Journal of Official Statistics, Vol. 23, No. 1, 2007, pp. 51–68

A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption

Barry Schouten[1]

Estimates for population statistics can be seriously biased if response rates are low and the response to a survey is selective. Methods like poststratification or propensity score weighting are often employed in order to adjust for bias due to nonresponse.

One problem which many adjustment methods have in common is that of which of the available auxiliary variables to use. In the case of poststratification it must be decided what strata are defined. In the case of propensity score weighting adjustment cells must be formed that have comparable response probabilities.

In this article we propose a selection strategy of weighting variables. The strategy simultaneously accounts for the relation with response behaviour and the relation with the important survey questions. The selection strategy aims at the minimisation of the absolute bias under the assumption that the slope parameter in the linear regression estimator is approximately preserved.

The selection strategy is applied to the Integrated Survey on Household Living Conditions (POLS) 1998.

Key words: Bias; nonresponse adjustment; weighting model; missing-at-random; generalised regression estimator; linear weighting.

1. Introduction

Nonresponse can affect the quality of estimates if nonrespondents are different from respondents in respect of the topics of a survey. This threat may be serious if the size of the nonresponse is large relative to the sample size. In the Netherlands response rates are often rather low. Adjustment methods for potentially selective response, therefore, play an important role in improving the quality of population estimates.

Adjustment methods are based on auxiliary information from population databases, censuses and registers. This information may be available on the population level or on the individual level. In the case of the former, the distribution of for instance age, gender, and marital status is known for the target population of the survey. More ideally, however, the auxiliary information is available on a personal or household level and can be linked directly to the sample. In this article we assume the latter situation, i.e., auxiliary variables can be linked to both respondents and nonrespondents.

Under nonresponse we may distinguish three groups of variables, namely the survey questions of interest, the auxiliary variables that are linked from external sources, and the response indicator. The response indicator is a $0-1$-variable indicating whether a sampled

[1] Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, The Netherlands. Email: bstn@cbs.nl

person responded or not and stands by itself. In the survey literature a lot of research has been devoted to the relation between the response indicator and auxiliary variables. For an introduction see Groves, Dillman, Eltinge, and Little (2002). Furthermore, it has been known for a long time that weighting methods using auxiliary variables that are correlated with the important survey questions may considerably reduce the variance of estimators. As a consequence also the relation between survey questions and auxiliary variables has been analysed. The only relation, however, that is not and cannot be investigated directly is the relation between the response indicator and the target variables of the survey.

In the literature various adjustment methods are given that incorporate auxiliary variables. For a recent overview and a comparison of methods, see Kalton and Flores-Cervantes (2003) and Rocco, Salvati, and Pratesi (2004). An estimator that is often used is the generalised regression estimator modified to nonresponse (see Bethlehem 1988). When using only crossings of categorical auxiliary variables this estimator reduces to poststratification. The population is divided into a number of subpopulations, the so-called strata, and the missing answers of the nonrespondents are predicted by the "average" answers of the respondents in the same stratum. Another method that is often used is propensity score weighting. This technique was introduced by Rosenbaum and Rubin (1983) in the setting of studies of causal effects. In the nonresponse setting the propensity score is the response probability, i.e., the probability that a person or household responds if selected in the sample. The response probability is usually fitted by means of a logit or probit model. The answers by the respondents are weighted by the inverse of their estimated response probabilities.

One problem that many adjustment methods have in common is the selection of informative auxiliary variables. In the case of poststratification it must be decided how to choose strata. In the case of propensity score weighting, groups must be formed that have comparable response probabilities in order to keep a low variance. Since there are three groups of variables involved, the choice of strata or response groups is often performed in two steps. In the first step auxiliary variables are selected that explain the response indicator. In the second step a further selection of variables is performed in which variables are chosen that also relate to the important target variables of the survey. Little (1986) proposes the formation of so-called adjustment cells by modelling the response probability, forming response groups and clustering response groups based on the differences between the "average" answers to the survey questions (see also Rosenbaum and Rubin 1984, Ekholm and Laaksonen 1991 and Czajka et al. 1992). Eltinge and Yansaneh (1997) compare several criteria for the formation of adjustment cells. Geuzinge, Van Rooijen, and Bakker (2000) propose using the product of the correlation between the response indicator and the auxiliary variables and the correlation between a target variable and the auxiliary variables as a measure of the relevance of auxiliary variables in a weighting model.

Crucial in the adjustment for nonresponse are the assumptions that are made about the nonresponse or missing data mechanism. The nonresponse mechanism is called Missing Completely at Random (MCAR) whenever the probability of response is independent of the survey questions. If the probability of response is independent of the survey questions when conditioned on a set of auxiliary variables, the mechanism is called Missing at Random (MAR). For most surveys the MCAR assumption does not hold for the auxiliary variables. In practice it is usually assumed that the nonresponse can be made "sufficiently" MAR by incorporation of the available auxiliary variables in a weighting model.

In this article we use the generalised regression estimator to adjust for nonresponse and we select auxiliary variables by minimising the maximal absolute bias under a weaker assumption than MAR.

As a criterion we use minimisation of the maximal absolute bias of an estimator and we employ linear weighting to recover sample means. In the analysis no population totals or means are used. Thus the response is calibrated to the sample for the auxiliary variables that are selected in the weighting model. Additional weights may be used to calibrate the sample to the population in a second step. However, we did not perform this additional step. We concentrated on the bias because we believe that nonresponse affects especially the location of means and not their variation. We believe that variance reduction is most effective in the calibration from sample to population.

The absolute bias falls within an interval of known form but unknown size. The proposed criterion favours one estimator over another estimator if it corresponds to a smaller interval.

We propose a selection strategy of weighting variables that minimises the width of the bias interval. The strategy is a forward inclusion-backward elimination algorithm of auxiliary variables similar to algorithms applied in regression analysis. Variables are included or eliminated based on jackknife estimates of the mean and standard deviation of the change in interval width. We apply the selection strategy to the 1998 Integrated Survey on Household Living Conditions abbreviated POLS (*Permanent Onderzoek Leefsituatie* in Dutch).

In Section 2 we first give some theoretical background and introduce the selection strategy in more detail. In Section 3 we give results for POLS 1998. Finally, in Section 4 we discuss the outcomes.

