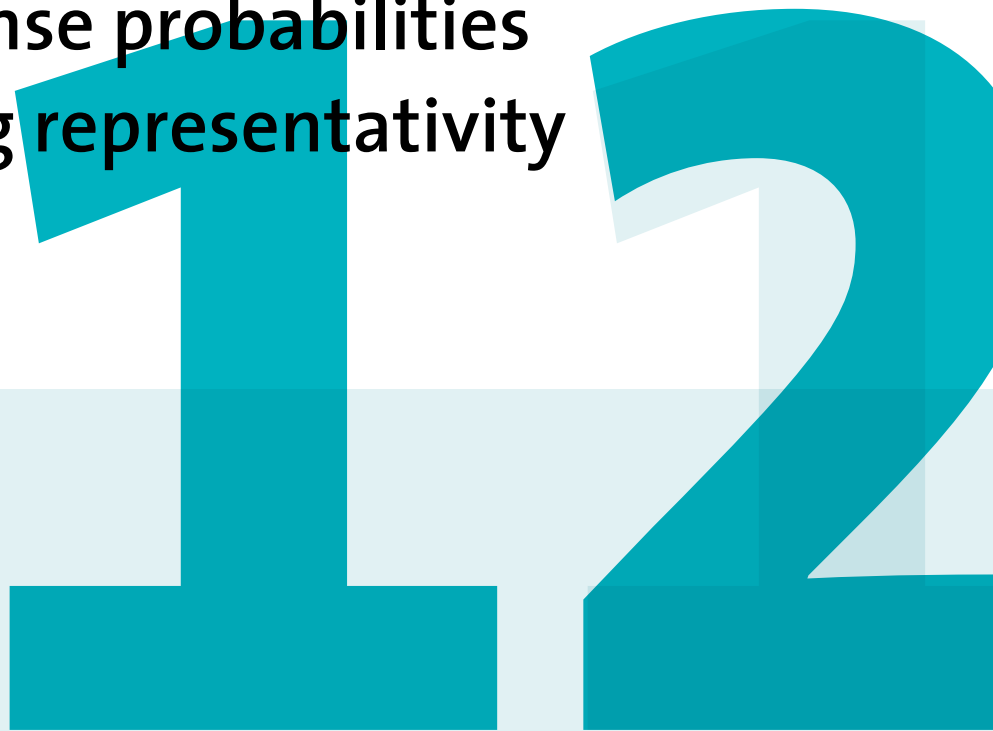


# Using response probabilities for assessing representativity



*Jelke Bethlehem*

The views expressed in this paper are those of the author(s)  
and do not necessarily reflect the policies of Statistics Netherlands

**Discussion paper (201212)**



## Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10– 2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

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

**Publisher**  
Statistics Netherlands  
Henri Faasdreef 312  
2492 JP The Hague

**Prepress**  
Statistics Netherlands  
Grafimedia

**Cover**  
Teldesign, Rotterdam

**Information**  
Telephone +31 88 570 70 70  
Telefax +31 70 337 59 94  
Via contact form:  
[www.cbs.nl/information](http://www.cbs.nl/information)

**Where to order**  
E-mail: [verkoop@cbs.nl](mailto:verkoop@cbs.nl)  
Telefax +31 45 570 62 68

**Internet**  
[www.cbs.nl](http://www.cbs.nl)

ISSN: 1572-0314

© Statistics Netherlands,  
The Hague/Heerlen, 2012.  
Reproduction is permitted,  
provided Statistics Netherlands is quoted as source.

# Using response probabilities for assessing representativity

Jelke Bethlehem

*Summary: Nonresponse in surveys may effect representativity, and therefore lead to biased estimates. A first step in exploring a possible lack of representativity is to estimate response probabilities. This paper proposes using the coefficient of variation of the response probabilities as an indicator for the lack of representativity. The usual approach for estimating response probabilities is by fitting a logit model. A drawback of this model is that it requires the values of the explanatory variables of the model to be known for all nonrespondents. This paper shows this condition can be relaxed by computing response probabilities from weights that have been obtained from some weighting adjustment technique.*

*Keywords: Nonresponse, response probability, representativity*

## 1. Introduction

There are various ways for selecting a sample for a survey, but over the years it has become clear that the only scientifically sound way to do this is by means of a probability sample. Objects (person, household, business) must have a non-zero probability of selections, and all these selection probabilities must be known. Only then can accurate, unbiased estimates of population characteristics be computed. And only then can the accuracy of these estimated by quantified, for example by means of a confidence interval.

In practice, the situation usually is not so perfect. One of the phenomena causing problems, is nonresponse. This means that no information is obtained from a number of objects in the sample. The questionnaire form remains empty.

One of the effects of nonresponse is that the sample size is smaller than expected. This would lead to less accurate, but still valid, estimates of population characteristics. This is not a serious problem as it can be taken care of by taking the initial sample size larger. A far more serious effect of nonresponse is that estimates of population characteristics may be biased. This situation occurs if, due to non-response, some groups in the population are over- or under-represented, and these groups behave differently with respect to the characteristics to be investigated. Consequently, wrong conclusions will be drawn from the survey data. Such a situation must be avoided. Therefore the amount of non-response in the fieldwork must be reduced as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of non-response usually remains. See e.g. Bethlehem, Cobben & Schouten (2011) for an overview of the nonresponse problem.

Although probability is the preferred way to select a sample, some researchers use different ways. Particularly for web surveys, often self-selection sampling is used.

The questionnaire is put on the web, and it is left to the visitors of the website to decide whether or not they will participate in the survey. No random sampling is involved. Survey respondents are those people that happen to know that 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 opt-in surveys is completely unknown and unclear, it is impossible to compute reliable estimates of population characteristics.

