

Working with Response Probabilities

Jelke Bethlehem¹

Sample surveys are often affected by nonresponse. These surveys have in common that their outcomes depend at least partly on a human decision whether or not to participate. If it would be completely clear how this decision mechanism works, estimates could be corrected. An often used approach is to introduce the concept of the response probability. Of course, these probabilities are a theoretical concept and therefore unknown. The idea is to estimate them by using the available data. If it is possible to obtain good estimates of the response probabilities, they can be used to improve estimators of population characteristics.

Estimating response probabilities relies heavily on the use of models. An often used model is the logit model. In the article, this model is compared with the simple linear model.

Estimation of response probabilities models requires the individual values of the auxiliary variables to be available for both the respondents and the nonrespondents of the survey. Unfortunately, this is often not the case. This article explores some approaches for estimating response probabilities that have less heavy data requirements. The estimated response probabilities were also used to measure possible deviations from representativity of the survey response. The indicator used is the coefficient of variation (CV) of the response probabilities.

Key words: Nonresponse; adjustment weighting; response propensity; representativity.

1. Introduction

There are various ways of selecting a sample for a survey, but over the years it has become clear that the scientifically sound way to do this is by means of selecting a *probability sample*. Objects (persons, households, businesses) must have a non-zero probability of selection, and all these selection probabilities must be known. This makes it possible to compute precise, unbiased estimates of population characteristics. Also, the precision of these estimates can be quantified, for example by means of a *confidence interval*, or a *margin of error*. These are the fundamental principles of survey sampling.

In practice, the situation is often not so ideal. All kinds of problems may affect the quality of the estimates. One of those problems is (unit) *nonresponse*. This means that no information is obtained about a number of objects in the sample. The questionnaire form remains empty for these objects. One of the effects of nonresponse is that the sample size is smaller than expected. This leads to less precise, but still valid, estimates of population characteristics. This is not a serious problem as it can be taken care of by increasing the initial sample size. A far more serious effect of nonresponse is that estimates of population characteristics may be *biased*. This occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with

¹ Leiden University, Institute of Political Science, Albert Verweystraat 21, 2394 TK Hazerswoude-Rijndijk, The Netherlands. Email: jelkeb@xs4all.nl.

respect to the characteristics being investigated. Consequently, wrong conclusions will be drawn from the survey data. Such a situation must be avoided as much as possible. Therefore, the amount of nonresponse must be kept small as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of nonresponse usually remains. There are several books that treat the nonresponse problem in more detail. A general overview is given by [Bethlehem et al. \(2011\)](#). [Stoop \(2005\)](#) shows when nonresponse can cause bias and investigates causes of and reasons of nonresponse. [Särndal and Lundström \(2005\)](#) focus on estimation techniques that improve the accuracy of survey estimates. A more recent reference is [Valliant et al. \(2018\)](#). The goal of this book is to present a set of tools for handling nonresponse. They also show how existing software can be used to solve survey problems.

Although probability sampling is the preferred way of sample selection, some researchers use different selection techniques. Particularly for online surveys, *self-selection sampling* is used. The questionnaire is made available on the internet. Respondents are those visitors of the website who spontaneously decide to participate in the survey. No random sampling is applied. Respondents are those who happen to know the survey is being conducted, happen to have internet access, decide to visit the survey website, and complete the questionnaire. As the selection mechanism of these online surveys is completely unknown and unclear, it is impossible to compute precise and valid estimates of population characteristics. For more on self-selection surveys, see, for example [Bethlehem and Biffignandi \(2012\)](#).

Both probability sampling (affected by nonresponse) and self-selection sampling have in common that their outcomes depend, at least partly, on a human decision whether or not to participate. If it would be completely clear how this decision mechanism works, the estimates could be corrected. Unfortunately, such information is not available. An often used approach to analyse the effects of and to correct for biased human participation decisions is to introduce the concept of the *response probability*. It is assumed that every object in the target population of the survey has a certain probability to respond in the survey if asked to do so. Of course, these probabilities are a theoretical concept and therefore unknown. The idea is now to estimate the response probabilities using the available data. If it is possible to obtain good estimates of the response probabilities, they can be used to improve estimators of population characteristics.

Estimated response probabilities can be used by survey researchers in several ways. The focus in this article is on:

- *Analysis of nonresponse.* By analysing the relationships between response probabilities and other survey variables, insight is obtained in how the nonresponse mechanism works.
- *Correction for nonresponse.* Once precise estimates of response probabilities are available, they can be used in weighting adjustment techniques that reduce the bias caused by nonresponse.
- *Representativity indicator.* The response probabilities can be used to construct a representativity indicator, which shows, in one number, how good or bad a survey is.

Estimating response probabilities relies heavily on the use of models. An often used model is the *logit model*. It attempts to predict the response probabilities by using a set of

auxiliary variables. This seems to work well in practical survey situations. Other models are the *probit model* and the *linear model*. These models are compared.

Estimation of these response probabilities models requires the individual values of the auxiliary variables to be available for both the respondents and the nonrespondents in the survey. Unfortunately, this is often not the case. This article explores some approaches for estimating response probabilities that have less heavy data requirements. The idea is to start by computing weights with some weighting adjustment technique. These weights can be seen as a kind of inverted response probabilities, and therefore they can be used to estimate response probabilities. Weighting techniques have more modest data requirements. They can compute weights without having the individual data of the nonrespondents. Two weighting techniques are considered: generalised regression estimation and raking ratio estimation.

By taking the logit model as a benchmark, it is explored whether approximately the same estimated response probabilities can be obtained using techniques requiring less information:

1. The linear model for response probabilities;
2. Transforming weights that have been obtained by generalised regression estimation into estimated response probabilities;
3. Transforming weights that have been obtained by raking ratio estimation into estimated response probabilities.

The various approaches are tested on a real survey data set of Statistics Netherlands. This is an anonymised data set that will be called here the General Population Survey (GPS). The sample for this survey was selected from the population register of the Netherlands. Therefore, auxiliary variables in the register are available for both respondents and nonrespondents. Moreover, the sample data file was linked to some registers, providing even more auxiliary variables. So logit models could be fitted, and they could be compared with approaches requiring less data.

The estimated response probabilities were used to measure possible deviations from representativity of the survey response. The indicator used is the *coefficient of variation* (CV) of the response probabilities. This CV can be seen as a normalised measure of dispersion of the response probabilities. The larger the value of the CV, the more the response probabilities will vary, and the more the survey response therefore will lack representativity. The CV also turns up as a component of the bias of estimators that are affected by nonresponse.

The weighting adjustment approach makes it possible to estimate response probabilities in situations in which the logit model cannot be used. An example is given. Response probabilities and the CV are computed for a self-selection panel. This is the 'EenVandaag Opiniepanel' of the Dutch national public television channel 'NOPI'.

2. The Concept of Response Probability

2.1. Nonresponse in a Simple Random Sample

Let the finite *survey population* U consist of a set of N identifiable objects that are labelled $1, 2, \dots, N$. Associated with each object k is an unknown value Y_k of the *target variable*.

The vector of all values of the target variable is denoted by

$$Y = (Y_1, Y_2, \dots, Y_N)'. \quad (1)$$

The symbol ' denotes transposition of a matrix or vector. The objective of the sample survey is assumed to be the estimation of the population mean

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k. \quad (2)$$

To estimate this population characteristic, a simple random sample of size n is selected without replacement. The sample can be represented by the N -vector

$$a = (a_1, a_2, \dots, a_N)' \quad (3)$$

of indicators, where $a_k = 1$ if object k is selected in the sample, and otherwise $a_k = 0$.

In case of simple random sampling without replacement, the sample mean

$$\bar{y} = \frac{1}{n} \sum_{k=1}^N a_k Y_k \quad (4)$$

is an unbiased estimator of the population mean.

Now suppose there is unit nonresponse in the survey. It is assumed that each object k in the population has a certain, unknown probability ρ_k of response. If object k is selected in the sample, a random mechanism is activated that results with probability ρ_k in response and with probability $1 - \rho_k$ in nonresponse. A vector R of response indicators

$$R = (R_1, R_2, \dots, R_N)' \quad (5)$$

is introduced, where $R_k = 1$ if the corresponding objects k responds, and where $R_k = 0$ otherwise. So, $P(R_k = 1) = \rho_k$, and $P(R_k = 0) = 1 - \rho_k$.

