimation of R-indexes would only lead to variance and not to bias. We would be able to estimate population R-indexes without bias. Hence, for small sample sizes the estimators would have a small accuracy but this could be accounted for by using confidence intervals instead of merely point estimators.

The implications for the estimation of response probabilities are, however, different because of model selection and model fit. There are two alternatives. Either, one imposes a model to estimate propensities fixing the covariates beforehand, or one lets the model be dependent on the significant contribution of covariates with respect to some predefined level. In the first case, again no bias is introduced but the standard error may be considerable because of over fitting. In the second case, the model for the estimation of response propensities depends on the size of the sample; the larger the sample, the more interactions that are accepted as significant. Although it is standard statistical practice to fit models based on a significance level, model selection may introduce bias and variance to the estimation of R-indexes. This can be easily understood by going to the extreme of a sample of say size 10. For such a small sample no interaction between response behaviour and auxiliary characteristics will be accepted, leaving an empty model and an estimated R-index of 1. Small samples simply do not allow for the estimation of response propensities. In general, a smaller sample size will, thus, lead to a more optimistic view on representativity.

We should make a further subtle distinction. It is possible that for one survey a lot of interactions contribute to the prediction of response propensities but only very little each, while in another survey there is only one but strong contribution of a single interaction. None of the small contributions may be significant, but together they are as strong as the one large contribution that is significant. Hence, we would be more optimistic in the first example even if sample sizes would be comparable.

The potential bias of R-indexes puts the use of these indicators for comparison of surveys under pressure. This is especially true, whenever different surveys have very different sample sizes.

The bias and variance of R-indexes are topics of future research. We need to be able to compute confidence intervals for the R-indexes and we need to be able to account for sample size. Here, we identify two approaches to deal with the bias of R-indexes. In the first approach we do fit models incorporating only significant interactions between response and auxiliary variables. In case a single survey is evaluated, it may be key to fit models to response propensities and publish the R-index together with the set of covariates and the sample size. Empirical validation should lead to knowledge about reasonable values of the R-indexes for various sample sizes. In case a number of surveys is compared directly, one may subsample surveys in order to get similar sample sizes and then fit models to all subsamples. In the second approach one chooses a stratification beforehand and uses this stratification to estimate response propensities regardless of the sample size.

4.3 Normalization

The third important feature is normalizability. We want to be able to attach bounds to an R-index so that the scale of an R-index, and, hence, changes in the R-index get a meaning. Clearly, the interpretation issues that we raised in the previous section also strongly affect the normalization of the R-index. Therefore, in this section we assume the ideal situation that we can estimate response propensities without bias. This assumption holds for very large surveys. We discuss the normalization of R-indexes $\hat{R}_1$ and $\hat{R}_2$. Normalization of $\hat{R}_3$ may be a topic of future research.

Let Y be some variable that is measured in a survey and let $\hat{\bar{y}}$ be the Horvitz-Thompson estimator for the population mean based on the survey response, i.e.

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{N} y_i \frac{r_i}{\pi_i}}{\sum_{i=1}^{N} \frac{r_i}{\pi_i}}. \tag{21}$$

From the literature it is known that the non-response bias of (21) is equal to

$$B(\hat{\bar{y}}) = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})(y_i - \bar{y})}{\bar{\rho}}, \tag{22}$$

the quotient of the covariance between the response propensities and Y, and the average response propensity.

Now, suppose, hypothetically, that in the survey we measure the individual response propensities ρ_i. From (22) it easily follows that the bias of the Horvitz-Thompson for the average population response propensity $\bar{\rho}$ equals the variance of the response propensities divided by that same average, i.e.

$$B(\hat{\bar{\rho}}) = \frac{S^2(\tilde{\rho})}{\bar{\rho}}. \tag{23}$$

This result gives a direct opportunity to normalize R-indexes $\hat{R}_1$ and $\hat{R}_2$ by bounding the bias of the average response propensity in case we would observe it. If $\hat{R}_1 = \alpha$, then $B(\hat{\bar{\rho}}) = \frac{(1-\alpha)^2}{4\bar{\rho}}$. If $\hat{R}_2 = \beta$, then $B(\hat{\bar{\rho}}) = \frac{1-\beta}{4\bar{\rho}}$.