2. Estimators for the Sample Mean

We want to analyse the effects of nonresponse and correct for bias due to selective nonresponse. Since nonresponse takes place after a sample is selected, we focus on the estimation of the sample mean. In this section we discuss estimators for the sample mean. In Section 2.1 we introduce notation and make some basic assumptions. Next in Section 2.2 we describe the generalised regression estimator.

2.1. Notation and Basic Assumptions

In the following we distinguish stochastic variables from their realisations by using upper-case and lower-case letters, i.e., the realisation of a stochastic variable Z is denoted by z. We let μ_Z and σ_Z be, respectively, the expectation and standard deviation of variable Z, and $c(Z_1,Z_2)$ and $\gamma(Z_1,Z_2)$ correspond to, respectively, the covariance and correlation between variables Z_1 and Z_2.

We adopt a superpopulation model, i.e., we assume that the survey yields an independent, identically distributed sample of some unknown distribution. We want to estimate the expectation of this unknown distribution. Due to nonresponse part of the sample data is missing, however.

Let the sample in a survey consist of n units labelled 1 to n. In the following we use index i when we refer to unit i in the sample. Let R_i represent the 0-1-indicator for response, $X_i = (X_{1,i}, X_{2,i}, \ldots, X_{m,i})'$ be a set of m background characteristics and Y_i be a survey question. We assume that the background characteristics are available for both respondents ($R_i = 1$) and nonrespondents ($R_i = 0$). Furthermore, the auxiliary variables and survey questions are assumed to be complete, i.e., assumed to contain no missing data.

We assume that the triplets $(R_i, X_i, Y_i)_{1 \leq i \leq n}$ are independent and identically distributed. After the survey is conducted we have realisations (r_i, x_i, y_i) of (R_i, X_i, Y_i) for all respondents. For the nonrespondents we only have the realisations r_i and x_i of R_i and X_i, respectively. We denote the available data after a survey by $(R_i, X_i, R_i Y_i)_{1 \leq i \leq n}$. We let p_R be the probability of response, i.e., $P[R_i = 1]$.

Our interest in this article is in the estimation of the expectation of the survey question Y_i, i.e., μ_Y, using the available data $(R_i, X_i, R_i Y_i)_{1 \leq i \leq n}$.

The sample and response mean vectors of X_i are denoted by $\overline{X}$ and $\overline{X}^*$, respectively. Here, $\overline{X}^*$ stands for

$$\overline{X}^* = \frac{\sum_{i=1}^{n} X_i R_i}{\sum_{i=1}^{n} R_i}$$

Equivalently, we let $\overline{Y}$ and $\overline{Y}^*$ represent the sample and response means of the survey question. The sample and response covariance matrices of X_i are given by S_X^2 and S_X^{*2}. For Y_i the sample and response variance are denoted by S_Y^2 and S_Y^{*2}. We use $C(X, Y)$ as notation for the vector of sample covariances between the auxiliary variables X_i and the survey question Y_i, while $C(X, R)$ and $C(Y, R)$ represent the sample covariances between the background characteristics and the response indicator and the sample covariances between the survey question and the response indicator, respectively. Again we use an asterisk as additional index to indicate covariances based on the response only. Finally, we denote sample correlations by Γ. For example $\Gamma(X, Y)$ is the vector of sample correlations between the background characteristics and the survey question.

We assume, furthermore, that the expectation of Y_i given the realisation $X_i = x_i$ can be described by a linear combination of x_i, i.e.,

$$E(Y_i | X_i = x_i) = \alpha + \beta' x_i \tag{1}$$

where α, the intercept, and $\beta = (\beta_1, \ldots, \beta_m)'$, the slope vector, are unknown constants. Also, we suppose that $c(Y_i - \alpha - \beta' X_i, X_i) = 0$, or in other words that the error term is orthogonal to the background statistics.

We assume that the set of background characteristics is linearly independent, i.e., that there exists no constant vector $\lambda = (\lambda_1, \ldots \lambda_m)'$ such that $\lambda' X_i = 0$. If such a λ should exist, then one of the auxiliary variables is a linear combination of the others. If the auxiliary variables X_i are linearly dependent, then it also holds that $E(Y_i | X_i = x_i) = \alpha + (\beta - j\lambda)' x_i$ for all $j \in \mathbf{Z}$. Hence, the slope parameter would not be unique.

Finally, we assume that the set of auxiliary variables does not contain a constant neither explicitly nor implicitly. This means there does not exist a constant vector $\kappa = (\kappa_1, \ldots, \kappa_m)'$ such that $\kappa' X_i = 1$. If the auxiliary variables contain a constant vector or sum up to one, then it follows that $E(Y_i | X_i = x_i) = (\alpha + j) + (\beta - j\kappa)' x_i$ for all $j \in \mathbf{Z}$. Again, the parameters will not be unique.

2.2. *The Generalised Regression Estimator for the Sample Mean*

An estimator of μ_Y that does not make use of the available background characteristics is the response mean $\overline{Y}*$. When using the response mean as an estimator, the answers of the nonrespondents are predicted by the average answer of the respondents.

We may, however, use the available background characteristics to predict the answers of the nonrespondents. Based on the sample we can estimate α and β in (1) using ordinary least squares

$$\begin{aligned} \hat{\alpha} &= \overline{Y} - \hat{\beta}'\overline{X} \\ \hat{\beta} &= \left(S_X^2\right)^{-1} C(X, Y) \end{aligned} \tag{2}$$

However, since we do not know $\overline{Y}$ and $C(X, Y)$, we must use the response-based estimators

$$\begin{aligned} \hat{\alpha}^* &= \overline{Y}^* - \hat{\beta}^{*\prime}\overline{X}^* \\ \hat{\beta}^* &= \left(S_X^{*2}\right)^{-1} C^*(X, Y) \end{aligned} \tag{3}$$