The survey response only consists of those elements k for which $a_k = 1$ (in the sample) and $R_k = 1$ (responds). Hence, the number of respondents is equal to

$$n_R = \sum_{k=1}^N a_k R_k. \quad (6)$$

Likewise, the number of nonrespondents is $n_{NR} = n - n_R$.

The values of the target variable only become available for the n_R responding objects. The mean of these values can be denoted by

$$\bar{y}_R = \frac{1}{n_R} \sum_{k=1}^N a_k R_k Y_k. \quad (7)$$

Bethlehem (2009) shows that the expected value of the response mean is approximately equal to

$$E(\bar{y}_R) \approx \frac{1}{N} \sum_{k=1}^N \frac{\rho_k}{\bar{\rho}} Y_k, \quad (8)$$

where

$$\bar{\rho} = \frac{1}{N} \sum_{k=1}^N \rho_k \quad (9)$$

is the mean of all response probabilities in the population. Expression (8) shows that, generally, the expected value of the response mean is unequal to the population mean to be estimated. Therefore, this estimator is biased. This bias is approximately equal to

$$B(\bar{y}_R) = E(\bar{y}_R) - \bar{Y} \approx \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (10)$$

where $R_{\rho Y}$ is the correlation between the response probabilities ρ and a target variable Y of the survey, S_{ρ} is the standard deviation of the response probabilities ρ , and S_Y is the standard deviation of the target variable Y . From Equation (10) a number of conclusions can be drawn:

- The bias vanishes if there is no relationship between the target variable of the survey and the response behaviour. Then $R_{\rho Y} = 0$. The stronger the relationship between the target variable and response behaviour, the larger the bias will be.
- The bias vanishes if all response probabilities are equal. Then $S_{\rho} = 0$. Indeed, in this situation the nonresponse is not selective. It just reduces the sample size. The more the values of the response probabilities vary, the larger the bias will be.
- The magnitude of the bias increases as the mean of the response probabilities decreases. The response rate is an unbiased estimator of the mean response probability. Translated in practical terms, this means that low response rates will lead to large biases.

It is clear that analysis of the estimates of the response probabilities provides insight into the possible effects of nonresponse on the possible bias of estimates of population characteristics. Some authors (for example [Groves 2006](#)) discuss a possible relationship between the response rate of a survey and the bias of its estimates. They fear increased biases if response rates decline. Equation (10) shows that the magnitude of the bias is determined by more than just the response rate. Just as important are the variation of the response probabilities (S_{ρ}) and the correlation between the response probabilities and the target variable ($R_{\rho Y}$).

2.2. Nonresponse in a Self-Selection Sample

Self-selection means that researchers are not in control of the sample selection process. They just make the survey questionnaire available, and wait and see what happens. A typical example is a web survey where everyone can complete the questionnaire on the internet. Also people outside the target population of the survey can participate. It is sometimes even possible to fill in the questionnaire more than once.

Participation in a self-selection web survey requires that respondents are aware of the existence of the survey. Moreover, they must have access to the internet, they have to visit the website (for example by following up a banner, an e-mail message, or a commercial on radio or TV), and they have to decide to fill in the questionnaire. This means that each

object k in the population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, \dots, N$.

Assuming there are no under-coverage problems, everyone has a nonzero probability of participating in the survey. The survey response is denoted by the vector of indicators

$$R = (R_1, R_2, \dots, R_N)', \quad (11)$$

where $R_k = 1$ if object k participates, and otherwise $R_k = 0$, for $k = 1, 2, \dots, N$. The expected value $\rho_k = E(R_k)$ is the *response probability* of element k . The realised sample size is denoted by

$$n_S = \sum_{k=1}^N R_k \quad (12)$$

Lacking any knowledge about the values of the response probabilities, a naïve researcher would implicitly assume all these probabilities to be equal. In other words: simple random sampling is assumed. Consequently, the sample mean

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N R_k Y_k \quad (13)$$

is used as an estimator for the population mean. [Bethlehem \(2009\)](#) shows that the expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \frac{1}{N\bar{\rho}} \sum_{k=1}^N \rho_k Y_k \quad (14)$$

where $\bar{\rho}$ is the mean of all response probabilities.

It is clear from Equation (14) that, generally, the expected value of this sample mean is not equal to the population mean. One situation in which the bias vanishes is that in which all response probabilities are equal. In terms of the theory of missing data, this comes down to Missing Completely At Random (MCAR). This is the situation in which the cause of missing data is completely independent of all variables measured in the survey. For more information on MCAR and other missing data mechanisms, see [Little and Rubin \(2002\)](#). Indeed, in the case of MCAR, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability.

[Bethlehem \(2009\)](#) shows that the bias of the sample mean in Equation (13) is approximately equal to

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y} \approx \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (15)$$

in which $R_{\rho Y}$ is the correlation between the values of target variable Y and the response probabilities ρ , S_{ρ} is the standard deviation of the response probabilities, S_Y is the standard deviation of the target variable, and $\bar{\rho}$ is the average response probability.

Equation (10) for the bias in a random sample affected by nonresponse is identical to Equation (15) for the bias in a self-selection survey. However, in practical situations their values will be substantially different. For example, the probability samples for surveys of Statistics Netherlands had response rates of around 60%. This means that the average

response probability was 0.6. There have been self-selection web surveys in the Netherlands with large samples. An example is *21minuten.nl*. Approximately 170,000 people completed the questionnaire in 2006. Assuming the target population to consist of all Dutch citizens from the age of 18, the average response probability was $170,000 / 12,800,000 = 0.0133$. This is a much lower value than the 0.6 of probability sampling based surveys. So there is a risk of a much large bias in self-selection surveys.

From Equations (10) or (15) an upper bound for the bias can be computed. Given the mean response probability $\bar{\rho}$, there is a maximum value that the standard deviation S_{ρ} of the response probabilities cannot exceed:

$$S_{\rho} \leq \sqrt{\bar{\rho}(1 - \bar{\rho})}. \quad (16)$$

This implies that in the worst case, S_{ρ} assumes its maximum value if the correlation coefficient $R_{\rho Y}$ is equal to either $+1$ or -1 . Then the absolute value of the bias will be

$$|B_{max}| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}. \quad (17)$$

In case of a survey based on probability sampling with a response rate of around 60%, the maximum absolute bias is equal to $0.816 \times S_Y$. In case of a self-selection survey a size 170,000 from a population of size 12,800,000, the maximum absolute bias is $8.619 \times S_Y$. This is more than ten times as large.

3. Estimating Response Probabilities

3.1. Models for Response Probabilities

Response probabilities are unknown. Therefore they must be estimated using the available data. To this end, the concept of the *response propensity* is introduced. Following [Little \(1986\)](#) and [Bethlehem et al. \(2011\)](#), the response propensity of object k is defined by

$$\rho_k(X_k) = P(R_k = 1 | X_k) \quad (18)$$

where R_k is the response indicator, and $X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$ is a vector of values of p auxiliary variables. So the response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, but they can be estimated provided the values of the auxiliary variables are available for both the respondents and nonrespondents. The estimated response propensity is denoted by $\hat{\rho}_k(X_k)$. If the set of auxiliary variables is sufficient to explain the response probabilities, the (estimated) response propensities will resemble the response probabilities.

To be able to estimate the response propensities, a model must be chosen. The most frequently used one is the *logistic regression model*. It assumes the relationship between response propensity and auxiliary variables can be written as

$$\text{logit}(\rho_k(X_k)) = \log\left(\frac{\rho_k(X_k)}{1 - \rho_k(X_k)}\right) = \sum_{j=1}^p X_{kj}\beta_j, \quad (19)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of p regression coefficients. The *logit* transformation ensures that estimated response propensities are always in the interval $[0, 1]$.

Another model sometimes used is the *probit model*. It assumes the relationship between the response propensity and auxiliary variables can be written as

$$\text{probit}(\rho_k(X_k)) = \Phi^{-1}(\rho_k(X_k)) = \sum_{j=1}^p X_{kj}\beta_j, \quad (20)$$

in which Φ^{-1} is the inverse of the standard normal distribution function. Both models are special cases of the *generalised linear model* (GLM)

$$g(\rho_k(X_k)) = \sum_{j=1}^p X_{kj}\beta_j, \quad (21)$$

where g is called the *link function* that has to be specified. Another special case of the link function is the *identity* link function. This means the relationship between the response propensity and the auxiliary variables can be written as

$$\rho_k(X_k) = \sum_{j=1}^p X_{kj}\beta_j. \quad (22)$$

This is a simple *linear model*. It has advantages and disadvantages. A first advantage of the linear model is that coefficients are much easier to interpret. They simply represent the effects of the auxiliary variables on the response propensity. These effects are ‘pure’ effects. The coefficient of an auxiliary variable is corrected for the interdependencies of the other auxiliary variables in the model. Interpretation of a logit or probit model is not so straightforward. The logit or probit transformation complicates the interpretation of the model parameters.