Both sample selection mechanisms mentioned have in common that they rely at least partly on a human decision whether or not to participate. If it would be completely clear how this decision mechanisms works, this knowledge could be used to correct the estimates. Unfortunately, this knowledge is not available. An often used approach to diminish the effects of human participation decisions is to introduce the concept of the *response probability*. It is assumed that every member of 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 also 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.

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 using auxiliary variables. This seems to work well in practical survey situations. However, this approach requires the individual values of these variables to be available for both the participants and the non-participants of the survey. Unfortunately, this is often not the case.

This paper explores some other approaches to estimate response probabilities that do not have such heavy data requirements. The first approach is to replace the logit model by a simple linear model. The second approach is to focus on weighting adjustment techniques. These techniques produce weights that correct for the lack of representativity of the survey response. Since the weights can be seen as a kind of inverted response probabilities, 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. So they present a promising approach for estimating response probabilities. Two weighting techniques are considered: generalized regression estimation and raking ratio estimation.

By taking the logit model as a benchmark, this paper investigates 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 generalized regression estimation into estimated response probabilities;
- (3) Transforming weights that have been obtained by raking ration estimation into estimated response probabilities;

The various approaches are tested using a real survey data set of Statistics Netherlands. This is an anonymized 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 also linked to some registers, providing even more auxiliary variables. So logit models can be fitted, and they can be compared with approaches requiring less data.

The estimated response probabilities are 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 normalized measure of dispersion of the response probabilities. The larger the value of the CV is, 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.

There are other indicators for the lack of representativity. Schouten, Cobben & Bethlehem (2009) propose the R-indicator, which also measures the variation of the response probability. Särndal (2011) discusses measures of imbalance based on the differences of the response means and sample means of auxiliary variables.

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 so-called 'EenVandaag Opiniepanel' of the Dutch national television station 'Nederland 1'.

## 2. The concept of response probabilities

### 2.1 Nonresponse in a simple random sample

Let the finite *survey population*  $U$  consist of a set of  $N$  identifiable elements, which are labelled 1, 2, ...,  $N$ . Associated with each element  $k$  is an unknown value  $Y_k$  of the *target variable*. The vector of all values of the target variable is denoted by

$$Y = (Y_1, Y_2, \dots, Y_N)' \quad (2.1.1)$$

Objective of the sample survey is assumed to be estimation of the population mean

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k \quad (2.1.2)$$

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

$$a = (a_1, a_2, \dots, a_N)' \quad (2.1.3)$$

of indicators. The symbol ' denotes transposition of a matrix or vector. The  $k$ -th indicator  $a_k$  assumes the value 1 if element  $k$  is selected in the sample, and otherwise it assumes the value 0.

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

$$\bar{y} = \frac{1}{n} \sum_{k=1}^N a_k Y_k \quad (2.1.4)$$

is an unbiased estimator of the population mean.

Now suppose there is nonresponse in the survey. It is assumed that each element  $k$  in the population has a certain, unknown probability  $r_k$  of response. If element  $k$  is selected in the sample, a random mechanism is activated that results with probability  $\rho_k$  in response and with probability  $1 - \rho_k$  in nonresponse. Under this model, a set of response indicators

$$R_1, R_2, \dots, R_N \quad (2.1.5)$$

can be introduced, where  $R_k = 1$  if the corresponding element  $k$  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  $a_k = 1$  and  $R_k = 1$ . Hence, the number of available cases is equal to

$$n_R = \sum_{k=1}^N a_k R_k, \quad (2.1.6)$$

Note that this realized sample size is a random variable. Likewise, the number of nonrespondents is equal to

$$n_{NR} = \sum_{k=1}^N a_k (1 - R_k), \quad (2.1.7)$$

where  $n = n_R + n_{NR}$ . The values of the target variable only become available for the  $n_R$  responding elements. The mean of these values is denoted by

$$\bar{y}_R = \frac{1}{n_R} \sum_{k=1}^N a_k R_k Y_k. \quad (2.1.8)$$

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

$$E(\bar{y}_R) \approx \tilde{Y}, \quad (2.1.9)$$

where

$$\tilde{Y} = \frac{1}{N} \sum_{k=1}^N \frac{\rho_k}{\bar{\rho}} Y_k \quad (2.1.10)$$

and

$$\bar{r} = \frac{1}{N} \sum_{k=1}^N r_k \quad (2.1.11)$$

is the mean of all response probabilities in the population. Expression (2.1.9) 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

$$B(\bar{y}_R) = \tilde{Y} - \bar{Y} = \frac{C_{rY}}{\bar{r}} = \frac{R_{rY} S_r S_Y}{\bar{r}}, \quad (2.1.12)$$

where

$$C_{rY} = \frac{1}{N} \sum_{k=1}^N (r_k - \bar{r})(Y_k - \bar{Y}) \quad (2.1.13)$$

is the covariance between the response probabilities and the values of the target variable,  $R_{rY}$  is the corresponding correlation coefficient,  $S_Y$  is the standard deviation of the variable  $Y$ , and  $S_r$  is the standard deviation of the response probabilities. From this expression of the bias 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 behavior. This implies  $R_{rY} = 0$ . The stronger the relationship between target variable and response behavior, the larger the bias will be.
- The bias vanishes if all response probabilities are equal. Then  $S_r = 0$ . 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. Translated in practical terms, this means that lower response rates will lead to larger biases.

It is clear that estimates of the values of the response probabilities provide insight into the possible effects of nonresponse on a possible bias of estimates of population characteristics.

## 2.2 Nonresponse in a self-selection sample

Self-selection means that the researcher is not in control of the sample selection process. He just makes the survey questionnaire available, and waits and sees 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 do it. 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 element  $k$  in the population has unknown probability  $\rho_k$  of participating in the survey, for  $k = 1, 2, \dots, N$ .