Instead of using the response mean as a prediction, we may now predict the answers of the nonrespondents using their auxiliary variables and the estimated regression parameters in (3). This is the generalised regression estimator and it has the form

$$\begin{aligned} \overline{Y}^*_{gr} &= \frac{1}{n}\left(\sum_{i=1}^{n} R_i Y_i + \sum_{i=1}^{n} (1 - R_i)(\hat{\alpha}^* + \hat{\beta}^{*\prime} X_i)\right) \\ &= \overline{RY}^* + \frac{1}{n}\sum_{i=1}^{n} (1 - R_i)(\overline{Y}^* - \hat{\beta}^{*\prime}\overline{X}^* + \hat{\beta}^{*\prime} X_i) \\ &= \overline{RY}^* + \overline{Y}^* - \hat{\beta}^{*\prime}\overline{X}^* + \hat{\beta}^{*\prime}\overline{X} - \frac{1}{n}\sum_{i=1}^{n} R_i(\overline{Y}^* - \hat{\beta}^{*\prime}\overline{X}^* + \hat{\beta}^{*\prime} X_i) \\ &= \overline{RY}^* + \overline{Y}^* - \hat{\beta}^{*\prime}\overline{X}^* + \hat{\beta}^{*\prime}\overline{X} - \overline{RY}^* + \overline{R}\hat{\beta}^{*\prime}\overline{X}^* - \overline{R}\hat{\beta}^{*\prime}\overline{X}^* \\ &= \overline{Y}^* + \hat{\beta}^{*\prime}(\overline{X} - \overline{X}^*) \end{aligned} \tag{4}$$

Note that in (4) we use sample means of the auxiliary variables and not population means. The auxiliary information may thus also include fieldwork characteristics of the survey, e.g., the interviewer district or the interviewer experience.

Instead of sample means we may also use population means in (4). The use of population means, however, serves variance reduction but it does not affect the bias. In this article we are primarily interested in bias. We therefore calibrate to sample means. Calibration to population means can be performed in an additional weighting step.

3. Bias and Assumptions About the Nonresponse Mechanism

In the literature various assumptions about the nonresponse mechanism are discussed. These assumptions have different implications for the adjustment of nonresponse (see for instance Little and Rubin 2002).

If the answer to a survey question is independent of the response to the survey, then the nonresponse mechanism is called Missing-Completely-at-Random (MCAR). This is a strong assumption which implies that response means are unbiased. If the answer to a survey question is independent of the response to the survey conditionally on a set of auxiliary variables, then the nonresponse mechanism is called Missing-at-Random (MAR). The MAR assumption is weaker than MCAR and the auxiliary variables for which the nonresponse mechanism is MAR can be used to construct unbiased estimators. Assumptions that are weaker than MAR are called Not-Missing-at-Random (NMAR).

However, in most cases no explicit assumption is made about the nonresponse mechanism. By applying a particular adjustment method for a survey it is implicitly assumed that this method yields unbiased estimators.

In the next sections we elucidate assumptions about the nonresponse mechanism and discuss the consequences for the bias of the estimators of Section 2.

3.1. Response Propensities

Assumptions about the nonresponse mechanism are often stated with respect to the so-called response propensities. We define a response propensity as follows:

$$\rho_{x,y} = P[R_i = 1|X_i = x, Y_i = y] \tag{5}$$

i.e., the probability of response given the background characteristics and the answer to the survey question. Note that the response propensity is only defined for values of x and y with a positive (joint) probability density.

By $\rho_{X,Y}$ we denote the stochastic variable that represents the conditional probability of response:

$$\rho_{X,Y} = P[R_i = 1|X_i, Y_i] \tag{6}$$

The MCAR assumption means that $\rho_{X,Y}$ is deterministic, or in other words $\rho_{X,Y}$ is a constant and $\rho_{x,y} = p_R$, for all values of x and y. The MAR assumption implies that $\rho_{X,Y}$ is stochastic, however, only in the background characteristics. For fixed x the response propensities are constant in y. Under MAR we omit the index for Y, i.e., we use ρ_X instead of $\rho_{X,Y}$.

The covariance $c(Y, R)$ between Y and R equals

$$\begin{aligned} c(Y,R) &= E(YR) - \mu_Y p_R \\ &= E(E(\rho_{X,Y} Y|X)) - \mu_Y p_R \end{aligned} \tag{7}$$

where the first expectation is over X and the second expectation is over Y conditionally on X.

Hence, from (7) it can be seen that the MCAR assumption leads to noncorrelated Y and R. If the MAR assumption holds, then (7) reduces to

$$\begin{aligned} c(Y,R) &= E(\rho_X E(Y|X)) - \mu_Y p_R \\ &= E(\rho_X (E(Y|X) - \mu_Y)) \end{aligned} \tag{8}$$

Under the linear model defined in (1) the covariance in (8) becomes

$$\begin{aligned} c(Y,R) &= E(\rho_X(\alpha + \beta' X - \alpha - \beta' EX)) \\ &= E(\rho_X \beta'(X - EX)) \\ &= c(\beta' X, R) \end{aligned} \tag{9}$$

3.2. Bias of the Response Mean

It is not difficult to show that under the condition that at least one individual did respond in the survey, i.e., $\sum_{i=1}^{n} R_i > 0$, the bias of the response mean equals

$$\begin{aligned} B(\overline{Y}^*) &= E(\overline{Y}^*) - \mu_Y \\ &= \frac{c(Y,R)}{p_R} \\ &= \gamma(Y,R)\sigma_Y \frac{\sigma_R}{p_R} \\ &= \gamma(Y,R)\sigma_Y \sqrt{\frac{1-p_R}{p_R}} \end{aligned} \tag{10}$$

where in the last step we use that $\sigma_R = \sqrt{p_R(1-p_R)}$. From (10) it follows that the bias of the response mean is zero if either the probability of response is one or the response to the survey and the answer to the survey question are uncorrelated. The latter holds under the MCAR assumption.

