

Särndal, Carl-Erik; Traat, Imbi; Lumiste, Kaur

## Article

# INTERACTION BETWEEN DATA COLLECTION AND ESTIMATION PHASES IN SURVEYS WITH NONRESPONSE

Statistics in Transition New Series

## Provided in Cooperation with:

Polish Statistical Association

*Suggested Citation:* Särndal, Carl-Erik; Traat, Imbi; Lumiste, Kaur (2018) : INTERACTION BETWEEN DATA COLLECTION AND ESTIMATION PHASES IN SURVEYS WITH NONRESPONSE, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 19, Iss. 2, pp. 183-200, <https://doi.org/10.21307/stattrans-2018-011>

This Version is available at:

<https://hdl.handle.net/10419/207894>

### Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

### Terms of use:

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

# INTERACTION BETWEEN DATA COLLECTION AND ESTIMATION PHASES IN SURVEYS WITH NONRESPONSE

Carl-Erik Särndal<sup>1</sup>, Imbi Traat<sup>2</sup>, Kaur Lumiste<sup>3</sup>

## ABSTRACT

Inference in surveys with nonresponse has been studied extensively in the literature with a focus on the estimation phase. Propensity weighting and calibrated weighting are among the adjustment methods used to reduce the nonresponse bias. The data collection phase has come into focus more recently; the literature on adaptive survey design emphasizes representativeness and degree of balance as desirable properties of the response obtained from a probability sample. We take an integrated view where data collection and estimation are considered together. For a chosen auxiliary vector, we define the concepts *incidence* and *inverse incidence* and show their properties and relationship. As we show, incidences are used in balancing the response in data collection; the inverse incidences are important for weighting adjustment in the estimation.

**Key words:** adaptive survey design, auxiliary vector, incidence, inverse incidence, nonresponse adjustment, response imbalance.

## 1. Introduction

Weighting techniques are important in producing statistics from sample surveys. Units under-represented in the sample ought to be given a higher weight in the estimation, those over-represented should get a lower weight. This intuitive understanding was probably practiced well before theoretical advancement in the 1930's made it formal: Unbiased estimation in stratified sampling calls for weighting units by the inverse of the stratum sampling rate; the rates may differ considerably between strata. Later and more generally, the Horvitz-Thompson estimator principle established that if the sampling design gives inclusion probability  $\pi_k$  to unit  $k$ , then the weights  $1/\pi_k$  will grant *design unbiased estimation* of a population total. That holds in the absence of nonresponse. This principle has had a great impact on survey methodology for at least 60 years, and continues to be a backbone for methodology, particularly in national statistical institutes, despite heavy unit nonresponse affecting many surveys today, especially those of individuals and households (Bethlehem et al., 2011).

---

<sup>1</sup> Statistics Sweden. E-mail: carl.sarndal@telia.com.

<sup>2</sup> Institute of Mathematics and Statistics, University of Tartu, Estonia. E-mail: imbi.traat@ut.ee.

<sup>3</sup> Questro Analytics Ltd., Tartu, Estonia. E-mail: kaur.lumiste@eesti.ee.

When we come to surveys with nonresponse, specifically to NMAR (not missing at random) nonresponse, weighting techniques continue to be attractive and important, but are less successful in that estimates are no longer unbiased. An inspection of the realized set of respondents may reveal that certain types of sample units are markedly under- or over-represented. Weighting is used to compensate for this, then called “weighting adjustment”. Intuitively, this can reduce bias, perhaps considerably, compared with a passive attitude of a flat weighting, as when we simply use the respondent mean multiplied by the population size. But weighting adjustment will not fully eliminate the bias.

A comprehensive review of nonresponse weighting adjustment was presented by Brick (2013). He identifies three major themes in nonresponse research: (a) Study of the response mechanism; (b) Data collection methods to reduce damage by nonresponse, (c) Adjustment of the survey weights to adjust for survey nonresponse. We are concerned in this article with (b) and (c), and more particularly with the interaction between them. As Brick (2013, p. 347) also notes, a deeper understanding of nonresponse in surveys is prevented by the complexity of the survey process; many unknown factors contribute to it.

With the considerable attention paid recently to responsive (or adaptive) survey design, the practice of weighting comes into a new light. Such designs can bring a more appropriate final set of respondents, compared with a stationary design where the data collection obeys a fixed unchanging protocol from beginning to end. A better balanced response is, potentially, a better starting point for the weighting adjustment in the estimation phase. A review of the literature of adaptive and responsive survey designs is found in Tourangeau et al. (2017). They also suggest directions for further improvement of such designs, and for data collection management more generally.