A second advantage of the linear link function is that the computations are simpler. Estimates of the coefficients can be obtained by ordinary least squares. Estimation of the logit and probit models requires maximum likelihood estimation.

An advantage of the probit and logit models is that estimated response propensities are always in the interval $[0, 1]$. The linear model does not prevent estimated probabilities to be negative or larger than 1. However, according to [Keller et al. \(1984\)](#) the probability of estimates outside the interval $[0, 1]$ vanishes asymptotically if the model is correct and all response probabilities are strictly positive. If a linear model produces estimated response propensities outside $[0, 1]$, this is often an indication that the model does not fit very well.

It should be noted that the linear model is not necessarily a worse approximation of reality than the probit or logit model. Particularly the logit transform was introduced for convenience only, and not because this model was ‘more likely’.

[Figure 1](#) contains the graphs of the logit and probit function. It can be observed that both functions are more or less linear for values of p between, say, 0.2 and 0.8. So, the linear link function can be seen as an approximation of the other two link functions.

3.2. Application of the Logit and Linear Model

As an example, the logit model and linear model are applied in the General Population Survey (GPS). The GPS was a face-to-face survey. The target population consisted of

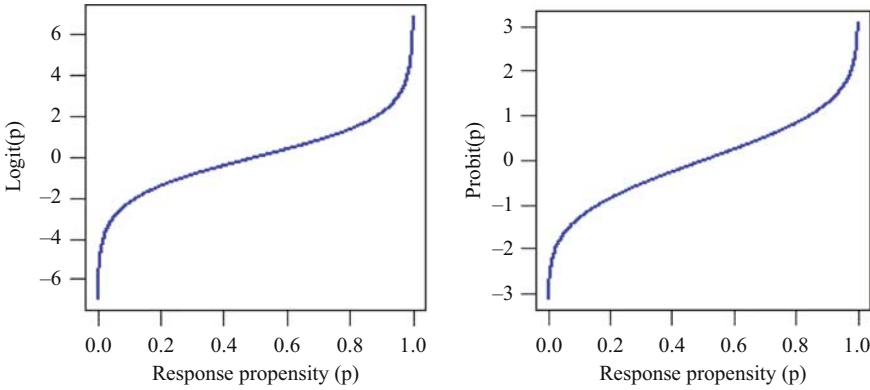


Fig. 1. The logit and probit link functions.

persons of age 12 and older. Persons were selected by means of a stratified two-stage sample. All persons had the same selection probability. The initial sample size was 32,019 people. The response consisted of 18,792 people. So, the response rate was 58.7%.

The GPS sample was linked to the Social Statistics Database (SSD) of Statistics Netherlands. The SSD contains a large set of variables for every person living in the Netherlands. These variables have been retrieved from registers and other administrative sources. By linking the GPS to the SSD, the values of all these variables became available for both respondents and nonrespondents. [Table 1](#) lists the variables that have been used in this article.

Not all auxiliary variables were included in the models for the response propensities. A simple selection procedure was used to determine only the relevant ones. These are variables having a relationship with response behaviour. The strength of this relationship was measured with Cramér's V . It is defined by

$$V = \sqrt{\frac{\chi^2}{n \times \min(r - 1, c - 1)}}. \tag{23}$$

χ^2 is the chi-square statistic for the contingency table obtained by crossing two categorical variables, n is the total number of observations in the table, r is the number of rows and c is

Table 1. Auxiliary variables that were available for the GPS.

| Variable | Description | Categories |
|----------|--------------------------------------|------------|
| Gender | Gender | 2 |
| Married | Is married (yes / no) | 2 |
| Age13 | Age in 13 age groups | 13 |
| Ethnic | Ethnic background | 5 |
| HHType | Type of household | 5 |
| Phone | Has listed phone number (yes / no) | 2 |
| HasJob | Has a job (yes / no) | 2 |
| HouseVal | Average house value in neighbourhood | 5 |
| Region | Region of the country | 5 |
| Urban | Degree of urbanisation | 5 |

Table 2. Cramér’s *V* for the auxiliary variables.

| Variable | Cramér’s <i>V</i> | In model |
|----------|-------------------|----------|
| Region | 0.163 | Yes |
| Urban | 0.153 | Yes |
| Phone | 0.150 | Yes |
| HouseVal | 0.112 | Yes |
| Ethnic | 0.112 | Yes |
| HHType | 0.106 | Yes |
| Married | 0.096 | Yes |
| Age13 | 0.061 | No |
| HasJob | 0.037 | No |
| Gender | 0.011 | No |

the number of columns. *V* always assumes a value in the interval [0, 1]. *V* = 1 means a perfect relationship, and *V* = 0 means no relationship at all. Here, one of the variables is the auxiliary variable, and the other variable is the response variable (with categories *Yes* and *No*). Table 2 contains the values of Cramér’s *V* for all available auxiliary variables.

It is clear from the table that response behaviour has the strongest relationship with the region in which people live (variable *Region*). The second variable is degree of urbanisation (variable *Urban*), which could partly measure the same aspect as region: people in rural areas are more likely to respond than people in urban areas. The relatively high value for the variable *Phone* (has a listed phone number) implies that people with a listed phone number are more likely to respond than those without it.

It was decided (rather arbitrarily) to include the seven auxiliary variables in the logit model for which Cramér’s *V* has value larger than 0.090. Hence, the model contained the first seven variables in Table 2. It should be noted that this selection technique is only a simple one. There are more advanced techniques, like stepwise inclusion techniques that only add variables having a significant contribution, see, for example Bethlehem et al. (2011, chap. 9).

Note that all auxiliary variables are categorical. To be able to include them in the model, each variable is split into a set of dummy variables. There are as many dummy variables as the variable has categories. So there is a dummy variable for each category of each explanatory variable. Unique identification of this model requires some restrictions to be imposed. This can be done in various ways. Here, the coefficient of one of the dummy variables is set to 0. All other coefficients in the set represent deviations from this fixed value.

The logistic regression model was fitted, and subsequently used to estimate the response propensities. Figure 2 shows the distribution of the estimated response propensities. There is a substantial variation. The probabilities vary between 0.128 and 0.732.

Estimated response propensities can be used for the analysis of the nonresponse. Such a (numerical or graphical analysis) can give insight into possible relationships between response behaviour and auxiliary variables. Figure 3 shows an example. It is a boxplot of the response propensities by degree of urbanisation. There is a clear pattern: the lower the degree of urbanisation, the higher the response rate. Response is low in urban areas and high in rural areas.

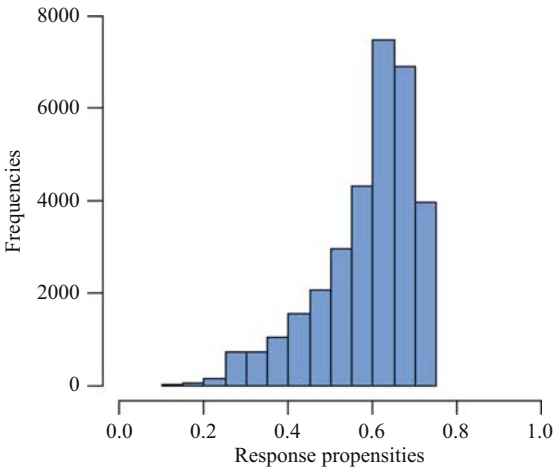


Fig. 2. Histogram of the estimated response propensities (logit model).

The response probabilities were also estimated with a linear model. The same auxiliary variables were included as for the logit model. Again, each explanatory variable was split into dummy variables, and extra restrictions were imposed to allow for unique identification: the last coefficient for each set of dummy variables was set to zero. For example, the variable *Phone* (has listed phone number) had two categories: *No* and *Yes*. The coefficient of *Yes* was set to 0. The estimate of the coefficient for *No* turned out to be equal to -0.108. So, not having a listed phone number reduced the response propensity by 0.108.