Let for the moment X_i be one-dimensional, i.e., $m = 1$. For independent, identically distributed triplets $(R_i, X_i, Y_i)_{1 \le i \le n}$ with finite variances it can be proved that $\gamma(Y,R)$ can be bounded using $\gamma(X,R)$ and $\gamma(X,Y)$ in the following way:

$$\begin{aligned} \gamma(X,Y)\gamma(X,R) - \sqrt{1-\gamma^2(X,Y)}\sqrt{1-\gamma^2(X,R)} &\le \gamma(Y,R) \\ &\le \gamma(X,Y)\gamma(X,R) + \sqrt{1-\gamma^2(X,Y)}\sqrt{1-\gamma^2(X,R)} \end{aligned} \tag{11}$$

The bounds in (11) can be proved to hold by using the fact that a covariance matrix is nonnegative definite (see for instance Strang 1986). The bounds are also sharp. For every value in interval (11) a joint probability distribution for (R_i, X_i, Y_i) can be formulated in such a way that $\gamma(Y,R)$ equals this value, while the joint probability distribution of the observed variables, $(R_i, X_i, R_i Y_i)_{1 \le i \le n}$, remains fixed. In other words we cannot distinguish between those joint probability distributions based on the observations alone.

Combining (10) with (11) we arrive at the following interval for the bias of the response mean:

$$\begin{aligned}&\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\Big(\gamma(X,Y)\gamma(X,R)-\sqrt{1-\gamma^2(X,Y)}\sqrt{1-\gamma^2(X,R)}\Big)\leq B(\overline{Y}^*)\\&\quad\leq\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\Big(\gamma(X,Y)\gamma(X,R)+\sqrt{1-\gamma^2(X,Y)}\sqrt{1-\gamma^2(X,R)}\Big)\end{aligned}\qquad(12)$$

The midpoint of interval (12) is $\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\gamma(X,Y)\gamma(X,R)$ and its width is

$$2\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\sqrt{1-\gamma^2(X,Y)}\sqrt{1-\gamma^2(X,R)}\qquad(13)$$

Hence, any auxiliary variable *X* defines an interval for the bias of the response mean for survey question *Y*. Since (12) is true for any auxiliary variable we may take the intersection over the intervals for various auxiliary variables. Let us, however, search for a composite auxiliary variable that minimises the width (13).

3.3. Bias of the Generalised Regression Estimator for the Sample Mean

Let us first suppose that we know the true vector of slope parameters β defined in Section 2.1, and that we estimate the sample mean of survey question *Y* using the estimator

$$\overline{Y}_{gr}=\overline{Y}^*+\beta'(\overline{X}-\overline{X}^*)\qquad(14)$$

The bias of (14) equals

$$\begin{aligned}B(\overline{Y}_{gr})&=E(\overline{Y}^*)-\mu_Y+E\beta'(\overline{X}-\overline{X}^*)\\&=B(\overline{Y}^*)-\beta'B(\overline{X}^*)\\&=B(\overline{Y}^*)-B(\beta'\overline{X}^*)\end{aligned}\qquad(15)$$

Hence, the bias of (14) reduces to the bias of the response mean of *Y* minus the bias of the slope vector times the response mean of *X*. Using (7) and (10) we get

$$\begin{aligned}B(\overline{Y}_{gr})&=\frac{c(Y,R)}{p_R}-\frac{c(\beta'X,R)}{p_R}\\&=\frac{c(Y,R)}{p_R}-\frac{c(\beta'X,Y)c(\beta'X,R)}{\sigma^2_{\beta'X}p_R}\\&=(\gamma(Y,R)-\gamma(\beta'X,Y)\gamma(\beta'X,R))\sqrt{\frac{1-p_R}{p_R}}\sigma_Y\end{aligned}\qquad(16)$$

In the first step of (16) we make use of representation (1) and we assume that the error term in the linear regression is orthogonal to the background characteristics, i.e., $c(Y,\beta'X)=c(\alpha+\beta'X,\beta'X)=\sigma^2_{\beta'X}$

If we assume that the response mechanism is MAR, then from (9) we see that regression estimator (14) is unbiased.

If we do not assume MAR, then the bias of estimator $\overline{Y}_{gr}$ is the same as the bias of the response mean in (10) except for a shift of size $\gamma(\beta' X, Y)\gamma(\beta' X, R)\sqrt{(1-p_R)/p_R}\,\sigma_Y$. This shift equals the midpoint of (12) if we take $\beta' X$ as (composite) auxiliary variable in (11). However, since the interval for $\gamma(Y, R)$ in (11) is constructed while fixing the other two covariances, we have

$$\begin{aligned}
-\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\sqrt{1-\gamma^2(\beta' X, Y)}\sqrt{1-\gamma^2(\beta' X, R)} &\le B(\overline{Y}_{gr}) \\
\le \sigma_Y\sqrt{\frac{1-p_R}{p_R}}\sqrt{1-\gamma^2(\beta' X, Y)}\sqrt{1-\gamma^2(\beta' X, R)} &
\end{aligned} \tag{17}$$

and, hence, the width of the bias interval is the same as that of the response mean and equals

$$2\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\sqrt{1-\gamma^2(\beta' X, Y)}\sqrt{1-\gamma^2(\beta' X, R)} \tag{18}$$

So, even if we know the true slope vector, (14) does not produce a bias interval that is smaller than that of the response mean.

The bias of the response-based regression estimator $\overline{Y}^*_{gr}$ defined in (4) equals

$$\begin{aligned}
B\left(\overline{Y}^*_{gr}\right) &= B(\overline{Y}^*) + E(\hat{\beta}^{*\prime}(\overline{X} - \overline{X}^*)) \\
&= B(\overline{Y}^*) - E(\hat{\beta}^{*\prime}(\overline{X}^* - \overline{X})) \\
&= \frac{c(Y, R)}{p_R} - E(\hat{\beta}^{*\prime}(\overline{X}^* - \overline{X}))
\end{aligned} \tag{19}$$

If we assume MAR, then using (9) and (10) the bias in (19) can be rewritten as

$$\begin{aligned}
B\left(\overline{Y}^*_{gr}\right) &= \frac{c(\beta' X, R)}{p_R} - E(\hat{\beta}^{*\prime}(\overline{X}^* - \overline{X})) \\
&= \beta' B(\overline{X}^*) - E(\hat{\beta}^{*\prime}(\overline{X}^* - \overline{X})) \\
&= -E(\hat{\beta}^* - \beta)'(\overline{X}^* - \overline{X})
\end{aligned} \tag{20}$$

Under MAR the response indicators and the answers to the survey questions are independent conditionally on the auxiliary variables. Consequently, it can be shown that (20) is equal to zero under the assumption that there are at least two respondents. This can be shown to hold by conditioning on the auxiliary variables $\{X_i\}_{1\le i\le n}$ so that terms concerning the $\{Y_i\}_{1\le i\le n}$ and $\{R_i\}_{1\le i\le n}$ are independent and expectations can be derived separately. Hence, estimator $\overline{Y}^*_{gr}$ is unbiased if the response mechanism is MAR.