An adaptive data collection does not follow a stationary protocol. Interventions may take place during the data collection period. Representativeness and low imbalance are general objectives for the ultimate set of respondents. The *R-indicator* of Schouten et al. (2009) is a measure of the former concept. In a similar vein, Särndal (2011), Särndal and Lundquist (2014) used the *Imbalance* statistic to monitor the data collection. Representativeness and balance are related. Both are measured with respect to an auxiliary vector composed of auxiliary variable values known at least for the sample units, possibly for all population units.

Response propensity is another important concept for the data collection. It is a conditional response probability, given the auxiliary vector (Schouten et al., 2011). It is thus a theoretical quantity, defined either at the population level or at the sample level. It can be estimated from a response set. In adaptive design, the response propensity of the sample units is evolving during the data collection period, in tune with of the recruitment protocol changes (Olson and Groves 2012, Schouten et al. 2011).

Until recently, the data collection phase and the estimation phase have been seen largely as separate fields of research. Estimation under nonresponse has a long history and a large literature, namely, on how to apply statistical estimation theory to get the best possible – least biased – estimates, with the “frozen” set of respondents that the data collection happened to give, let alone how “good” or “representative” that set of respondents may be.

With the recent attention paid to adaptive design for the data collection, the need has arisen to know more about how a representative or well-balanced response may help the search for less biased estimates. Designs that optimize collection and adjustment simultaneously need to be developed (Kaminska 2013, p. 356).

We discuss terminology and concepts important for the two phases, the data collection and the estimation. We focus on the interrelation of the two phases and explore the connections that exist, via a multivariate auxiliary vector, between a realized set of respondents and the full (unrealized) probability sample. We see the response not as fixed and frozen but as dynamic, subject to change through the adaptive data collection. Important concepts introduced and studied in Sections 2 and 3 of the paper are *incidence* (of different types of sample units) and *inverse incidence*. The former is used for balancing the response during the data collection period, see Section 4, the latter for weighting responding units at the estimation stage, see Section 5. The two concepts do not necessarily assume a probabilistic response mechanism. A concluding discussion is the topic in Section 6.

## 2. Response Set and Sample Set: One Reflected in the Other

Suppose the survey data collection has resulted in a non-empty response set  $r$ , out of a probability sample  $s$  drawn from the population  $U = \{1, \dots, k, \dots, N\}$ ;  $r \subset s \subset U$ . The response  $r$  is the set of units  $k$  having delivered the value  $y_k$  of the study variable  $y$ . The survey may have several study variables; the discussion and the formulas will necessarily focus on one. The sample  $s$  is drawn from  $U$  so that unit  $k$  has the known inclusion probability  $\pi_k > 0$  and the sampling weight  $d_k = 1/\pi_k$ . The mechanism that generates  $r$  from  $s$  is unknown. The (sample-weighted) survey response rate is  $P = \sum_{k \in r} d_k / \sum_{k \in s} d_k$ , where  $0 < P < 1$  is assumed.

### 2.1. The Auxiliary Vector

In the nonresponse context, three types of variables play a role: The study variable (continuous or categorical)  $y$  has values  $y_k$  observed for  $k \in r$  only, and used to estimate the population total  $Y = \sum_{k \in U} y_k$ . The response indicator  $I$  has value  $I_k = 1$  for  $k \in r$  and  $I_k = 0$  for  $k \in s - r$ .

The auxiliary vector  $x$  with value  $x_k$  is available at least for  $k \in s$ , possibly for  $k \in U$ . The  $J \geq 1$  variables in the vector  $x$  can be continuous or categorical. They are recorded from registers or available as paradata from the data collection process. An early use of the latter information is in Politz and Simmons (1949), a more recent one in Beaumont (2005).

Since  $x_k$  is known for  $k \in s$  we can note, in an ongoing data collection, which values  $x_k$  of the sample units are over-represented (have high incidence) in the realized response  $r$ , and which are under-represented (have low incidence). At the end of data collection, we can analyse the final response outcome with respect the specified vector  $x$ .

In an important special case, all auxiliary variables are categorical. We denote the number of distinct values  $x_k$  by  $M$ , a number possibly different from the vector

dimension  $J$ . More particularly,  $x$  can be a group vector, that is, of the form  $x_k = (0, \dots, 1, \dots, 0)'$  with a single entry "1" to indicate the group membership of  $k$ . Then  $J = M$ . For other kinds of  $x$ -vector,  $J < M$ , where  $M$  may be considerably greater than  $J$ .

To illustrate, if  $x$  represents a crossing of 2 sexes, 3 exhaustive education categories and 4 exhaustive age categories, then  $x$  is a group vector with dimension  $J = 2 \times 3 \times 4 = 24$  and  $J = M = 24$ . If the same three variables are used to define instead the auxiliary vector  $x$  with sex and education crossed, while the categorical age is coded as one of (1,0,0), (0,1,0), (0,0,1) and (0,0,0), then the dimension is only  $J = 2 \times 3 + 3 = 9$ , but  $M$  is unchanged at 24.