It is assumed in this section that there are no under-coverage problems. So in principle everyone has a nonzero probability of participating in the survey. The responding elements are denoted by the set of indicators

$$R_1, R_2, \dots, R_N, \quad (2.2.1)$$

where the  $k$ -th indicator  $R_k$  assumes the value 1 if element  $k$  participates, and otherwise it assumes the value 0, for  $k = 1, 2, \dots, N$ . The expected value  $\rho_k = E(R_k)$  is the response probability of element  $k$ . The realized sample size is denoted by

$$n_S = \sum_{k=1}^N R_k \quad (2.2.2)$$

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

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N R_k Y_k \quad (2.2.3)$$

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

$$E(\bar{y}_S) \approx \tilde{Y} = \frac{1}{N\bar{\rho}} \sum_{k=1}^N \rho_k Y_k \quad (2.2.4)$$

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

It is clear from expression (2.2.4) 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 Missing 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 & 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 (2.2.3) can be written as

$$B(\bar{y}_S) = \tilde{Y} - \bar{Y} = \frac{C_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (2.2.5)$$

in which  $C_{\rho Y}$  is the covariance between the values of target variable and the response probabilities as defined by (2.1.13) and  $\bar{\rho}$  is the average response probability. Furthermore,  $R_{\rho Y}$  is the correlation coefficient between target variable and the



response behavior,  $S_\rho$  is the standard deviation of the response probabilities and  $S_Y$  is the standard deviation of the target variable.

Expression (2.1.12) for the bias in case of random sample affected by nonresponse is identical to expression (2.2.5) 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 have response rates of around 60%. This means that the average response probability is 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. This means there is a risk of a much large bias in a self-selection survey.

From expression (2.1.13) or (2.2.5) an upper bound for the bias can be computed. Given the mean response probability  $\bar{\rho}$ , there is a maximum value the standard deviation  $S_\rho$  of the response probabilities cannot exceed:

$$S(\rho) \leq \sqrt{\bar{\rho}(1-\bar{\rho})}. \quad (2.2.6)$$

This implies that in the worst case  $S_\rho$  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

$$|B_{max}| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}. \quad (2.2.7)$$

In case of a survey based on probability sampling with a response rate of around 60%, the maximum absolute bias is therefore equal to  $0.816 \times S_Y$ . In case of a self-selection survey a size 170,000 from a population of size 12,800,00, the maximum absolute bias is  $8.619 \times S_Y$ . This is more than 10 times as large.

### 3. Estimating response probabilities

#### 3.1 Models for response probabilities

The 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), the response propensity of element  $k$  is defined by

$$\rho_k(X) = P(R_k = 1 / X_k), \quad (3.1.1)$$

where  $R_k$  is the response indicator, and  $X_k = (X_{k1}, X_{k2}, \dots, X_{kp})'$  is a vector of values of, say,  $p$  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 non-respondents.

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

$$\text{logit}(\rho_k(X)) = \log\left(\frac{\rho_k(X)}{1 - \rho_k(X)}\right) = \sum_{j=1}^p X_{kj} \beta_j, \quad (3.1.2)$$

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 response propensity and auxiliary variables can be written as

$$\text{probit}(\rho_k(X)) = \Phi^{-1}(\rho_k(X)) = \sum_{j=1}^p X_{kj} \beta_j, \quad (3.1.3)$$

in which  $\Phi^{-1}$  is the inverse of the standard normal distribution function. Both models are special cases of the generalized linear model (GLM)

$$g(\rho_k(X)) = \sum_{j=1}^p X_{kj} \beta_j, \quad (3.1.4)$$

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 response propensity and auxiliary variables can be written as

$$\rho_k(X) = \sum_{j=1}^p X_{kj} \beta_j. \quad (3.1.5)$$

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 a 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.

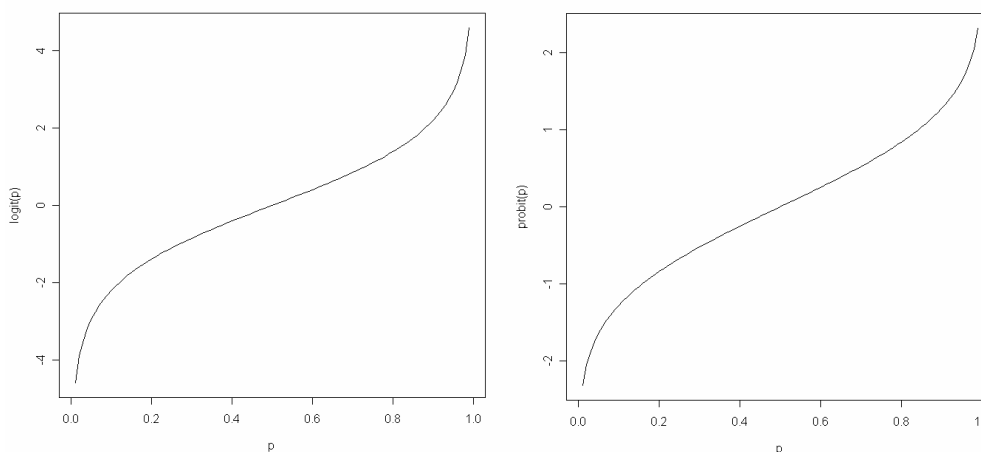
A second advantage is that the computations are simpler. Estimates of the coefficients can be obtained by ordinary least squares. Estimation of the logit and probit model requires maximum likelihood estimation.

An advantage of the probit and logit model 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 often is an indication 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 3.1.1 contains the graphs of the logit and the 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.

*Figure 3.1.1. The logit and probit link functions*



### 3.2 Application of the logit and linear model

Both the logit model and the linear model are now applied to General Population Survey (GPS). The GPS was a face-to-face survey. The target population consisted of persons of age 12 and older. Persons were selected by a stratified, two-stage sample. All persons had the same selection probability.