The estimated response propensities for the linear model varied between 0.050 and 0.738. So, all estimates were within the interval [0, 1]. Note that the smallest response probability for the linear model (0.050) is somewhat smaller than the one for the logit model (0.128). To see how much the estimated response propensities differ for the two models, they were plotted in a scatter plot, see [Figure 4](#).

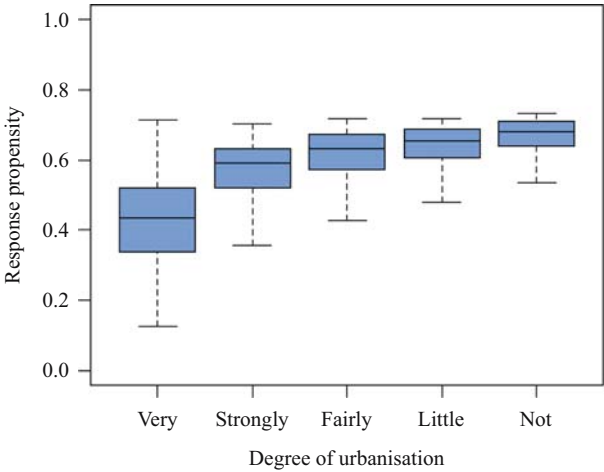


Fig. 3. Boxplot of estimated response propensities by degree of urbanisation for the General Population Survey (GPS).

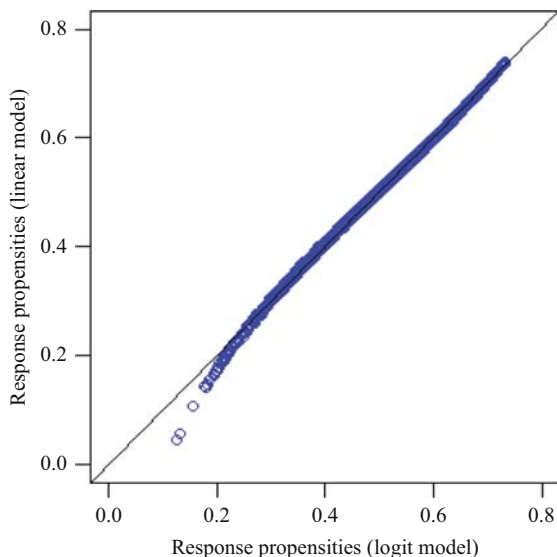


Fig. 4. Response propensities of the logit model and the linear model.

There is an almost perfect linear relationship between the response propensities of both models. This is confirmed by the value of the correlation coefficient, which is equal to 0.999573. Hence, one can conclude that, at least in this example, both models result in almost the same response propensities.

4. Adjustment Weighting With Probabilities

4.1. Weighting Adjustment

Selective nonresponse may cause estimators to be biased. To correct for such a bias, usually some *weighting adjustment technique* is applied. The basic idea is to assign weights to responding elements in such a way that over-represented groups get a weight smaller than one and under-represented groups get a weight larger than one.

There are several types of weighting techniques. The most frequently used ones are *post-stratification*, *generalised regression estimation*, and *raking ratio estimation*. Weighting is based on the use of *auxiliary information*. Auxiliary information is defined here as a set of variables that have been measured in the survey (auxiliary variables), and for which the distribution in the population, or in the complete sample, is available.

It is also possible to use response propensities for weighting adjustment. This can be done in several ways. Two approaches are described in this chapter. The first approach is *response propensity weighting*. It is based on the principle of [Horvitz and Thompson \(1952\)](#) that always an unbiased estimator can be constructed if the selection probabilities are known. In case of nonresponse, selection depends on both the sample selection mechanism and the response mechanism. The idea is now to adapt the Horvitz-Thompson estimator by including the (estimated) response probabilities.

A second approach is *response propensity stratification*. It is based on the fact that estimates will not be biased if all response probabilities are equal. In this case, selection problems will only lead to fewer observations, but the composition of the sample is not affected. The idea is to divide the sample in strata in such a way that all elements within a stratum have (approximately) the same response probability. Consequently, unbiased estimates can be computed within strata. Next, stratum estimates are combined into a unbiased population estimate.

First, the three traditional weight adjustment techniques (post-stratification, generalised regression estimation, and raking ratio estimation) are described. Then it is shown how response propensities can be used for weighting. The two mentioned approaches (response propensity weighting and response propensity stratification) are described. This section concludes with an example of the application of these approaches.

4.2. Post-Stratification

Post-stratification is a well-known and often used weighting technique, see, for example [Cochran \(1977\)](#) or [Bethlehem \(2002\)](#). To carry out post-stratification, categorical variables are needed. By crossing these variables, population and sample are divided into a number of non-overlapping subpopulations, called *strata*.

All objects in one stratum are assigned the same weight, and this weight is equal to the population proportion in that stratum divided by the response proportion in that stratum. Suppose that crossing the stratification variables produces L strata. The number of population objects in stratum h is denoted by N_h , for $h = 1, 2, \dots, L$. Hence, the population size is equal to $N = N_1 + N_2 + \dots + N_L$. The weight w_k for an object k in stratum h is now defined by

$$w_k = \frac{N_h/N}{m_h/m}, \quad (24)$$

where m_h is the number of respondents in stratum h (with $m_h < n_h$), and n is the total number of respondents (with $m < n$). If the values of the weights are taken into account, the result is the post-stratification estimator

$$\bar{y}_{ps} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h \quad (25)$$

where \bar{y}_h is the response mean in stratum h . So, the post-stratification estimator is equal to a weighted sum of response means in the strata. The bias of the post-stratification estimator is equal to

$$B(\bar{y}_{ps}) = \frac{1}{N} \sum_{h=1}^L N_h B(\bar{y}_h) = \frac{1}{N} \sum_{h=1}^L N_h \frac{R_{\rho Y, h} S_{\rho, h} S_{Y, h}}{\bar{\rho}_h} \quad (26)$$

where $R_{\rho Y, h}$ is the correlation coefficient between the response probability and the target variable in stratum h , $S_{\rho, h}$ is the standard deviation of the response probabilities in stratum h , $S_{Y, h}$ is the standard deviation of the target variable in stratum h , and $\bar{\rho}_h$ is the average response probability in stratum h .

It can be concluded that the bias of weighted estimates is small if there is a strong relationship between the target variable and the stratification variables. The variation in the values of the target variable should manifest itself between strata, but not within strata. In other words, strata should be homogeneous with respect to the target variables. In nonresponse correction terminology, this situation comes down to Missing At Random (MAR).

The bias of the estimator will also be small if the variation of the response probabilities is small within strata. This implies that there must be strong relationships between the auxiliary variables and the response probability.

In conclusion, application of post-stratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy the following conditions:

- They must be measured in the survey;
- Their population (or complete sample) distribution must be available;
- They must be strongly correlated with all target variables;
- They must be strongly correlated with the response behaviour.

Unfortunately, such variables are often not available. If weakly correlated variables are used, the bias will only be partly removed.

4.3. Generalised Regression Estimation

Post-stratification is a simple and straightforward weighting technique. Unfortunately, it is not always possible to apply post-stratification. For example, if there are many auxiliary variables, cross-classifying them may result in empty strata. It is not possible to compute weights for such strata. These problems can be avoided by applying more advanced weighting adjustment techniques. Such techniques are described in [Bethlehem \(2002\)](#) and [Särndal and Lundström \(2005\)](#). One of these techniques is *generalised regression estimation*. It is sometimes also called *linear weighting*.

Generalised regression estimation assumes there is a set of auxiliary variables X_1, X_2, \dots, X_p that can be used to predict the values of a target variable Y . The generalised regression estimator is defined by

$$\bar{y}_{GR} = \bar{y} + (\bar{X} - \bar{x})'b, \quad (27)$$

in which \bar{y} is the sample mean of the target variable. \bar{X} is the vector of population means of the auxiliary variables, and \bar{x} is the vector of response means of these variables. Furthermore, b is the (estimated) vector of regression coefficients. The estimator reduces the bias if the underlying regression model fits the data well.

Post-stratification is a special case of generalised regression estimation. If the stratification is represented by a set of dummy variables, where each dummy variable denotes a specific stratum, Equation (27) reduces to Equation (25).

By rewriting Equation (27), it can be shown that generalised regression estimation is a form of weighting adjustment, see, for example, [Bethlehem et al. \(2011\)](#). The value of a weight for a specific respondent is determined by using the corresponding values of the auxiliary variables.