As a consequence, if we choose an upper bound for the bias of the response propensity itself, then we get lower bounds for the R-indexes. The lower bound depends on the size of the average response propensity; the higher the response rate, the lower the lower bounds for the R-indexes. Instead, and somewhat analogous to *p*-values in statistical tests, one may compute the response propensity bias for the comparison of different surveys.

An alternative normalization is found by the inequality of Cauchy-Schwarz. This inequality states that the covariance between any two variables is bounded in absolute sense by the product of the standard deviations of the two variables. We can translate this to bounds for the bias (22) of an arbitrary survey variable Y

$$|B(\hat{\bar{y}})| \leq \frac{S(\tilde{\rho})S(y)}{\bar{\rho}}. \tag{24}$$

If the variable of interest is the proportion of units falling in some category z, i.e. $Y = \delta_{z,i}$ is the 0-1-indicator function for that category, then $S(y) \leq 1/2$ and

$$|B(\hat{\bar{y}})| \leq \frac{S(\tilde{\rho})}{2\bar{\rho}}. \tag{25}$$

If $\hat{R}_1 = \alpha$, then $|B(\hat{\bar{y}})| \leq \frac{1-\alpha}{4\bar{\rho}}$. If $\hat{R}_2 = \beta$, then $|B(\hat{\bar{y}})| \leq \frac{\sqrt{1-\beta}}{2\bar{\rho}}$.

Conversely, if we demand that

$$\frac{S(\tilde{\rho})}{2\bar{\rho}} \leq \gamma, \tag{26}$$

then (26) implies that

$$\hat{R}_1 \geq 1 - 4\bar{\rho}\gamma, \tag{27}$$

and

$$\hat{R}_2 \geq 1 - 16\bar{\rho}^2\gamma^2. \tag{28}$$

From (27) and (28) we can see that again the lower bounds go up in case the response rate also goes up.

Figure 4.3.1 contains lower bounds (27) and the observed R-indexes $\hat{R}_1$ for the example of section 3.4. Two values of γ are chosen, $\gamma = 0{,}1$ and $\gamma = 0{,}05$. Figure 6 indicates that after the second contact attempt, the values of the R-index exceed the lower bound corresponding to $\gamma = 0{,}1$. However, the values never exceed the other lower bound that is based on $\gamma = 0{,}05$.

In figure 4.3.2 the maximal absolute bias is derived from the observed R-indexes. After the third contact attempt the R-index converges to a value around 0.08. In other words, the maximal absolute bias of an estimated proportion cannot exceed 8%.

Figure 4.3.1: Lower bounds for R-index $\hat{R}_1$ and observed values for the first six contact attempts of POLS 1998. Lower bounds are based on $\gamma = 0{,}1$ and $\gamma = 0{,}05$.