*Table 3.2.1. Available auxiliary variables*

Variable	Description	Categories
Gender	Gender	2
Married	Is married yes / no	2
Age13	Age in 13 age groups	13
Ethnic	Type of non-native	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 urbanization	5

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 become available for respondents and nonrespondents. Table 3.2.1 lists the variables that have been used in this paper.

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

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}} \quad (3.2.1)$$

$\chi^2$  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 the number of columns.  $V$  always assumes a value in the interval [0,1].  $V$  is equal to 1 in case of a perfect relationship, and it is equal to 0 if there is 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 3.2.2 contains the values of Cramér's  $V$  for all available auxiliary variables.

Table 3.2.2. Cramér's  $V$  for the auxiliary variables

Variable	Cramér's $V$	In model
Phone	0.149	Yes
Married	0.096	Yes
Region	0.082	Yes
Urban	0.076	Yes
Ethnic	0.056	Yes
HouseVal	0.056	Yes
HHType	0.053	Yes
HasJob	0.037	No
Age13	0.030	No
Gender	0.011	No

It is clear from the table that response behaviour has the strongest relationship with having a registered phone number (variable *Phone*). People with a registered number are more likely to respond than those without it. The second variable is *Married*. Apparently, married people respond (on average) better than unmarried people. The variables *Urban* (degree of urbanisation) and *Region* also have an effect on response behaviour. A familiar phenomenon is observed here that people in urban areas are less likely to respond than those in rural areas.

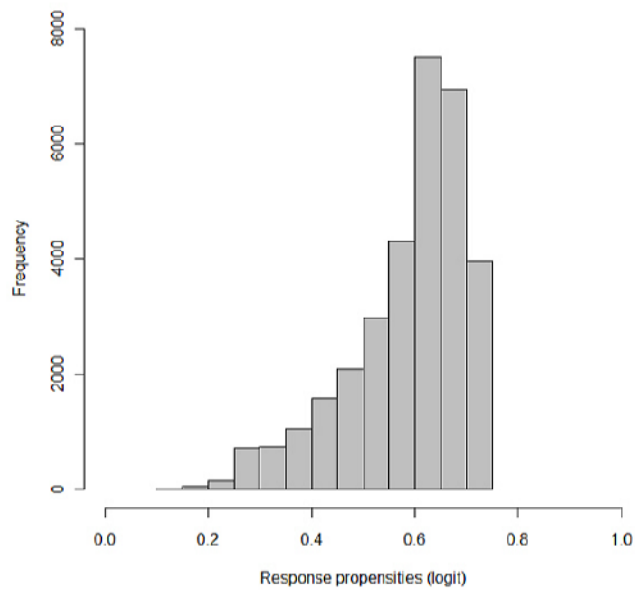
It was decided (rather arbitrarily) to include all auxiliary variables in the logit model for which Cramér's  $V$  has value larger than 0.050. Hence, the model contained the first seven variables in table 3.2.2. It should be noted there are more advanced and better techniques to select variables for a model, like stepwise inclusion techniques that only add variables having a significant contribution

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 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 model was fitted, and subsequently used to estimate the response propensities. Figure 3.2.1 shows the distribution of the estimated response propensities. There is a substantial variation. The probabilities vary between 0.128 and 0.732.

Figure 3.2.1. Histogram of the estimated response propensities (logit model)



To obtain some insight in which people have high or low response propensity, table 3.2.3 contains the values of the seven explanatory variables in the model for the lowest and highest response propensity.

Table 3.2.3. The lowest and the highest response propensity

Variable	Value for minimum probability	Value for maximum probability
Response probability	0.128	0.732
Has registered phone	No	Yes
Is married	No	Yes
Region	Big city	East
Degree of urbanization	Very	Not
Ethnic background	First gen. Non-western	Second gen. western
House value (x 1000)	<100	200-250
Household type	Other	Couple+
HasJob	No	Yes
Age class	45-49	50-54
Gender	Male	Male

The person with the lowest response propensity is a middle-age male in a cheap house in a big city. He was not born in The Netherlands but in another, non-western,

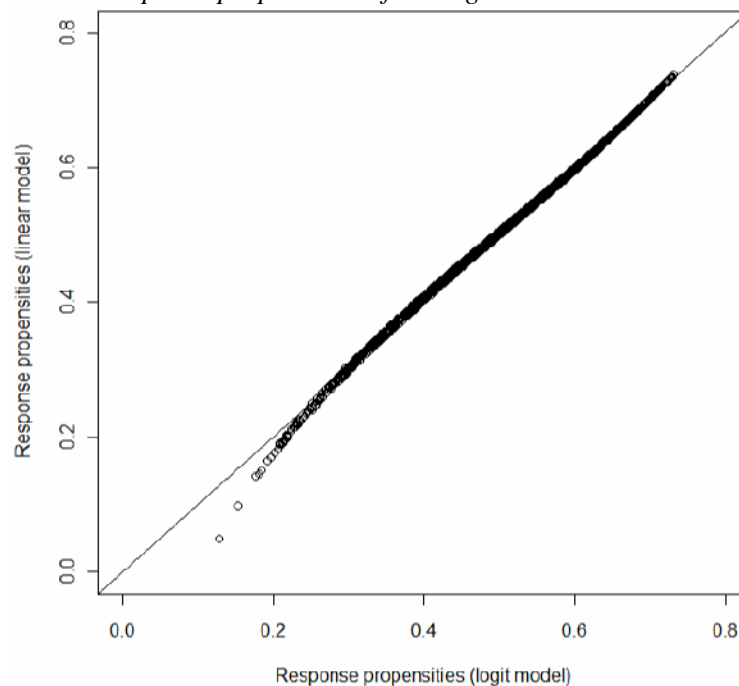
country. He is not married, does not have a job, and does not have a registered phone number.

The person with the highest response propensity is a somewhat older male in a more expensive house in a rural part of the country. He is married and has children. His parents were born in a different, western, country. He has a job and a registered phone number.