The bias of $\overline{Y}^*_{gr}$ cannot easily be simplified in the general case. However, earlier it was concluded that even if we use the true slope parameter the regression estimator leads to a bias interval of the same width as that of the response mean. We conjecture, therefore, that the bias interval of the regression estimator using the response-based slope parameter cannot be smaller in general as it makes use of less information. This conjecture forms the basis for our selection strategy in Section 4. In the next section it is explained why a missing data assumption is made which is weaker than MAR.

3.4. A Not-Missing-at-Random Assumption

In Section 3.3 it was shown that the bias interval of the regression estimator using the true slope parameter β has the same width as the bias interval of the response mean. If we assume MAR then the bias of the regression estimator is zero while the bias of the response mean may not be zero.

In this article we do not assume MAR for three reasons. First, it is our experience that when new auxiliary variables become available through registers the explanatory power of models for the nonresponse mechanism and the key survey questions often increases significantly. In Section 5 we employ a number of auxiliary variables that have recently become available at Statistics Netherlands and we find that adding those variables leads to different estimates and an increased explained variance. If we make the MAR assumption then we should believe that we already have sufficient background information to make respondents resemble nonrespondents. However, in that case there would be no need to search for auxiliary variables that better explain the differences between respondents and nonrespondents. We can never preclude that auxiliary variables will become available in the future that indicate that there is still bias left.

Secondly, even if we accept that we have sufficient background information, we still need to find a criterion in order to construct weighting models. In practice it is not usually possible to use the full model of auxiliary variables without letting the variance of the estimators become very large. Hence, in setting up a weighting model one has to choose which auxiliary variables to add to the model and which to omit. This choice is not an easy one since one needs to account simultaneously for the relation between nonresponse and background characteristics and between survey questions and background characteristics. Little (1986), for instance, proposes forming adjustment cells by merging strata with respondents that give similar answers. The width (18), however, gives an easy criterion for constructing and comparing weighting models directly.

Finally, when adding auxiliary variables in weighting models to adjust for nonresponse, it can often be observed that the estimates move in one direction. Each time a variable is added the estimate shifts further away from the response mean but in the same direction. This conforms to the idea that there is some background characteristic that separates respondents from nonrespondents when it comes to the survey question under investigation and that we seem to grasp better but never to its full extent.

4. A Selection Strategy

In this section we form a selection strategy based on the findings in the previous section. In Section 4.1 we first derive a criterion for the comparison of weighting models. Next, in Section 4.2 we propose an algorithm in order to minimise this criterion and to build weighting models for cases where all auxiliary variables are categorical.

4.1. The Selection Criterion

We return to the bias of the generalised regression estimator. In Section 3.3 it was shown that the maximal absolute bias of the generalised regression estimator using the true regression parameter β is the width of bias interval (18)

$$2\sigma_Y\sqrt{\frac{1-p_R}{p_R}}\sqrt{1-\gamma^2(\beta'X,Y)}\sqrt{1-\gamma^2(\beta'X,R)} \tag{21}$$

Let us search for the composite vector of auxiliary variables that minimises (21). Since the first part of (21), $2\sigma_Y\sqrt{(1-p_R)/p_R}$, is independent of the choice of auxiliary variables, it suffices to minimise

$$w(X)=\sqrt{1-\gamma^2(\beta'X,Y)}\sqrt{1-\gamma^2(\beta'X,R)} \tag{22}$$

The criterion given by (22) cannot be computed unless we know the true regression parameter β and the true correlations between the composite variable $\beta'X$ and the response indicators R and between the composite variable $\beta'X$ and the survey question Y.

Clearly, we know neither the regression parameter nor the correlations. We must, therefore, rely on estimators based on the response and the sample. Let $w^*(X)$ be the estimator for $w(X)$ based on the response slope vector $\hat{\beta}^*$ and on estimated correlations

$$w^*(X)=\sqrt{1-\Gamma^2(\hat{\beta}^{*\prime}X,R)}\sqrt{1-(\Gamma^*(\hat{\beta}^{*\prime}X,Y))^2} \tag{23}$$

In (23) we estimate the correlation between $\beta'X$ and R by the sample correlation between $\hat{\beta}^{*\prime}X$ and the response indicator R. The correlation between $\beta'X$ and Y is estimated by the response correlation between $\hat{\beta}^{*\prime}X$ and Y.

We computed deviances between $w^*(X)$ and $w(X)$ for a number of auxiliary variables that we artificially treated as survey answers. For these variables we found only small deviances.

Note that for the computation of (23) it is sufficient to estimate the covariance matrix of Y, R and X. Hence, it is not necessary to have a gross sample file linked to administrative data. It would suffice to have sample totals. This implies that the selection criterion can also be applied to calibrate the response directly to population totals. In that case the response indicator is one if a unit was sampled and did respond.

4.2. The Forward-backward Algorithm

We set up an algorithm for the case where all auxiliary variables are categorical. The generalised regression estimator then reduces to multiway stratification (see Bethlehem and Kersten 1985).

We first introduce some additional notation. Let the available auxiliary variables be labelled 1 to m. The set $\mathrm{M}=\{1,2,\ldots,m\}$ represents the labels of auxiliary variables that are available for weighting. For any two subsets $\mathrm{M}_1 \subseteq \mathrm{M}$ and $\mathrm{M}_2 \subseteq \mathrm{M}$ we let $\Delta w^*(\mathrm{M}_1, \mathrm{M}_2)$ denote the difference of the widths (23) between the two models, i.e.,

$$\Delta w^*(\mathrm{M}_1, \mathrm{M}_2) = w^*(\{X_l\}_{l \in \mathrm{M}_1}) - w^*(\{X_l\}_{l \in \mathrm{M}_2}) \tag{24}$$

Furthermore, $\hat{s}_{\Delta w^*(\mathrm{M}_1,\mathrm{M}_2)}$ will represent the jackknife estimator (see Miller 1974), for the standard deviation of (24) and $\xi_{1-\alpha}$ is the $100(1 - \alpha)\%$-quantile of the standard normal distribution. For convenience, we omit the two models in the index of the estimator $\hat{s}_{\Delta w^*(\mathrm{M}_1,\mathrm{M}_2)}$, since it will be clear from the context what models are taken. Finally, the "empty" model is denoted by ϕ. Note that $w^*(\phi) = 1$.