We assume that all  $x$ -vectors used here have the following feature: There exists constant vector  $\mu$  (not depending on  $k$ ) such that

$$\mu'x_k = 1 \text{ for all } k. \quad (1)$$

Most vectors of interest satisfy this requirement. When  $x$  is a group vector, the vector  $\mu$  with all elements equal to "1" satisfies (1). In the example above, where  $x$  has sex and education crossed, and age contributing three more positions, the vector  $\mu = (1,1,1,1,1,1,0,0,0)'$  satisfies (1). The reason for the requirement is convenience in many derivations.

## 2.2. The Response Described by the Incidence of the Sampled Units

To say that the response  $r$  is a subset of the sample  $s$ , and to say, inversely, that the set  $s$  contains  $r$ , are weak and uninformative descriptions of the relationship between  $r$  and  $s$ . Their relationship is made more explicit through the intermediary of chosen vector  $x$  and its values  $x_k$  known for  $k \in s$ . No assumptions about the probabilistic nature of the response mechanism are needed in this description.

Given  $r$  and an  $x$ -vector, we ask: What values  $f_k$ , attached to the sample units  $k \in s$ , will give agreement with the observed response mean  $\bar{x}_r = \sum_{k \in r} d_k x_k / \sum_{k \in r} d_k$ ? We seek  $f_k$  for  $k \in s$  to satisfy

$$\sum_{k \in s} d_k f_k x_k / \sum_{k \in s} d_k = \bar{x}_r. \quad (2)$$

Further specification is needed to get a unique solution. One is obtained by letting  $f_k$  be linear in the  $x$ -vector:  $f_k = A'x_k$  for some  $J$ -vector  $A$ . Inserting into (2), and solving, we get  $A' = \bar{x}_r' \Sigma_s^{-1}$ , where the  $J \times J$  matrix

$$\Sigma_s = \sum_{k \in s} d_k x_k x_k' / \sum_{k \in s} d_k \quad (3)$$

is assumed non-singular. Therefore,

$$f_k = \bar{x}_r' \Sigma_s^{-1} x_k, \quad k \in s. \quad (4)$$

We call  $f_k$  the incidence (factor) of unit  $k$ . The mean incidence over  $s$ , as a consequence of (1), is  $\bar{f}_s = \sum_{k \in s} d_k f_k / \sum_{k \in s} d_k = 1$ . The variance over  $s$ ,

$\sum_{k \in s} d_k (f_k - \bar{f}_s)^2 / \sum_{k \in s} d_k$ , is minimal under the constraint in (2). The proof is in the Appendix.

Units with the same value of  $x_k$  share the same incidence  $f_k$ . In the simple example where gender is the only  $x$ -variable, we have  $J = M = 2$ ,  $x_k = (1,0)'$  for all men,  $x_k = (0,1)'$  for all women. Then (4) says that all sampled men have the

incidence  $f_k = P_{\text{men}}/P$ , all sampled women have  $f_k = P_{\text{women}}/P$ , where  $P_{\text{men}}$  and  $P_{\text{women}}$  are the gender response rates and  $P$  the overall rate. This crude kind of response analysis describes how the response for men differs from that of women.

For  $x$ -vectors typically used in practice, the number  $M$  of distinct values can be large. The response rate within groups of units with the same  $x_k$ -value is replaced by the wider concept generalized response rate,  $P_k = P \times f_k$ , which can also be seen as an estimated response propensity for unit  $k$  characterized by  $x_k$ . The mean of  $P_k$  over  $s$  is  $P\bar{f}_s = P$ , the overall response rate.

### 2.3. The Sample Described by the Inverse Incidence of the Responding Units

After a completed data collection, the composition of the response  $r$  can no longer be changed or influenced. We can describe the relationship between  $r$  and  $s$  by the *inverse incidence*. The direction here is to make the smaller set  $r$  conform to the larger set  $s$ , by weighting the units in  $r$ .