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 registered phone number) has 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 -0.109. So, not having a registered phone number reduces the response propensity by 0.109.

The estimated response propensities for the linear model varied between 0.050 and 0.738. So all estimates are within the interval [0, 1]. Not that the smallest response probability for the linear model (0.050) is somewhat smaller than that 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 3.2.2.

Figure 3.2.2. 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. So, one can conclude that, at least in this application, both models result in almost the same response propensities.

## 4. Transforming weights into response probabilities

### 4.1 Weighting adjustment

Selective nonresponse will 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 1 and under-represented groups get a weight larger than 1.

There is a relationship between response probabilities and weights: large weights correspond to small response probabilities, and vice versa. Therefore it should be possible to transform weights into estimates for response probabilities.

There are several types of weighting techniques. The most frequently used ones are post-stratification, generalized 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, 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 elements. This is in contrast to the techniques discussed in chapter 3.

It is explored here whether it is possible to estimate the response probabilities using weightings that are produced by a weighting model that only uses the marginal distributions of a set of auxiliary variables.

### 4.2 Generalized regression estimation

Generalized regression estimator is based on a linear model that attempts to explain a target variable of the survey from one or more auxiliary variables. The weights resulting from generalized regression estimation make the response representative with respect to the auxiliary variables in the model. In the terminology of Särndal (2011), one can also so that these weights make the response more balanced.

In principle, the auxiliary variables in the linear model have to be continuous variables, i.e. they measure a size, value or duration. However, it is also possible to use categorical variables. The trick is to replace a categorical variable by a set of dummy variables, where each dummy variable represents a category, i.e. it indicates whether or not a person belongs to a specific category.

Suppose there are  $p$  (continuous) auxiliary variables available. The  $p$ -vector of values of these variables for element  $k$  is denoted by

$$X_k = (X_k^{(1)}, X_k^{(2)}, \dots, X_k^{(p)})'. \quad (4.2.1)$$

Let  $Y$  be the  $N$ -vector of all values in the population of the target variable, and let  $X$  be the  $N \times p$ -matrix of all values of the auxiliary variables. The vector of population means of the  $p$  auxiliary variables is defined by

$$\bar{X} = (\bar{X}^{(1)}, \bar{X}^{(2)}, \dots, \bar{X}^{(p)})'. \quad (4.2.2)$$

This vector represents the population information assumed to be available. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector  $B = (B_1, B_2, \dots, B_p)'$  of regression coefficients for a best fit of  $Y$  on  $X$ , the residuals  $E = (E_1, E_2, \dots, E_N)'$ , defined by

$$E = Y - XB \quad (4.2.3)$$

will vary less than the values of the target variable itself. In the ideal case of a perfect relation between  $Y$  and  $X$ , all residuals will be 0. Application of ordinary least squares results in

$$B = (X'X)^{-1}XY' = \left( \sum_{k=1}^N X_k X_k' \right)^{-1} \left( \sum_{k=1}^N X_k Y_k \right). \quad (4.2.4)$$

For a simple random sample without replacement, the vector  $B$  can be estimated by

$$b = \left( \sum_{k=1}^N a_k R_k X_k X_k' \right)^{-1} \left( \sum_{k=1}^N a_k R_k X_k Y_k \right) = \left( \sum_{i=1}^{n_R} x_i x_i' \right)^{-1} \left( \sum_{i=1}^{n_R} x_i y_i \right), \quad (4.2.5)$$

where

$$x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})'. \quad (4.2.6)$$

denotes the  $p$ -vector of values of the  $p$  auxiliary variables for responding element  $i$  (for  $i = 1, 2, \dots, n_R$ ). The variable  $a_k$  indicates whether or not element  $k$  is selected in the sample, and the variable  $R_k$  indicates whether or not element  $k$  responds.. The *generalized regression estimator* is now defined by

$$\bar{y}_{GR} = \bar{y} + (\bar{X} - \bar{x})'b, \quad (4.2.7)$$

where  $\bar{x}$  is the vector of response means of the auxiliary variables.

Bethlehem and Keller (1987) have shown that the generalized regression estimator can be rewritten in the form of the weighted estimator. The adjustment weight  $w_i$  for observed element  $i$  is equal to  $w_i = v'X_i$ , and  $v$  is a vector of weight coefficients defined by

$$v = n_R \left( \sum_{i=1}^{n_R} x_i x_i' \right)^{-1} \bar{X}. \quad (4.2.8)$$

From expression (4.2.7) it is clear that computation of the weight does not require the individual values of the nonresponding elements. It is sufficient to have the population means of the auxiliary variables.

As an illustration, the case of one auxiliary variable with  $C$  categories is considered. Then  $C$  dummy variables  $X^{(1)}, X^{(2)}, \dots, X^{(C)}$  are defined. For an observation in a category  $h$ , the corresponding dummy variable  $X^{(h)}$  is assigned the value 1, and all other dummy variables are set to 0. Consequently, the vector of population means of these dummy variables is equal to



$$\bar{X} = \left( \frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_C}{N} \right), \quad (4.2.9)$$

where  $N_j$  is the number of elements in category  $j$  (in the population), for  $j = 1, 2, \dots, C$ . The vector  $v$  of weight coefficients is equal to

$$v = \frac{n_R}{N} \left( \frac{N_1}{n_{R,1}}, \frac{N_2}{n_{R,2}}, \dots, \frac{N_C}{n_{R,C}} \right). \quad (4.2.10)$$

where  $n_{R,j}$  is the number of responding elements in category  $j$ .

The generalized regression estimator will be unbiased if nonresponse does not affect the regression coefficients. Practical experience shows that nonresponse often seriously affects estimators like means and totals, but less often causes estimates of relationships to be biased. Particularly, if relationships are strong (the regression line fits the data well), the risk of finding wrong relationships is small. For a more detailed description of generalized regression estimation, see e.g. Bethlehem, Cobben & Schouten (2011).