We propose to use the following forward-backward selection algorithm that is similar to stepwise regression with forward inclusion and backward elimination:

- Take the auxiliary variable k for which $\frac{\Delta w^*(\phi,\{k\})}{\hat{s}_{\Delta w^*}}$ is largest for all the auxiliary variables, but only if $\frac{\Delta w^*(\phi,\{k\})}{\hat{s}_{\Delta w^*}} > \xi_{1-\alpha}$. Set $i = 1$ and let $\mathrm{M}_1 = \{k\}$. If a variable is added to the empty model go to Step 2, otherwise go to Step 4.
- Add the auxiliary variable l for which $\frac{\Delta w^*(\mathrm{M}_i, M_i \cup \{l\})}{\hat{s}_{\Delta w^*}}$ is largest for all the remaining auxiliary variables, but only if $\frac{\Delta w^*(\mathrm{M}_i, M_i \cup \{l\})}{\hat{s}_{\Delta w^*}} > \xi_{1-\alpha}$, and let $\tilde{\mathrm{M}}_{i+1} = \mathrm{M}_i \cup \{l\}$. Otherwise, let $\tilde{\mathrm{M}}_{i+1} = \mathrm{M}_i$.
- Remove auxiliary variable $m \in \mathrm{M}_i$ for which $\frac{\Delta w^*(\tilde{\mathrm{M}}_{i+1} \backslash \{m\}, \tilde{\mathrm{M}}_{i+1})}{\hat{s}_{\Delta w^*}}$ is smallest, but only if $\frac{\Delta w^*(\tilde{\mathrm{M}}_{i+1} \backslash \{m\}, \tilde{\mathrm{M}}_{i+1})}{\hat{s}_{\Delta w^*}} < \xi_{1-\alpha}$, and let $\mathrm{M}_{i+1} = \tilde{\mathrm{M}}_{i+1} \backslash \{m\}$. Otherwise, let $\mathrm{M}_{i+1} = \tilde{\mathrm{M}}_{i+1}$.
- If no auxiliary variable was added or removed then stop, otherwise repeat from Step 2 with $i := i + 1$.

The proposed selection strategy starts with a simple model with only one weighting variable, namely the variable that minimises (23). In the following steps variables are iteratively added and removed. Variables are only added or removed if the difference in width (24) is larger than $\xi_{1-\alpha}$ estimated standard deviations. The significance level α may be chosen differently for the addition and removal step.

It can be shown that the algorithm cannot retrace its own steps if there are a finite number of available auxiliary variables. This implies that the algorithm stops after a finite number of steps.

However, the algorithm does not necessarily converge to the optimal subset of auxiliary variables. We found some examples where subsets other than the subsets found by the algorithm led to smaller bias intervals. Differences were very small and the composition of these subsets was very much the same. It is important to stress that the number of possible subsets of auxiliary variables equals 2^m. In the examples we investigated the algorithm needed at most six iterations.

5. Results

Here we apply the proposed selection strategy to the 1998 Integrated Survey on Household Living Conditions (*Permanent Onderzoek Leef Situatie* in Dutch, or POLS). It is a large continuing survey with questions about issues like health, social participation, justice and recreational activities. For a detailed description of POLS we refer to Vousten and De Heer (1998).

The survey is modular and consists of a base questionnaire and a number of questionnaires that deal with separate topics. The base questionnaire is to be filled in by all persons. However, each person only fills in one topical questionnaire. The base questionnaire contains general questions and a number of basic questions that are used for allocation of the topical questionnaires. These basic questions are also used in weighting models for the topical questions. Here, we will focus on questions from the base questionnaire.

The survey uses a two-stage sample, in which the clusters in the first stage are formed by municipalities. From the clusters simple random samples without replacement are drawn consisting of persons. The first-order inclusion probabilities differ only for age. All persons 12 years and older have the same probability ending up in the sample. In this article we consider all persons 12 years and older and omit only the nonresponse due to frame errors. The sample then consists of 36,136 persons.

The 1998 POLS had a fieldwork period of two months. The first month is CAPI, and the second month is a mixture of CAPI and CATI. After two months the size of the response was 21,571 persons, i.e., a response rate close to 60%.

From the POLS 1998 survey we selected two survey questions, Owner of a house (yes or no) and Owner of a PC or laptop (yes or no), and one auxiliary variable, Receiving some form of social allowance (yes or no). We treat the last variable as if it were a survey question, i.e., we omit the nonrespondents.

In the forward-backward algorithm the following auxiliary variables are used: (A) gender (male or female), (B) age, (C) marital status (not married, married, divorced or widowed), (D) ethnic group (native, Moroccan, Turkish, Surinam, Netherlands Antilles/Aruba, other nonwestern nonnative or other western nonnative), (E) ethnic generation (native, nonnative 1st generation, nonnative 2nd generation one parent or 2nd generation two parents), (F) having a job (yes or no), (G) province of residence, (H) region in the Netherlands (north, east, west, or south), (I) children in the household (yes or no), (J) household type (single, couple, couple with children, single parent or other), (K) household size (1, 2, 3, 4, or >4), (L) degree of urbanisation (5 levels), (M) size of town (8 levels), (N) interviewer district (27 districts), (O) having a listed telephone number (yes or no), (P) average value of houses in 6-digit postal code area, (Q) percentage of nonnatives in 6-digit postal code area, and (R) receiving a social allowance (yes or no). Furthermore, we crossed age and marital status into a new variable (S), age x marital status, in which some of the categories are clustered. In the following we will refer to the labels (A), (B). . .(S) for convenience. In the case of the variable receiving a form of social allowance we did not use (R), of course.

Next we illustrate the selection strategy for the selected variables. Tables 1 to 3 describe the selection process for these variables. The final weighting models are depicted by bold letters. We use the jackknife method with group size 100 to estimate the standard deviations in the selection strategy and take $\alpha = 0.01$ as the significance level for both additions and removals.