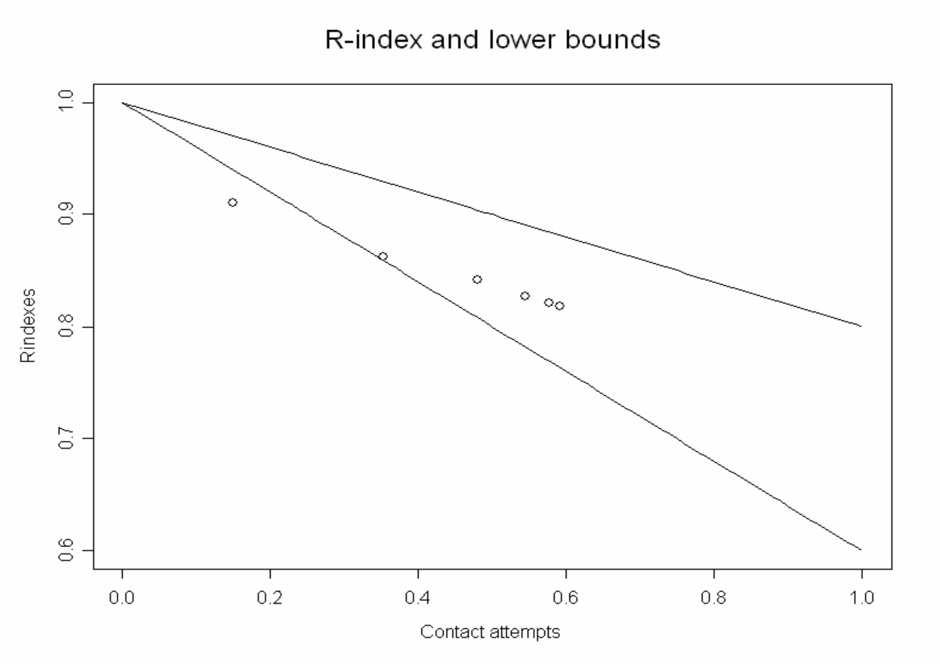
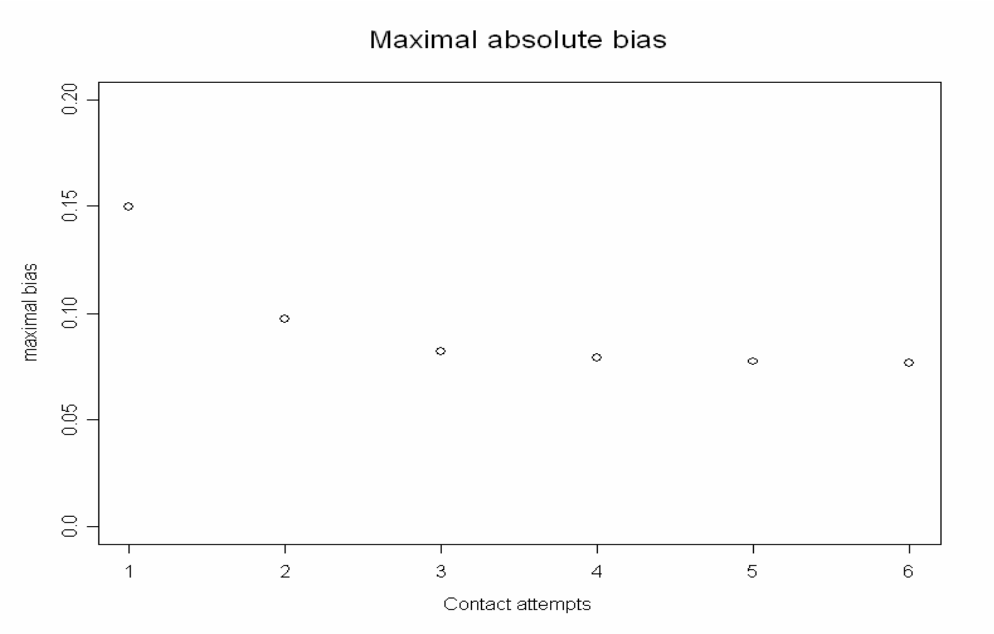


Figure 4.3.2: Maximal absolute bias of an estimated proportion for the first six contact attempts of POLS 1998.



It is important to remark that the proposed R-indexes are directed at measuring variation in response propensities and not at measuring non-response bias. However, the proposed R-indexes lead to general bounds for the non-response bias. Although these bounds are very conservative, they offer an opportunity for the normalization of the R-indexes.

5. Discussion and future research

We had two main objectives in this paper: the construction of potential indicators for representativity and the identification of research issues for the near future. We proposed a number of R-indexes and marginal R-indexes and illustrated the indicators by application to a real data set. Furthermore, we discussed the features of R-indexes and their implications for the comparison of surveys. These implications are the main input for future research into R-indexes.