### 4.3 Raking ratio estimation

In case of generalized regression estimation, weights are obtained by taking a linear combination of *weight coefficients*. It is also possible to compute weights in a different way, namely as the product of a number of *weight factors*. This weighting technique is usually called *raking ratio estimation*, *raking* or *multiplicative weighting*. Here it is denoted by *raking ratio estimation*. Weights are obtained as the product of a number of factors contributed by the various auxiliary variables in the model.

Raking ratio estimation can be applied in the same situations as generalized regression estimation as long as only categorical auxiliary variables are used. In this paper weighting is restricted to using only the marginal distributions of the auxiliary variables. No cross-classifications are introduced.

Weights are obtained by means of an iterative procedure. The resulting weights are the product of factors contributed by all variables in the model. To compute the weight factors, the following scheme must be carried out:

- Step 1: Introduce a weight factor for each category of each auxiliary variable in the model. Set the initial values of all factors to 1.
- Step 2: Adjust the weight factors for the first auxiliary variable so that the weighted sample becomes representative with respect to this variable.
- Step 3: Adjust the weight factors for the next auxiliary variable so that the weighted sample becomes representative for this variable. Generally, this will disturb representativity with respect to the other variables in the in the model.
- Step 4: Repeat this adjustment process until all variables in the model have been dealt with.

Step 5: Repeat steps 2, 3, and 4 until the weight factors do not change any more

Generalized regression estimation has the advantages of a clear underlying linear model, straightforward computation of weights by means of ordinary least squares, and analytic expressions for the variances of estimates. A disadvantage is that weights can be negative. Although not incorrect from a theoretical point of view, this seems counter-intuitive, and also some statistical analysis packages cannot process negative weights.

Raking ratio estimation has no simple underlying model. Weights are computed by means of iterative proportional fitting. There are no simple expressions for the variance of estimates. An advantage is that weights are always positive.

#### 4.4 Estimating the response probabilities

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

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

$$X_k = (X_k^{(1)}, X_k^{(2)}, \dots, X_k^{(p)})' \quad (4.4.1)$$

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

Whether generalized regression estimation or raking ratio estimation is applied, all responding elements with the same set of values for the auxiliary variables will be assigned the same weight. Suppose an element is in category number  $k_1$  of the first variable, category  $k_2$  of the second variable, ..., 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

$$\hat{n}(k_1, k_2, \dots, k_p) = \frac{n}{n_R} \times w(k_1, k_2, \dots, k_p) \times r(k_1, k_2, \dots, k_p), \quad (4.4.2)$$

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

$$\hat{\rho}(k_1, k_2, \dots, k_p) = \frac{r(k_1, k_2, \dots, k_p)}{\hat{n}(k_1, k_2, \dots, k_p)} = \frac{n_R}{n} \times \frac{1}{w(k_1, k_2, \dots, k_p)}. \quad (4.4.3)$$

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

#### 4.5 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 generalized regression estimator is applied. The auxiliary variables are the same as those in the logit model and the linear model of section 3.2. There are seven variables: *Phone*, *Married*,

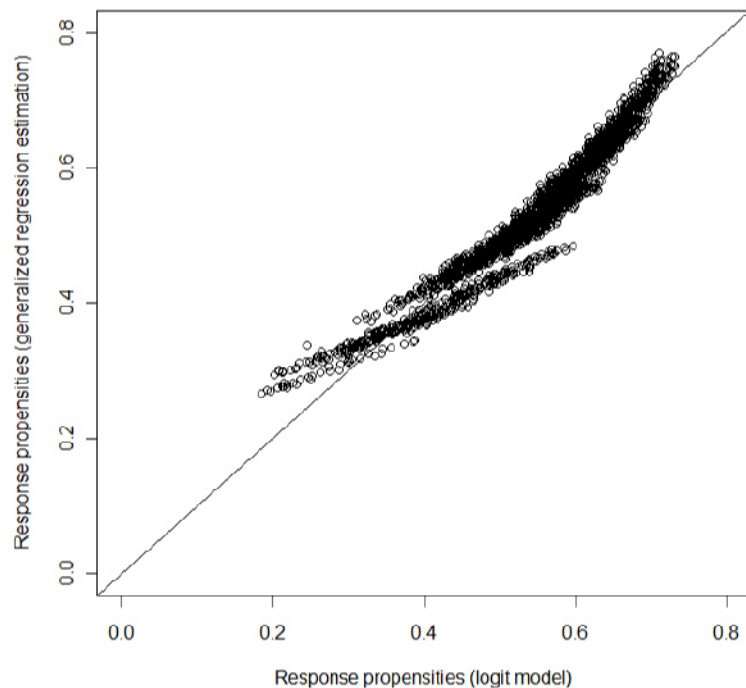
*Region, Urban, Ethnic, HouseVal, and HHType*. Only their marginal distributions are used for computing the weights. There are no interactions in the weighting model.

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. They have some margin of error and this may cause some bias in the CV, particularly for small samples.

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 section 3.2. As a consequence, response propensities can only be computed for the responding elements.

Figure 4.5.1 shows the relationship between the logit model response propensities and the generalized regression estimation response propensities. There is a strong relationship. The correlation coefficient is equal to 0.980154.