Generalised regression estimation can be applied in more situations than post-stratification. For example, post-stratification by age class and sex requires the population distribution of the crossing of age class by sex to be known. If just the marginal population distributions of age class and sex separately are known, post-stratification cannot be applied. At most, only one variable can be used. However, generalised regression estimation makes it possible to specify a regression model that contains both marginal distributions. In this way, more information is used, and this will generally lead to better estimates.

Generalised regression estimation has the disadvantage that some correction weights may turn out to be negative. Such weights are not wrong, but simply a consequence of the underlying model. Usually, negative weights indicate that the regression model does not fit the data too well. Some analysis software packages are able to take into account weights, but do not accept weights to be negative. This may be a reason not to apply generalised regression estimation.

It should be noted that generalised regression estimation will only substantially reduce the bias if Missing At Random (MAR) applies to the set of auxiliary variables used. For more about generalised regression estimation, see, for example [Bethlehem and Keller \(1987\)](#).

4.4. Raking Ratio Estimation

Correction weights produced by generalised regression estimation are the sum of a number of weight coefficients. It is also possible to compute correction weights in a different way, namely as the product of a number of weight factors. This weighting technique is usually called *raking ratio estimation*, *iterative proportional fitting*, *RIM weighting* (RIM stands for Random Iterative Method), or *multiplicative weighting*.

Raking ratio estimation can be applied in the same situations as generalised regression estimation, as long as only categorical auxiliary variables are used. Correction weights are the result of an iterative process, in which a series of post-stratifications is carried out repeatedly. This is shown schematically in [Figure 5](#). After post-stratification 1 is carried out, the survey has become representative with respect to the variables of this post-stratification. Then post-stratification 2 is carried out. This adapts the weights. The survey becomes representative with respect to the variables of this post-stratification, but representativity with respect to the variables of post-stratification 1 is lost. However, the

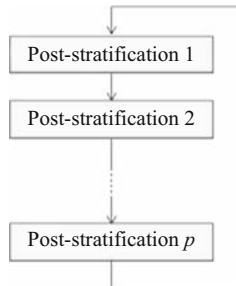


Fig. 5. Raking ratio estimation.

deviation from representativity is smaller than it was. After all (p) post-stratifications have been dealt with, the loop of p post-stratifications starts again. In every post-stratification in every loop the weights are adapted. The process stops if the values of the weights do not change any more. Then the weighting process has converged.

Multiplicative weighting has the advantage that computed weights are always positive. It has the disadvantage that there is no clear model underlying this approach. Moreover, there is no simple and straightforward way to compute standard errors of weighted estimates. Generalised regression estimation is based on a regression model, which allows for computing standard errors.

It should be mentioned that [Deville and Särndal \(1992\)](#) and [Deville et al. \(1993\)](#) have developed a general framework for weighting, of which raking ratio estimation and generalised regression estimation are special cases. They call this framework *calibration*. The paper by [Haziza and Beaumont \(2017\)](#) is also noteworthy. They present an overview of weighting adjustment procedures that are used by national statistical institutes.

4.5. Response Propensity Weighting

[Horvitz and Thompson \(1952\)](#) showed that it is always possible to construct an unbiased estimator if the following conditions are satisfied:

- The sample is selected by means of probability sampling;
- Each object in the target population has a positive probability of selection;
- All selection probabilities are known.

Again, let $a = (a_1, a_2, \dots, a_N)'$ denote the vector of sampling indicators, where $a_k = 1$ if object k is in the sample, and $a_k = 0$ otherwise. The selection probability π_k is defined by $P(a_k = 1)$. It is also called the *first order inclusion probability*. The Horvitz-Thompson estimator is now defined by

$$\bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N \frac{a_k Y_k}{\pi_k} \quad (27)$$

In case of full response, this is an unbiased estimator. In case of nonresponse, however, only the data of the responding objects can be used, and this results in a biased estimator. One way of solving this problem is to include the nonresponse mechanism in the estimator. Let $R = (R_1, R_2, \dots, R_N)'$ denote the response indicators, and $\rho = (\rho_1, \rho_2, \dots, \rho_N)$ the corresponding response probabilities, then

$$\bar{y}_{HT} = \frac{1}{N} \sum_{k=1}^N \frac{a_k R_k Y_k}{\rho_k \pi_k} \quad (28)$$

would be an unbiased estimator. However, it is not possible to use this estimator, since the values of the response probabilities are not known. The way out is to replace each response probability ρ_k by its corresponding estimated response propensity $\hat{\rho}_k(X_k)$. See Subsection 3.2 on how to estimate response propensities. This results in the adjusted

Horvitz-Thompson estimator

$$\bar{y}_{HT,R} = \frac{1}{N} \sum_{k=1}^N \frac{a_k R_k Y_k}{\hat{\rho}_k(X_k) \pi_k} \quad (29)$$

The better the estimated response propensities resemble the ‘true’ response probabilities, the smaller the bias of the estimator will be.

4.6. Response Propensity Stratification

It was already made clear in Subsection 4.2 that post-stratification can reduce the bias of estimates. The bias of the post-stratification estimator was shown to be equal to

$$B(\bar{y}_{ps}) = \frac{1}{N} \sum_{h=1}^L N_h B(\bar{y}_h) = \frac{1}{N} \sum_{h=1}^L N_h \frac{R_{pY,h} S_{p,h} S_{Y,h}}{\bar{\rho}_h} \quad (30)$$

This bias is small if the strata are homogeneous. This means that the target variable should vary between strata and not within strata. The same applies to the response probabilities: they should vary between strata and not within strata. So a post-stratification based on response probabilities helps to reduce the bias.

The idea is now to use one post-stratification variable and that is the response probability. Since the response probabilities are unknown, the estimated response propensities are used instead.

To construct strata based on estimated response propensities, a number of choices have to be made. One is how to construct the strata? They should at least be such that each stratum contains response propensities of approximately the same size. One way to do it is to divide the interval from 0 to 1, into a number of subintervals of equal length. This may result in some subintervals having many observations, and others only a few. Another way to do it is to make strata that all have the same amount of observations. More on this issue can be found in, for example, [Bethlehem et al. \(2011, chap. 11\)](#). Another choice to be made is for the number of strata to be constructed. According to [Cochran \(1968\)](#), five strata should be sufficient in most cases.

4.7. An Example

The various weighting techniques described in this section are applied in the GPS survey. The GPS was a face-to-face survey. The target population consisted of persons of age 12 and older. Persons were selected by means of a stratified two-stage sample. All persons had the same selection probability. The initial sample size was 32,019 people. The response consisted of 18,792 people. So, the response rate was 58.7%.

The auxiliary variables used were listed phone number (yes/no), married (yes/no), region, degree of urbanisation, ethnic background, house value, and type of household. Two target variables were considered: has a PC (yes/no) and owns a house (yes/no).

Five estimation approaches were applied: no adjustment, generalised regression estimation, raking ratio estimation, response propensity weighting, and response propensity stratification. The resulting estimates are summarised in [Table 3](#).

Table 3. Results of a number of weighting adjustment techniques.

| Weighting approach | Has PC | Owens house |
|------------------------------------|--------|-------------|
| No adjustment weighting | 57.4 % | 62.5 % |
| Generalised regression estimator | 55.7 % | 58.5 % |
| Raking ratio estimation | 55.7 % | 58.6 % |
| Response propensity weighting | 55.7 % | 58.6 % |
| Response propensity stratification | 58.8 % | 58.9 % |

For the variable *HasPC* all adjustment weighting approaches have approximately the same effect: they produce smaller estimates, and all adjusted estimates are similar. The same can be observed for the variable *Ownhouse*: weighting has an effect, and all adjusted estimates are similar.

At first sight, the results in Table 3 seem to suggest that there are no differences between the various adjustment weighting approaches. One could conclude that the type of weighting adjustment does not matter as long as the right auxiliary variables are used. Of course, this is only one example. It takes more research to establish whether or not this conclusion can be generalised.

5. From Weights to Response Probabilities

5.1. Weighting Adjustment

Estimation of response propensities requires the values of the auxiliary variable to be known for the nonrespondents. This information is not available for many surveys. So then it is not possible to work with estimated response propensities. Still, there is a trick to do this. It makes use of the relation between adjustment weights and response propensities: inverse response propensities can be seen as adjustment weights. The idea is to first carry out some weighting technique and then to transform the weights into response propensities.