In the first instance in the selection of weighting variables we also crossed auxiliary variables, but we found in all the investigated cases that the estimates based on weighting models including interaction effects differ at most 0.1% from the estimates based on

Table 1. Results of the selection strategy for ownership of a house. Also given are the correlations and the value of the selection criterion

Model	$\overline{Y}^*_{gr}$	$\Gamma^*_{\hat{\beta}^{*\prime}X,Y}$	$\Gamma_{\hat{\beta}X,R}$	w^*
ϕ (Empty)	0.633	–	–	1
P	0.612	0.47	0.11	0.875
P + J	0.605	0.52	0.13	0.849
P + J + Q	0.598	0.54	0.15	0.829
J + Q	0.602	0.44	0.16	0.886
P + J + Q + G	0.594	0.56	0.16	0.820
J + Q + G	0.599	0.45	0.17	0.881
P + Q + G	0.599	0.53	0.15	0.841
P + J + Q + G + S	0.593	0.57	0.16	0.810
J + Q + G + S	0.598	0.47	0.17	0.869
P + Q + G + S	0.594	0.56	0.16	0.816
P + J + G + S	0.596	0.55	0.16	0.825
P + J + Q + G + B	0.594	0.57	0.16	0.813
P + J + Q + G + C	0.594	0.56	0.16	0.819

weighting models that incorporate only the main effects. This important finding led to the decision not to model interaction effects at all. Consequently, in Tables 1–3 only models with main effects are given. It must be noted, however, that crossing auxiliary variables does narrow the interval for the bias in general. The only variables that we did cross were age and marital status. In Tables 1 to 3, if age x marital status (S) is selected, we consider

Table 2. Results of the selection strategy for ownership of a PC. Also given are the correlations and the value of the selection criterion

Model	$\overline{Y}^*_{gr}$	$\Gamma^*_{\hat{\beta}^{*\prime}X,Y}$	$\Gamma_{\hat{\beta}X,R}$	w^*
ϕ (Empty)	0.598	–	–	1
S	0.585	0.50	0.06	0.866
S + P	0.579	0.52	0.09	0.850
C + P	0.586	0.34	0.08	0.938
B + P	0.582	0.51	0.08	0.858
S + P + D	0.574	0.53	0.11	0.842
P + D	0.585	0.22	0.14	0.966
C + P + D	0.583	0.34	0.10	0.934
B + P + D	0.577	0.52	0.10	0.850
S + P + D + K	0.573	0.54	0.11	0.837
P + D + K	0.572	0.42	0.14	0.898
C + P + D + K	0.573	0.45	0.13	0.883
B + P + D + K	0.573	0.53	0.11	0.840
S + D + K	0.575	0.53	0.10	0.846
B + P + D + K + R	0.572	0.54	0.12	0.835
P + D + K + R	0.571	0.42	0.15	0.896
B + D + K + R	0.574	0.53	0.11	0.844
B + P + K + R	0.577	0.53	0.10	0.844

Table 3. Results of the selection strategy for receiving some form of social allowance. Also given are the correlations and the value of the selection criterion

Model	$\overline{Y}^*_{gr}$	$\Gamma^*_{\hat{\beta}^{*\prime}X,Y}$	$\Gamma_{\hat{\beta}X,R}$	w^*
ϕ (Empty)	0.104	–	–	1
S	0.109	0.34	−0.05	0.940
S + P	0.112	0.36	−0.08	0.930
C + P	0.109	0.22	−0.08	0.973
B + P	0.109	0.33	−0.06	0.943
S + P + O	0.114	0.37	−0.11	0.925
P + O	0.111	0.16	−0.16	0.975
C + P + O	0.111	0.23	−0.11	0.968
B + P + O	0.113	0.34	−0.10	0.937

age (B) and marital status (C) in the removal step. So (S) may be replaced by (B) or (C) if the increase in the criterion function is significantly small.

We will explain Table 1, the results of the selection process for ownership of a house. The variable that produces the smallest interval is the average house value in 6-digit postal code area (P). Introduction of this variable gives a considerable reduction of 0.125 of $w^*(X)$. This variable is thus added to the empty weighting model. Next all remaining variables are tested in combination with the average house value. The "best" variable is household type (J), which gives a further reduction of 0.026 of $w^*(X)$. By itself household type gives $w^*(X) = 0.939$, which is much larger than the $w^*(X) = 0.875$ of average house value. In the following step we add a third variable to average house value and household type. It turns out that the percentage of foreigners (Q) is the choice leading to the smallest interval. Variable $w^*(X)$ decreases from 0.849 to 0.829, and it is therefore added to the model. Next, the weighting models with one variable removed are compared to the three-variable model. However, both models lead to an increase of $w^*(X)$ that is not significant at $\alpha = 0.01$ and the variables are not removed. In the fourth and fifth iteration, respectively, the variables province of residence (G) and age x marital status (S) are added. In the fifth iteration the variable age × marital status is replaced by age (B), indicating that marital status does not significantly affect the width of the interval. Finally, in the sixth iteration no auxiliary variables can be found that significantly decrease $w^*(X)$ and the algorithm is stopped. Hence, the weighting model for ownership of a house becomes

average house value + household type + percentage foreign + province + age

The final model for ownership of a PC or laptop is

age + average house value + ethnic group + household size + social allowance

while for receiving some form of social allowance it is

age × marital status + average house value + telephone

The response means of the three variables are, respectively, 63.3%, 59.8% and 10.4%. The regression estimates corresponding to the final models are, respectively, 59.4%, 57.2% and 11.4%. The sample mean of receiving a form of social allowance equals 12.1%.

Tables 1 to 3 are good examples of series of estimates that move in one direction. The estimated proportion of owners of a house moves in steps of −2.1%, −0.7%, −0.7% and −0.4%, a total of −3.9%. Occasionally, however, we also found estimates that moved both up and down for some other survey questions.

It turned out that the resulting weighting model produces an estimate (11.4%) for the proportion of persons that receive some form of social allowance that is closer to the sample mean (12.1%) than that produced by the weighting model currently used for POLS. This is mostly due to the fact that we had a larger set of auxiliary variables available. Of course, it is impossible to know whether the same is true for the survey questions. However, the differences in estimates between the current weighting model and the models following from the selection strategy were close to 1% in some cases.