R-indexes estimate the dissimilarity between the respondent and sample pool with respect to auxiliary variables that are available from other sources than the survey itself. We call a response to a survey representative with respect to those variables if we cannot distinguish the composition of the response from that of the survey in a statistical sense. In other words we investigate whether the response mechanism is similar to a simple random sample without replacement taken out of the survey sample. The R-indexes that we propose measure the deviation from this ideal situation.

The R-indexes in this paper are promising because they can easily be computed and allow for interpretation and normalization in case response propensities can be estimated without error. The application to real survey data shows that the R-indexes confirm earlier analyses of the non-response composition. Other R-indexes can simply be constructed by choosing different distance functions between vectors of response propensities.

The R-indexes and graphical displays that we showed in this paper can be computed using most standard statistical software packages. However, for the transformation of regression parameters to centralized regression parameters, an additional macro must be written. Hence, software packages that allow for easy programming of such syntax are to be favoured.

The computation of R-indexes is sample-based and employs models for individual response propensities. Hence, R-indexes are random variables themselves and there are two estimation steps that influence their bias and variance. However, it is mostly the modelling of response propensities that has important implications. The restriction to the sample for the estimation of R-indexes, implies that those indicators are less accurate but this restriction does not introduce a bias. Model selection and model fit usually are performed by choosing a significance level and adding only those interactions to the model that give a significant contribution. The latter means that the size of the sample plays an important role in the estimation of response propensities. Bias may be introduced by the model selection strategy.

There are various obvious approaches to deal with the dependency on the size of the sample. One may restrain from model selection and fix a stratification beforehand. That way bias is avoided, but standard errors are not controlled and may be considerable. One may also let empirical validation be the input to develop 'best practices' for R-indexes.

We identify the following areas for future research:

- Search for other promising R-indexes
- Empirical validation of the proposed (marginal) R-indexes
- Estimators for the standard errors and confidence intervals of (marginal) R-indexes
- Interpretation and normalization of (marginal) R-indexes relative to sample size

References

Agresti, A. (2002), Categorical data analysis, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA.

Bertino, S. (2006), A measure of representativeness of a sample for inferential purposes, International Statistical Review, 74, 149 – 159.

Biemer, P.P., Link, M.W. (2006), A latent call-back model for nonresponse, Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

Biemer, P.P., Lyberg, L.E. (2003) Introduction to survey quality, John Wiley & Sons, New York, NY, USA.

Cobben, F., 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.

Curtin, R., Presser, S., Singer, E. (2000), The effects of response rate changes on the index of consumer sentiment, Public Opinion Quarterly 64, 413-428.

Fitzgerald, R., Fuller, L. (1982), I hear you knocking but you can't come in: the effects of reluctant respondents and refusers on sample survey estimates, Sociological Methods and Research, 11, 3 – 32.

Goodman, L.A., Kruskal, W.H. (1979), Measures of association for cross-classifications, Springer-Verlag, Berlijn, Duitsland.

Grijn, F. van der, Schouten, B., Cobben, F. (2006), Balancing representativity, costs and response rates in a call scheduling strategy, Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

Groves, R.M. (1989), Survey errors and survey costs, John Wiley & Sons, New York, NY, USA.

Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (2002), Survey nonresponse, Wiley Series in Probability and Statistics, Wiley: New York, NY, USA.

Groves, R.M., Heeringa, S.G. (2005), Responsive design for household surveys: tools for actively controlling survey nonresponse and costs, Technical paper, University of Michigan and Joint Program in Survey Methodology, USA.

Groves, R. M., Presser, S., Dipko, S. (2004), The role of topic interest in survey participation decisions, Public Opinion Quarterly 68, 2-31.

Hajék, J. (1981), Sampling from finite populations, Marcel Dekker, New York, USA.

Heerwegh, D., Loosveldt, G. (2006), A tentative assessment of marginal R-indices from the perspective of nonresponse error analysis, Technical paper, University of Leuven, Belgium.

Kalsbeek, W.D., Botman, S.L., Massey, J.T., Liu, P.W. (1994), Cost-efficiency and the number of allowable attempts in the National Health Interview Survey, Journal of Official Statistics, 10, 133 – 152.