There are several types of weighting adjustment techniques. The most frequently used ones are post-stratification, generalised regression estimation and raking ratio estimation. Weighting is based on the use of *auxiliary information*. This is the set of variables that have been measured in the survey, and for which the distribution in the population, or in the complete sample, is available. Note that the individual values of the auxiliary variables are not required for the nonresponding objects. This is in contrast to the techniques discussed in Section 3. It is explored here whether it is possible to estimate the response probabilities using weights that are produced by a weighting model that only uses the marginal distributions of a set of auxiliary variables.

5.2. Estimating the Response Probabilities

It is now shown how weights, computed by means of generalised regression estimation or raking ratio estimation, could be transformed into response propensities.

Let there be p categorical auxiliary variables. The values of these variables for object k are denoted by the vector

$$X_k = (X_k^{(1)}, X_k^{(2)}, \dots, X_k^{(p)}), \quad (31)$$

The number of categories of variable $X^{(j)}$ is denoted by C_j , for $j = 1, 2, \dots, p$. The categories are assumed to be numbered $1, 2, \dots, C_j$.

Whether generalised regression estimation or raking ratio estimation is applied, all responding objects with the same set of values for the auxiliary variables will be assigned the same weight. Suppose an object is in category k_1 of the first variable, category k_2 of the second variable, \dots , and category k_p of the p -th variable. Let $w(k_1, k_2, \dots, k_p)$ denote the corresponding weight. Furthermore, assume there are $r(k_1, k_2, \dots, k_p)$ respondents in this group. The number of sample elements $n(k_1, k_2, \dots, k_p)$ in the group can now be estimated by

$$\hat{n}(k_1, k_2, \dots, k_p) = \frac{n}{n_R} \times w(k_1, k_2, \dots, k_p) \times r(k_1, k_2, \dots, k_p), \quad (32)$$

where n is the sample size and n_R is the total number of respondents. The response propensity for all objects in the group can now be estimated by

$$\hat{\rho}(k_1, k_2, \dots, k_p) = \frac{r(k_1, k_2, \dots, k_p)}{\hat{n}(k_1, k_2, \dots, k_p)} = \frac{n_R}{n} \times \frac{1}{w(k_1, k_2, \dots, k_p)}. \quad (33)$$

Indeed, the response propensities are inversely proportional to the weights.

5.3. Application to the GPS

The data of the GPS survey are now used to explore the behaviour of response propensities that have been computed from weights. First, the generalised regression estimator is applied. The auxiliary variables are the same as those in the logit model and the linear model of Subsection 3.2. There are seven variables: *Phone*, *Married*, *Region*, *Urban*, *Ethnic*, *HouseVal*, and *HHType*. Only their marginal distributions are used for computing the weights. So there are no interactions in the weighting model.

Note that not the population distributions of the auxiliary variables are used to compute the weights, but the complete sample distributions. The sample distributions are unbiased estimates of the population distributions. So they have some margin of error.

It is assumed that the individual values of the auxiliary variables are only available for the responding elements, and not for the nonresponding elements. So less information is used than in the case of the logit or linear model in Subsection 3.2. As a consequence, response propensities can only be computed for the responding elements.

Figure 6 shows the relationship between the logit model response propensities and the generalised regression estimation response propensities. There is a strong relationship. The correlation coefficient is equal to 0.9801535.

The linear relationship is somewhat less strong than that between the logit model response propensities and the linear model response propensities (with a correlation of 0.999573). Three clusters of points can be distinguished in the scatter plot of Figure 6. Further analysis showed that the two line-shaped clusters with lower response propensities mainly contain people in highly urbanised areas. Persons living in rural areas can all be found in the banana-shaped cluster of high response propensities.

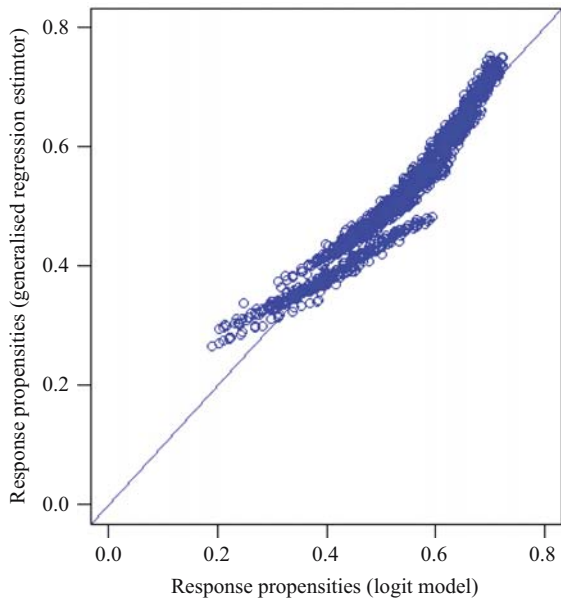


Fig. 6. Comparing response propensities produced by the logit model and generalised regression estimation.

The exercise was repeated using raking ratio estimation instead of the generalised regression estimation. Again, weights were transformed into response propensities. Figure 7 shows the relationship between the logit model response propensities and the raking ratio estimation response propensities. There is a strong relationship. The correlation coefficient is equal to 0.9937689.

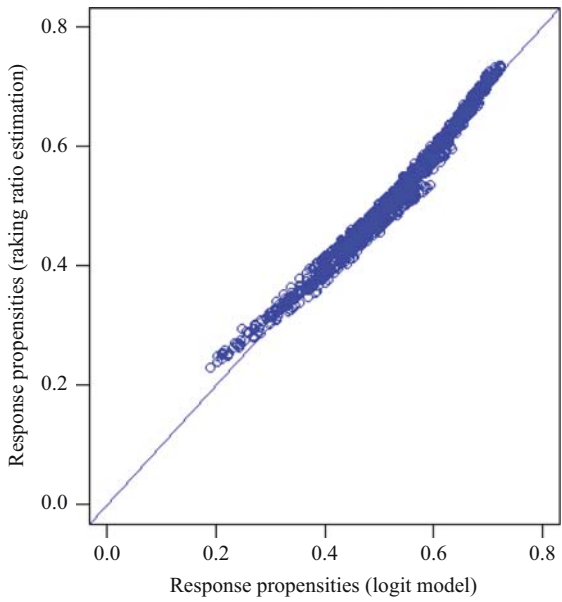


Fig. 7. Comparing response propensities produced by the logit model and raking ratio estimation.

In this example, raking ratio estimation seems to produce response propensities that are closer to those of the logit model than generalised regression estimation. Apparently, the individual values of the auxiliary variables are not needed in this case for estimating response propensities. However, this is just one example. More research is required to make clear whether or not this is a general phenomenon.

6. Using Response Propensities to Assess Representativity

6.1. The Coefficient of Variation

As was already described in Subsection 2.2, the bias of the response mean as an estimator for the population mean is equal to

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (34)$$

where $R_{\rho Y}$ is the correlation coefficient between target variable and the response behaviour, S_{ρ} is the standard deviation of the response probabilities, S_Y is the standard deviation of the target variable, and $\bar{\rho}$ is the average response probability. Equation (34) can be rewritten as

$$B(\bar{y}_R) = R_{\rho Y} \times S_Y \times CV_{\rho}, \quad (35)$$

where CV_{ρ} is the *coefficient of variation* (CV) of the response probabilities. It is the only component in the expression for the bias that purely depends on the response probabilities. A large coefficient of variation means that there is a potential risk of a large bias. How large the bias for a specific variable will be, depends on the strength of the relationship between the target variable and the response probabilities.

CV_{ρ} can be used as an indicator of representativity: the larger the value of CV_{ρ} , the larger the lack of representativity. A CV_{ρ} of 0 means that all response propensities are equal, which implies there is no bias.

Note that there are other indicators of representativity. The indicator presented in this section is based on the coefficient of variation of the (estimated) response probabilities. Schouten et al. (2009) propose the R-indicator, which is based on the standard deviation of the (estimated) response probabilities. Related to the concept of representativity is the concept of the imbalance of the response set, which was introduced by Särndal. More about this approach can be found in for example Särndal (2011), Lundquist and Särndal (2013), Särndal and Lundquist (2014a, 2014b, 2017) and Särndal et al. (2016).