Clearly, we can only draw strong conclusions when we have some idea of the size of the variance of the estimators. According to the jackknife estimates for the standard deviations of the differences in the criterion function w^* all additions were significant at the 1% level. It must be remarked that it is not at all straightforward to approximate variances under the NMAR assumption. We did approximate variances for some of the models under the stronger MAR assumption using bootstrap methods. These simulations revealed that most standard deviations are smaller than 0.001 (or 0.1%).

We manually applied the algorithm of Section 4.2. In the case of a large number of auxiliary variables, this can be quite cumbersome and time-consuming. The automation of the algorithm, however, is straightforward. The computation of the jackknife estimates may take a couple of minutes in each iteration step as the differences between the regression estimates of the current and proposed models need to be calculated for all subsets.

6. Discussion

The results indicate the usefulness of the proposed strategy for selecting auxiliary variables in weighting models. The selection of auxiliary variables is efficient and economical, because variables are only added when they significantly decrease the width of the bias interval. Furthermore, the algorithm is reasonably fast. If it is programmed and automated, then within 15 minutes a weighting model can be produced containing five auxiliary variables taken out of a set of 20 auxiliary variables and a sample of 36,000 persons. The calculation of jackknife estimates for the standard deviations of difference in criterion function accounts for the major part of the computational complexity.

Another benefit of the strategy is that the construction of strata can be done in one step. This strategy focuses simultaneously on the relation between auxiliary variables and survey questions and between auxiliary variables and response behaviour. Auxiliary variables are only interesting if both relations exist. In this article we used regression estimation as a method to adjust for bias. However, the selection strategy may equally well be used to form cells in propensity score weighting.

For all investigated cases we found that weighting models with interaction terms give estimates very similar to those given by weighting models with only main terms. This means we can substantially reduce the number of parameters in the model without affecting the outcomes and hence incorporate more auxiliary information. Also, this makes the addition and removal of auxiliary variables straightforward. If variables need to be crossed, then the algorithm becomes more complex. The restriction to main effects may not be suitable for all surveys, and must always be checked. A promising technique in this respect may be the classification tree method used in data mining (see Breiman, Friedman, Olshen, and Stone 1998).

In this article we used a gross sample file that was linked to administrative data. Furthermore, we implicitly assumed equal inclusion probabilities. We explicitly focused on the gross sample, since the selection criterion combines both adjustment for selective nonresponse and variance reduction. The proposed criterion, however, only needs estimates of covariances between auxiliary variables, survey answers and the response indicator. To compute the covariances, totals are sufficient. This implies that we can extend the selection criterion to calibrate directly from response to population totals. The response indicator then has a different meaning and equals one if a unit is sampled and responds. We do not need a gross sample but only population totals. If a gross sample file is available but inclusion probabilities are unequal, then an additional weighting step is necessary. The simplest solution would be to use the proposed weighting model to calibrate to population totals with the inclusion probabilities as inverse weights.

There are a number of issues that need to be resolved. First, we need to test other estimation methods for the standard deviations in the selection strategy. Also, it is necessary to investigate to what extent correlations between survey questions and auxiliary variables are affected by nonresponse as well as the selection criterion that we propose.

Furthermore, it is necessary to adapt the strategy to categorical survey questions with more than two categories. In the present form the selection strategy is not directly suited to these questions. The width of the bias interval is a vector if the survey question has more than two categories. If the variable is a nominal or ordinal variable, then correlations do not make much sense and it may be better represented by a vector of dummy variables for each category. However, if there are more categories, the intervals have different sizes in general. Hence, we can only prefer one auxiliary variable to another if we introduce some ordering, for instance by using the maximum or average width over the categories. This will be the topic of further research.

Another important aspect is the assumption underlying the response mechanism. Here we assumed the missing data are Not-Missing-at-Random. For auxiliary variables we have evidence that this assumption is closer to the truth than the usual Missing-at-Random (MAR) assumption. More empirical evidence is needed, however.

Finally, in the future we would also like to investigate whether estimates can be improved by using the composite auxiliary variable that minimises the width of the interval for the bias. In this article we used the parameter vector that follows from regression of the survey question on the auxiliary variables. In general the optimal composition of auxiliary variables will produce smaller intervals and thus a smaller maximal absolute bias.

7. References

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Bethlehem, J.G. and Kersten, H.M.P. (1985). On the Treatment of Nonresponse in Sample Surveys. Journal of Official Statistics, 1, 287–300.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1998). Classification and Regression Trees. Boca Raton: Chapman and Hall.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from Advance Data Using Propensity Modelling: An Application to Income and Tax Statistics. Journal of Business and Economic Statistics, 10, 117–131.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modelling in the Finnish Household Budget Survey. Journal of Official Statistics, 7, 325–337.

Eltinge, J.L. and Yansaneh, I.S. (1997). Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. Survey Methodology, 23, 33–40.

Everaers, P. and van der Laan, P. (2000). The Dutch Virtual Census, E-Proceedings of the 53th Session of the International Statistical Institute. Seoul, Korea.

Geuzinge, L., van Rooijen, J., and Bakker, B.F.M. (2000). The Use of Administrative Registers to Reduce Non-response Bias in Household Surveys. Netherlands Official Statistics, 2, 32–39.

Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (eds) (2002). Survey Nonresponse. New York: John Wiley and Sons.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. Journal of Official Statistics, 19, 81–97.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

Miller, R.G. (1974). The Jackknife – A Review. Biometrika, 61, 1–15.

Rocco, E., Salvati, N., and Pratesi, M. (2004). Participation in CATI Surveys: Traditional Nonresponse Adjustments versus Propensity Score Matching in Reducing the Nonresponse Bias. E-Proceedings of the European Conference on Quality and Methodology in Official Statistics. Mainz, Germany.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association, 79, 516–524.

Strang, G. (1986). Linear Algebra and Its Applications. San Diego: Harcourt Brace Jovanovich.

Vousten, R. and Heer, W. de (1998). Reducing Nonresponse: The POLS Fieldwork Design. Netherlands Official Statistics, 2, 16–19.

Received November 2004
Revised March 2006