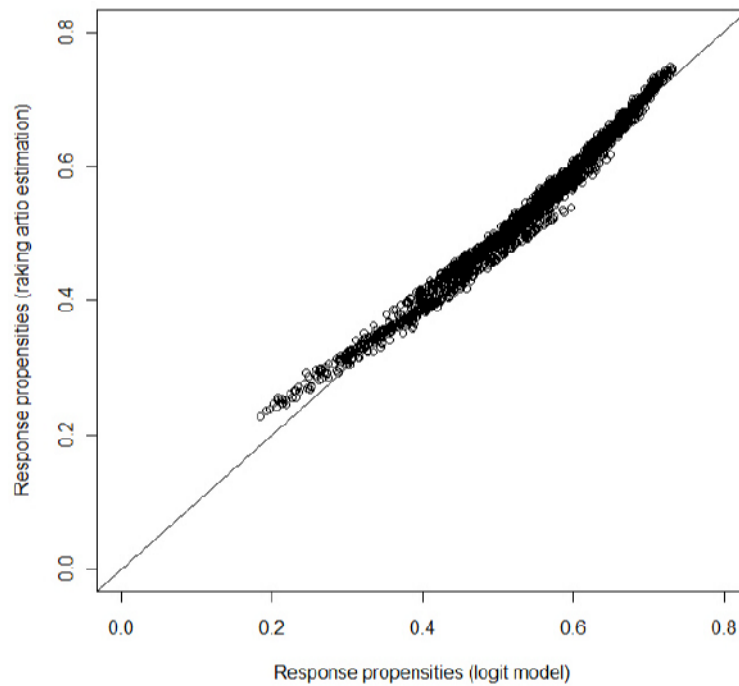
*Figure 4.5.1. Comparing response propensities produced by the logit model and generalized regression estimation.*



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 4.5.1. Further analysis showed that the two line-shaped clusters with lower response propensities mainly contain people in highly urbanized areas. Persons living in rural areas can all be found in the banana-shaped cluster of high response propensities.

The exercise was repeated using raking ratio estimation instead of the generalized regression estimation. Again, weights were transformed into response propensities. Figure 4.5.2 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 0.9937445.

Figure 4.5.2. 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 generalized regression estimation. Apparently, the individual values of the auxiliary variables are not needed in this case for estimating the response propensities. However, this is just one example. More research is required to make clear whether or not this a general phenomena.

## 5. Using response propensities to assess representativity

### 5.1 The coefficient of variation

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

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}}, \quad (5.1.1)$$

where  $R_{\rho Y}$  is the correlation coefficient between target variable and the response behavior,  $S_{\rho}$  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. Expression (2.2.5) can be rewritten as

$$B(\bar{y}_R) = R_{\rho Y} \times S_Y \times CV_{\rho}, \quad (5.1.2)$$

where  $CV_{\rho}$  is the *coefficient of variation* (CV) of the response probabilities. It is the only component in the expression of 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.

Suppose  $Y$  is an indicator variable. Then the maximum value of  $S_Y$  is equal to 0.5. This happens if the mean of  $Y$  is equal to 0.5. Consequently, the worst case bias is equal to  $CV_{\rho}/2$ .

It was already mentioned that the CV is just one indicator for the representativity of the survey response. Other indicators are the R-indicator proposed by Schouten, Cobben & Bethlehem (2009), and a measure of imbalance described by Särndal (2011).

## 5.2 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 section 3.1. can be applied. For each sample element, the response propensity can be estimated. Therefore, the CV can be computed for the sample, and this is an estimator of the population-based CV. Note that for small samples this indicator may be somewhat biased.

The data of the GPS survey are used as an illustration. Response propensities were estimated using the seven auxiliary variables *Phone*, *Married*, *Region*, *Urban*, *Ethnic*, *HouseVal*, and *HHType*. Only main effects were used in the logit and linear model. The computations for both models are summarized in table 5.2.1.

Table 5.2.1.. Computation of the CV for the GPS survey (Case 1)

Model	Estimated response propensities				CV
	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

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

The worst case bias for an indicator variable  $Y$  is approximately equal to  $0.191/2=0.1$ . Hence, the estimator for a percentage can have a bias of at most 1%.

### 5.3 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 4.4 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

$$\bar{r}_R = \frac{\sum_{k=1}^N a_k R_k r_k}{\sum_{k=1}^N a_k R_k} \quad (5.3.1)$$

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

$$E(\bar{r}_R) \approx \bar{r}_R + \frac{1}{N\bar{r}_R} \sum_{k=1}^N r_k^2 \quad (5.3.2)$$

By rewriting expression (5.3.2), it can be shown that the standard deviation of the response probabilities is equal to

$$S_r = \sqrt{\bar{r}_R (\bar{r}_R - \bar{r})} \quad (5.3.3)$$

In practice, the mean  $\bar{r}$  of the response probabilities is estimated by the response rate  $n_R / n$ . The quantity  $\bar{r}_R$  is estimated by the mean of the estimated response propensities of the respondents.

Table 5.3.1 summarizes the results of the computations for all four approaches considered in this paper.

*Table 5.3.1. Computation of the CV for the GPS survey*

Approach	Standard deviation	CV
Logit model	0.112	0.191
Linear model	0.112	0.191
Generalized regression estimation	0.117	0.199
Raking ratio estimation	0.114	0.194

Although less information is used, raking ratio estimation seems to perform almost as well as the logit and the linear model. Generalized regression estimation does not perform as well as 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.

#### 5.4 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 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. The CV takes the form

$$CV_{\rho} = \frac{\sqrt{\bar{\rho}(\bar{\rho}_R - \bar{\rho})}}{\bar{\rho}} = \sqrt{\frac{(\bar{\rho}_R - \bar{\rho})}{\bar{\rho}}}. \quad (5.4.1)$$

The mean response probability  $\bar{\rho}$  is estimated by  $n_S / N$ , where  $n_S$  is the size of the realized 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 of the respondents.

The theory is applied in an example. There are three nationwide public TV channels in The Netherlands. One of these channels (“Nederland 1”) 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 16,243 members of the panel of the age of 18 years and older.

The representativity of the response is affected by two phenomena. In the first place, the panel was constructed by means of self-selection. In the second place, not all members of the panel responded to the request to fill in the questionnaire.