6.2. Case 1: Individual Values for the Nonrespondents Are Available

If the individual values of the auxiliary variables are available for both respondents and nonrespondents, the logit model or the linear model, as described in Subsection 3.1, can be applied. For each sample element, the response propensity can be estimated. Therefore, CV_{ρ} can be computed for the sample, and this is an estimator of the population-based CV_{ρ} . Note that for small samples, this indicator may be somewhat biased.

The data of the GPS survey are used for an illustration. Response propensities were estimated using the seven auxiliary variables *Phone*, *Married*, *Region*, *Urban*, *Ethnic*,

Table 4. Computation of the CV for the GPS survey (case 1).

| Model | Estimated response propensities | | | | CV_{ρ} |
|--------|---------------------------------|---------|-------|--------------------|-------------|
| | Minimum | Maximum | Mean | Standard deviation | |
| Logit | 0.128 | 0.732 | 0.587 | 0.112 | 0.191 |
| Linear | 0.050 | 0.738 | 0.587 | 0.112 | 0.191 |

HouseVal, and *HHType*. Only main effects were used in the logit and linear model. The computations for both models are summarised in [Table 4](#)

Although the linear model produces a somewhat wider range of values for the response propensities, the values of the CV_{ρ} are approximately the same. At least in this example, the linear model can be used as an approximation of the logit model.

6.3. Case 2: Individual Values for the Nonrespondents are Not Available

If the individual values of the auxiliary variables are not available for the nonrespondents, the weighting approach may be considered for estimating response propensities. Section 5 describes how to do this. This approach requires the population distribution or the complete sample distribution to be known.

It should be noted that the response propensities can only be estimated for respondents and not for nonrespondents. These response propensities cannot be used without correction to estimate the standard deviation of the all response propensities in the sample. The reason is that elements with high response propensities will be over-represented in the response. Fortunately, there is a way out. Let

$$\bar{r}_R = \frac{\sum_{k=1}^N a_k R_k \rho_k}{\sum_{k=1}^N a_k R_k} \tag{36}$$

denote the response mean of the response probabilities. The expected value of this quantity is approximately equal to

$$E(\bar{r}_R) \approx \bar{\rho}_R \equiv \frac{1}{N\bar{\rho}} \sum_{k=1}^N \rho_k^2 \tag{37}$$

By rewriting Equation 37, it can be shown that the standard deviation of the response probabilities is equal to

$$S_{\rho} = \sqrt{\bar{\rho}(\bar{\rho}_R - \bar{\rho})} \tag{38}$$

In practice, the mean $\bar{\rho}$ of the response probabilities is estimated by the response rate n_R / n . The quantity $\bar{\rho}_R$ is estimated by the mean of the estimated response propensities of the respondents. This assumes simple random sampling. For unequal probability sampling designs the Horvitz-Thompson estimator should be used, which means that values are weighted with the inverse inclusion probabilities. [Table 5](#) summarises the results of the computations for all four approaches considered in this article.

Table 5. Computation of the CV_ρ for the GPS survey.

| Approach | Standard deviation | CV_ρ |
|-----------------------------------|--------------------|-----------|
| Logit model | 0.097 | 0.160 |
| Linear model | 0.097 | 0.160 |
| Generalised regression estimation | 0.107 | 0.176 |
| Raking ratio estimation | 0.102 | 0.168 |

Note that here computations are based on respondents only. This why the values for the logit and linear model differ from those in [Table 4](#).

Although less information is used, raking ratio estimation seems to perform almost as well as the logit and the linear model. Generalised regression estimation performs slightly less than raking ratio estimation, but still produces an estimate that is close to the logit estimates.

Again, it must be remarked that this conclusion is based on application to just one data set. More research is required to find out whether this holds in general.

6.4. Application to a Self-Selection Web Survey

The theory developed for estimating response propensities from adjustment weights can be applied to self-selection surveys. There is no sample selection for such a survey. There are no selection probabilities, but only response probabilities. To say it differently: the whole population is the sample.

Typically, the values of auxiliary variables are only available for the participants, and not for the non-participants. It is assumed that it is possible to obtain the population distributions of the auxiliary variables for weighting purposes. After weights have been computed, they are transformed into response propensities, and they can be used to compute the CV_ρ . The CV_ρ takes the form

$$CV_\rho = \frac{\sqrt{\bar{\rho}(\bar{\rho}_R - \bar{\rho})}}{\bar{\rho}} = \sqrt{\frac{(\bar{\rho}_R - \bar{\rho})}{\bar{\rho}}}. \quad (39)$$

The mean response probability $\bar{\rho}$ is estimated by n_S / N , where n_S is the size of the realised response and N is the size of the target population. The quantity $\bar{\rho}_R$ is estimated by the mean of the estimated response propensities for the respondents.

The theory is applied in an example. There are three nationwide public TV channels in the Netherlands. One of these channels ('NOP1') has a current affairs program called 'EenVandaag'. This program maintains a web panel. It is used to measure public opinion with respect to topics that are discussed in the program. The 'EenVandaag Opinion Panel' started in 2004. In 2008, it contained approximately 45,000 members. The panel is a self-selection panel. Participants were recruited from the viewers of the program. For these reasons, the panel lacks representativity.

In the period before the start of the Olympic Games in Beijing in August of 2008 there was a lot of discussion in the Netherlands about a possible boycott of the games. Suggestions ranged from not showing up at the opening ceremony to athletes not

participating in the games at all. This boycott was proposed because of the lack of respect of the Chinese for the human rights of the Tibetan people. One of the waves of the opinion panel was conducted in April 2008 in order to determine the public opinion of the Dutch with respect to this issue. The members of the panel were invited to complete a questionnaire. This questionnaire also contained topics about other issues, like preference for political parties. The questionnaire was completed by 19,392 members of the panel aged 18 years and older.

The representativity of the response was affected by two phenomena. Firstly, the panel was constructed by means of self-selection. Secondly, not all members of the panel responded to the request to fill in the questionnaire (nonresponse).

If persons apply for membership of the panel, they have to complete a basic questionnaire with a number of demographic questions. These demographic variables can be used as auxiliary variables. The following variables were used for weighting adjustment:

- Gender in two categories: male and female;
- Age in five categories: 18–24, 25–39, 40–54, 55–64, and 65+;
- Marital Status in four categories: never married, married, divorced, widowhood;
- Province of residence in twelve categories: Groningen, Friesland, Drenthe, Overijssel, Flevoland, Gelderland, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant and Limburg;
- Ethnic background in three categories: native, first-generation non-native, and second-generation non-native;
- Voting in the 2006 general elections in twelve categories: *CDA* (Christian-democrats), *PvdA* (social-democrats), *SP* (socialists), *VVD* (liberals), *PVV* (rightwing populists), *GroenLinks* (green party), *ChristenUnie* (right-wing Christians), *D66* (liberal-democrats), *PvdD* (party for the animals), *SGP* (right-wing Christians), other party, and did not vote.

The population distributions were available for all these variables. Note that the variables came from different sources, so that only marginal distributions could be used and not cross-classifications of variables.

The first step was to compute adjustment weights. Raking ratio estimation was used for this. The resulting weights turned out to have a large variation. The smallest weight was 0.089 and the largest was 34.570. This large variation clearly points to a substantial lack of representativity.

The next step was to estimate the response propensities using expression (Equation 29). The distribution of these response propensities is shown in [Figure 8](#). It is clear that all response propensities are small. They vary approximately between 0.000 and 0.017. This is not surprising, as only 19,000 people out of a population of more than 12 million people responded.

The computations for the coefficient of variation are summarised in [Table 6](#). The coefficient of variation is a little over one. This means that, compared to the GPS survey, the potential bias of the web survey is more than five times as large.

One should be careful when comparing the CV_p of different surveys. Differences are only meaningful if the estimated response probabilities are based on the same model. If

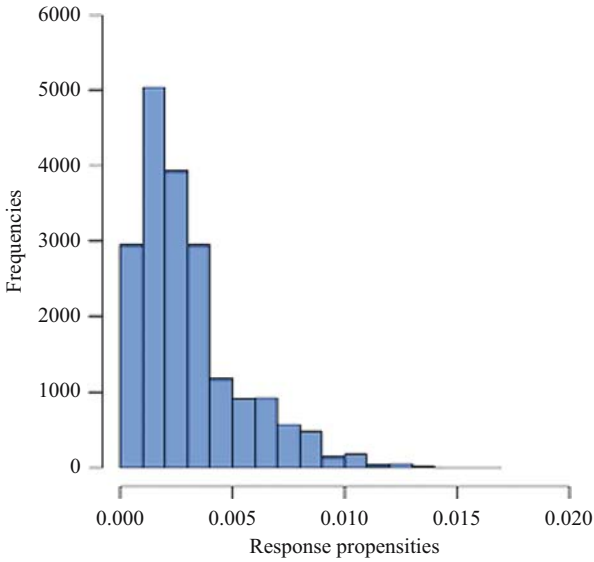


Fig. 8. Histogram of the response propensities in the self-selection survey.

this is not the case, differences may also be attributed to differences in models, and not to differences in the variation of the true response probabilities.