Kass, G.V. (1980), An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 29 no. 2, 119-127.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., Presser, S. (2000), Consequences of reducing nonresponse in a national telephone survey, Public Opinion Quarterly 64, 125-148.

Kruskal, W., Mosteller, F. (1979a), Representative sampling I: non-scientific literature, International Statistical Review, 47, 13-24.

Kruskal, W., Mosteller, F. (1979b), Representative sampling II: scientific literature excluding statistics, International Statistical Review, 47, 111-123.

Kruskal, W., Mosteller, F. (1979c), Representative sampling III: current statistical literature, International Statistical Review, 47, 245-265.

Lin, I.F., Schaeffer, N.C. (1995), Using survey participants to estimate the impact of nonparticipation, Public Opinion Quarterly, 59, 236 – 258.

Little, R.J.A., Rubin, D.B. (2002), Statistical analysis with missing data, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA.

Lynn, P., Clarke, P., Martin, J., Sturgis, P. (2002), The effects of extended interviewer efforts on nonresponse bias, In *Survey Nonresponse (eds R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little)*, John Wiley & Sons, New York, USA, 135 – 148.

Merkle, D.M., Edelman, M. (2002), Nonresponse in exit polls: a comprehensive analysis, In *Survey Nonresponse (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R..J.A.),* John Wiley & Sons, 243-258.

Schouten, B. (2004), Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998, Discussion paper 04001, CBS, Voorburg.

Schouten, B., Cobben, F. (2006), R-indexes for the comparison of different fieldwork strategies and data collection modes, Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

Stoop, I.A.L. (2005), The Hunt for the Last Respondent, PhD thesis, Faculty of Social Sciences, Utrecht University, The Netherlands.

Appendix: Centralized parameters in logistic regression

Suppose that in the logistic regression there are K categorical explanatory variables. The number of categories of variable k we denote by m_k. We include an intercept α in the model and use the last category of each variable as the category of reference. Let β_{kl} be the parameter for category l of variable k. Hence, $\beta_{km_k} = 0$ by definition as the last category is the reference category.

The logistic regression will produce estimates

$$\hat{\alpha}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{1,m_1-1}, 0, \ldots, \hat{\beta}_{K1}, \ldots, \hat{\beta}_{K,m_K-1}, 0$$

(with a zero for the last category). We want to transform these estimates in such a way that

$$\sum_{l=1}^{m_k} \hat{\beta}_{kl} = 0, \forall k,$$

and that the log-odds remain the same for all $\prod_{k=1}^{K} m_k$ cells.

This can be achieved by the transformation

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_{11} \\ \vdots \\ \hat{\beta}_{1,m_1-1} \\ 0 \\ \vdots \\ \hat{\beta}_{K1} \\ \vdots \\ \hat{\beta}_{K,m_K-1} \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \hat{\alpha} + \bar{\hat{\beta}}_1 + \ldots + \bar{\hat{\beta}}_K \\ \hat{\beta}_{11} - \bar{\hat{\beta}}_1 \\ \vdots \\ \hat{\beta}_{1,m_1-1} - \bar{\hat{\beta}}_1 \\ -\bar{\hat{\beta}}_1 \\ \vdots \\ \hat{\beta}_{K1} - \bar{\hat{\beta}}_K \\ \vdots \\ \hat{\beta}_{K,m_K-1} - \bar{\hat{\beta}}_K \\ -\bar{\hat{\beta}}_K \end{pmatrix},$$

with $\bar{\hat{\beta}}_k = \frac{1}{m_k} \sum_{l=1}^{m_k} \hat{\beta}_l$ the average over the estimated parameters for variable k.

The above transformation can easily be written in matrix form, say $T\theta$ with $\theta = (\hat{\alpha}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{1,m_1-1}, 0, \ldots, \hat{\beta}_{K1}, \ldots, \hat{\beta}_{K,m_K-1}, 0)'$. The estimated covariance matrix $\hat{\Sigma}$ needs to be transformed as well in case standard errors are needed. This can be done by $T\hat{\Sigma}T'$.