If persons apply for membership of the panel, they have to complete a basic questionnaire with a number of demographic questions. These demographic

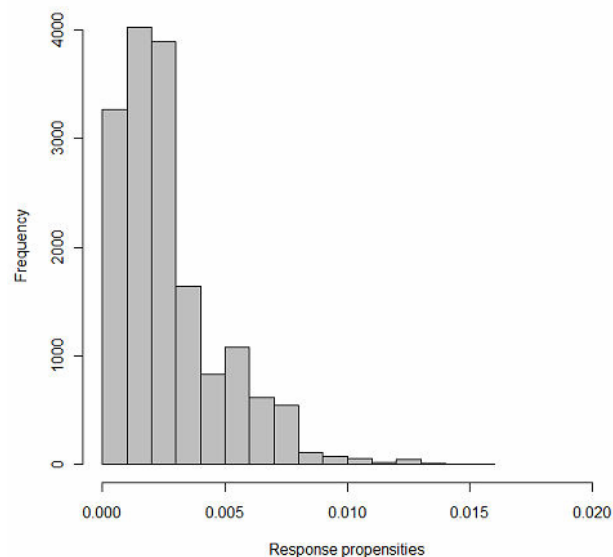
variables can be used as auxiliary variables. The following variables were used for weighting adjustment:

- Gender in two categories: male and female;
- Age in 5 categories: 18-24, 25-39, 40-54, 55-64 and 65+;
- Marital Status in 4 categories: never married, married, divorced, widowhood;
- Province of residence in 12 categories: Groningen, Friesland, Drente, Overijssel, Flevoland, Gelderland, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant and Limburg;
- Ethnic background in 3 categories: native, first-generation non-native and second-generation non-native;
- Voting at the 2006 general elections in 12 categories: CDA (Christian-democrats), PvdA (social-democrats), SP (socialists), VVD (liberals), PVV (right-wing 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 0.081 and the largest weight was 48.250. This clearly already points at a substantial lack of representativity.

*Figure 5.4.1. Histogram of the response propensities in the self-selection survey*





The next step was to estimate the response propensities using expression (4.4.3). The distribution of these response propensities is shown in figure 5.4.1. It is clear that all response propensities are small. They vary approximately between 0 and 0.015. This is not surprising as only 16,000 people out of a population of more than 12 million people responded.

The computations for the coefficient of variation are summarized in table 5.4.1. The coefficient of variation is a little over 1. This means that, compared to the GPS survey, the potential bias of the web survey is more than 5 times as large.

*Table 5.4.1. Computation of the CV for the self-selection web survey*

Quantity	Value
Minimum	0.000026
Maximum	0.015607
Mean	0.001263
Standard deviation	0.001350
Coefficient of variation	1.068492

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

In case of an indicator variable, the worst case bias is equal to  $1.068492/2 = 0.53$ . This means that the estimate for a percentage could have a bias of 53%. This shows that there is a potential risk of a large bias in this self selection survey.

## **6. Discussion**

This paper shows there are various ways to estimate response probabilities. The logit model is the most straightforward way to do this. However, the logit model requires the values of the explanatory variables to be known for all respondents and nonrespondents, and this is often not the case.

The linear model can be used as an alternative model for the logit model. At least for the applications in this paper, it does not matter much which model is used. Also application of the linear model requires the auxiliary variables to be known for respondents and nonrespondents.

Once estimates of response probabilities are available, they can be used to compute an indicator for the representativity of the survey response. The coefficient of variation (CV) is proposed for this purpose, as it is a major component in the expression of the bias of estimators.

If the values of the auxiliary variables are not available for the nonrespondents, the logit model and the linear model cannot be fitted. This paper proposes a different approach to solve this problem. The idea is to first apply some kind of weighting technique to compute weights that make the response representative with respect to

the auxiliary variables. Then, these weights can be transformed into estimates of response probabilities. The advantage of this approach is that it only requires the population distributions (or complete sample distributions) of the auxiliary variables to be known. This approach works for the sample data set in this paper. The CV's based on the logit model, the linear model and two different weighting techniques are almost the same for the data set used in this paper. Further research is necessary in order to establish if this holds in general.

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.

It should be noted that response probabilities can only be estimated properly if all relevant explanatory variables are included in the model. This corresponds to the situation that Little & Rubin (2002) call Missing at Random (MAR). The non-availability is caused by variables that are all included in the model. If not all variables are in the model, the resulting indicators of representativity may give a too optimistic picture.

It should also be noted that the representativity of two surveys can only be compared if the same models are used for the response probabilities. If this is not the case, differences can also be caused by differences in the model specifications. Therefore, one should be careful when comparing the CV of 0.191 for the GPS survey with the CV of 1.068 for the web panel.

A final remark is that the CV is a sample-based estimate. Therefore it has a margin of error and a small sample bias. When comparing CV's for different surveys (but based on the same model), this margin of error should be taken into account. Furthermore, for small sample sizes, the CV may have some bias.

## 7. References

- Bethlehem, J.G (2009), *Applied Survey Methods, A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.
- Bethlehem, J.G., Cobben, F. & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ
- Bethlehem, J.G & Keller, W.J. (1987), Linear weighting of sample survey data. *Journal of Official Statistics* 3, pp. 141-153.
- Keller, W.J., Verbeek, A. & Bethlehem, J.G. (1984), ANOTA: Analysis of Tables. CBS-report 5766-84-M1-3. Statistics Netherlands, Department for Statistical Methods, Voorburg, The Netherlands.
- Little, R.J.A & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second edition. New York: John Wiley & Sons.

- Little, R.J.A. (1986), Survey nonresponse adjustment for the estimates of means. *International Statistical Review* 54, pp. 139-157.
- Särndal, C.-E. (2011), The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official Statistics* 27, pp. 1-21.
- Schouten, B., Cobben, F. & Bethlehem, J.G. (2009), Measures for the Representativeness of Survey Response. *Survey Methodology* 36, pp. 101-113.