7. Conclusion

Nonresponse can have a serious impact on the quality of survey outcomes. Nonresponse affects the representativity of the survey and therefore the validity of its outcomes. Hence, it is important that survey researchers analyse the outcomes of their surveys. If there is a risk of biased outcomes, some kind of correction is called for.

One way of getting more insight into nonresponse is to introduce the concept of response probability. To that end, a model must be fit that is able to explain response probabilities from a set of auxiliary variables. The most frequently used model is the logistic regression model (or logit model). It is important that all relevant auxiliary variables are included in the model. It must have sufficient explanatory power.

Another model is the linear model. This is a simpler model. It can be seen as an approximation of the logistic regression model. Particularly when response probabilities

Table 6. Computation of the CV for the self-selection web survey.

| Quantity | Value |
|--|----------|
| Minimum response propensity (response) | 0.000044 |
| Maximum response propensity (response) | 0.016878 |
| Mean response propensity (response) | 0.003051 |
| Stand. dev. response propensity (response) | 0.002334 |
| CV_{ρ} (response) | 0.764931 |
| CV_{ρ} | 1.011636 |

are within the range from 0.20 to 0.08, both models produce almost the same predictions. Application in the GPS showed that it does not matter which model is used.

To be able to estimate response probabilities with the logistic regression model or the linear model, the values of the explanatory variables must be available for both respondents and nonrespondents. Sometimes this is the case, for example if the sample is selected from a population register or from a sampling frame that is linked to registers. More often the values for the nonrespondents are not available. The article proposes a technique to circumvent this problem: first, a weighting adjustment technique is applied, and which does not require the individual values of the nonrespondents. Examples are generalised regression estimation and raking ratio estimation. These techniques assign weights to respondents. These weights can be seen as reciprocal response probabilities, and therefore response probabilities can be computed.

Application of this reciprocal weights technique to the GPS data showed that it worked in practice. The estimated response probabilities were similar to those obtained by using the logistic model or linear model, even though less information was used (no interactions).

The approach of estimating response probabilities by means of weighting model models has the attractive property that it can also be applied in the case of self-selection surveys. Application of the theory in the example of the web panel shows that the worst case bias can be very large.

Various models for response probabilities were explored in this article and applied to the example of the GPS. The conclusion could be drawn that it does not seem to matter which correction technique is used, as long as the model contains the relevant auxiliary variables. Of course, this conclusion is based on just one example. More research is needed to find out whether or not this conclusion can be generalised.

Response probabilities can be used for various purposes. Not only for corrections, but also for analysis. It was shown how a graphical technique like a boxplot can give more insight into the relation between response behaviour and auxiliary variables.

Another application is to base representativity indicators on response probabilities. A well-known example is the R-indicator. This article proposes an indicator based on the coefficient of variation of the response probabilities. It seems to work for the example of a self-selection web panel. More research is necessary.

8. References

- Bethlehem, J.G. 2002. "Weighting Nonresponse Adjustments Based on Auxiliary Information." In *Survey Nonresponse*, edited by Groves, R.M., D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 275–288. New York: John Wiley & Sons.
- Bethlehem, J.G. 2009. *Applied Survey Methods, A Statistical Perspective*. Hoboken, NJ: John Wiley & Sons. DOI: <https://doi.org/10.1002/9780470494998>.
- Bethlehem, J.G. and S. Biffignandi. 2012. *Handbook of Web Surveys*. Hoboken, NJ: John Wiley & Sons. DOI: <https://doi.org/10.1002/978111812175>.
- Bethlehem, J.G., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons. DOI: <https://doi.org/10.1002/9780470891056>.

- Bethlehem, J.G. and W.J. Keller. 1987. "Linear Weighting of Sample Survey Data." *Journal of Official Statistics* 3: 141–153. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/linear-weighting-of-sample-survey-data.pdf> (accessed June 2020).
- Cochran, W.G. 1968. "The Effectiveness of Adjustments by Subclassification in Removing in Observational Studies." *Biometrics* 24: 205–213. DOI: <https://doi.org/10.2307/2528036>.
- Cochran, W.G. 1977. *Sampling techniques*. Third Edition. New York: John Wiley & Sons.
- Déville, J.C. and C.E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: <https://doi.org/10.1080/01621459.1992.10475217>.
- Déville, J.C., C.E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020. DOI: <https://doi.org/10.1080/01621459.1993.10476369>.
- Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Survey." *Public Opinion Quarterly* 70: 646–675. DOI: <https://doi.org/10.1093/poq/nfl033>.
- Hazizi, D. and J.F. Beaumont. 2017. "Constructions of Weights in Surveys." *Statistical Science* 32: 206–226. DOI: <https://doi.org/10.1214/16-STS608>.
- Horvitz, D.G. and D.J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685. DOI: <https://doi.org/10.1080/01621459.1952.10483446>.
- Keller, W.J., A. Verbeek, and J.G. Bethlehem. 1984. *ANOTA: Analysis of Tables*. Voorburg: Statistics Netherlands, Department for Statistical Methods. (CBS-report 5766-84-M1-3).
- Little, R.J.A. 1986. "Survey Nonresponse Adjustment for the Estimates of Means." *International Statistical Review* 54: 139–157. DOI: <https://doi.org/10.2307/1403140>.
- Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. Second edition. New York: John Wiley & Sons. DOI: <https://doi.org/10.1002/9781119013563>.
- Lundquist, P. and C.E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29: 557–582. DOI: <https://doi.org/10.1515/jos-2017-0033>.
- Särndal, C.-E. 2011. "The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection, in Estimation." *Journal of Official Statistics* 27: 1–21. Available at: <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-2010-morris-hansen-lecture-dealing-with-survey-nonresponse-in-data-collection-in-estimation.pdf> (accessed May 2020).
- Särndal, C.E., K. Lumiste, and I. Traat. 2016. "Reducing the Response Imbalance: Is the Accuracy of the survey estimates improved?" *Survey Methodology* 42: 219–238. Available at: <https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14663-eng.htm> (accessed June 2020).
- Särndal, C.E. and P. Lundquist. 2014a. "Balancing the Response and Adjusting Estimates for Nonresponse Bias: Complementary Activities." *Journal de la Société Française de Statistique* 155: 28–50. Available at: <https://www.semanticscholar.org/paper/Balancing-the-response-and-adjusting-estimates-for-S%C3%A4rndal-Lundquist/9686e52f8875ff2b0120303ed2e24b2008f463df> (accessed June 2020).

- Särndal, C.E. and P. Lundquist. 2014b. “Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation.” *Journal of Survey Statistics and Methodology* 2: 361–387. DOI: <https://doi.org/10.1093/jssam/smu014>.
- Särndal, C.E. and P. Lundquist. 2017. “Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance.” *Journal of Official Statistics* 33: 700–734. DOI: <https://doi.org/10.1515/jos-2017-0033>.
- Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. Chichester, UK: John Wiley & Sons.
- Schouten, B., F. Cobben, and J.G. Bethlehem. 2009. “Measures for the Representativeness of Survey Response.” *Survey Methodology* 36: 101–113. Available at: <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf?st=uZQq4P0J> (accessed June 2020).
- Stoop, I.A.L. 2005. *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office. Available at: https://www.researchgate.net/profile/Ineke_Stoop/publication/27686407_The_Hunt_for_the_Last_Respondent_Nonresponse_in_Sample_Surveys/links/569791b908ae1c427905094e/The-Hunt-for-the-Last-Respondent-Non-response-in-Sample-Surveys.pdf (accessed June 2020).
- Valliant, R., J.A. Dever, and F. Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. Heidelberg, Germany: Springer. DOI: <https://doi.org/10.1007/978-3-319-93632-1>

Received July 2018

Revised April 2019

Accepted September 2019