We ask: What numbers  $g_k$  applied to the responding units will reproduce the auxiliary sample mean  $\bar{x}_s = \sum_{k \in s} d_k x_k / \sum_{k \in s} d_k$ ? It is futile to ask that question for  $y_k$ , because it is missing for  $k \in s - r$ . This is the inverse of the question in the preceding section. We seek  $g_k$  for  $k \in r$  to satisfy

$$\sum_{k \in r} d_k g_k x_k / \sum_{k \in r} d_k = \bar{x}_s. \quad (5)$$

There is no unique solution. One solution is obtained by forming  $g_k$  as a linear combination of the  $x$ -variables: For some  $J$ -vector  $B$ , set  $g_k = B'x_k$ . Inserting into (5), solving for  $B$ , and assuming that

$$\Sigma_r = \sum_{k \in r} d_k x_k x_k' / \sum_{k \in r} d_k \quad (6)$$

is non-singular, we get

$$g_k = \bar{x}_s' \Sigma_r^{-1} x_k, \quad k \in r. \quad (7)$$

We call  $g_k$  the *inverse incidence (factor)*, or weight, of unit  $k \in r$ . The mean over  $r$  is  $\bar{g}_r = \sum_{k \in r} d_k g_k / \sum_{k \in r} d_k = 1$ , using (1). The variance over  $r$ ,  $\sum_{k \in r} d_k (g_k - \bar{g}_r)^2 / \sum_{k \in r} d_k$ , is minimal under the constraint in (5). The proof is analogous to the corresponding one for  $f_k$ , which is given in the Appendix. Note that  $g_k$  is computable for all  $k \in s$ , because  $x_k$  is available for  $k \in s$ .

## 3. Properties of Incidence and Inverse Incidence

### 3.1. The Moments and the Interrelation

The equation (2) makes a sample  $s$  conform to a realized response  $r$  through the *incidence factor*  $f$  with values  $f_k = \bar{x}_r' \Sigma_s^{-1} x_k$  given in (4) for  $k \in s$ . The equation (5) makes an “upweighted” response  $r$  conform with a given sample  $s$  through the *inverse incidence factor* (or weight factor)  $g$  with values  $g_k = \bar{x}_s' \Sigma_r^{-1} x_k$  given in (7) for  $k \in s$ . The values  $f_k \times g_k$  for  $k \in s$  define the *product factor*.

**Example.** Let  $x$  be a group vector of dimension  $J$ ,  $x_k = (0, \dots, 1, \dots, 0)'$ , coding the same number of different groups of sample units. Suppose that  $s$  is a self-weighting fixed size  $n$  sample. Then  $d_k = N/n$  for all  $k$ , and  $m\bar{x}_r =$

$(m_1, \dots, m_j, \dots, m_J)'$ , where  $m_j$  is the number of responding units in group  $j$ . Alternatively expressed,  $m_j$  is the size of the  $j$ th response group  $r_j$ , and  $m = \sum_{j=1}^J m_j$  is the size of  $r$ . From (4) and (7) we obtain  $f_k = P_j/P$ ,  $g_k = P/P_j$  for all units  $k$  in the same sample group  $s_j$ , where  $P = m/n$ ,  $P_j = m_j/n_j$  is the group  $j$  response rate and  $n_j$  is the size of  $s_j$ ,  $j = 1, \dots, J$ . Hence, when  $x$  is a group vector,  $g_k$  is the inverse of  $f_k$  in an exact numerical sense:  $f_k g_k = 1$  for every  $k$ .  $\square$

In practice, the incidences  $f_k$  for  $k \in s$  are used at the data collection phase, as tools for an adaptive data collection to create a well-balanced final response. This is reviewed in Section 4. The inverse incidences  $g_k$  are used in the estimation phase for weighting adjustment. This is the topic of Section 5. Here we present general properties of  $f_k$  and  $g_k$ .

We derive mean and variance of  $f_k$ ,  $g_k$  and of their product  $f_k \times g_k$ , over the response and over the full sample. For the  $f$  factor, these moments are defined as

$$\bar{f}_r = \text{mean}_r(f) = \sum_{k \in r} d_k f_k / \sum_{k \in r} d_k, \quad \bar{f}_s = \text{mean}_s(f) = \sum_{k \in s} d_k f_k / \sum_{k \in s} d_k, \quad (8)$$

$$\text{var}_r(f) = \sum_{k \in r} d_k (f_k - \bar{f}_r)^2 / \sum_{k \in r} d_k, \quad \text{var}_s(f) = \sum_{k \in s} d_k (f_k - \bar{f}_s)^2 / \sum_{k \in s} d_k. \quad (9)$$

For the corresponding moments of the  $g$  factor, replace  $f$  by  $g$ . For the product factor, replace  $f$  by  $f \times g$  and  $f_k$  by  $f_k \times g_k$  in (8) and (9).

The moments of the three factors are shown in Table 1 for an arbitrary vector  $x$ . Some of the table entries involve quadratic forms in the vector difference  $\bar{x}_r - \bar{x}_s$ :

$$Q_s = (\bar{x}_r - \bar{x}_s)' \Sigma_s^{-1} (\bar{x}_r - \bar{x}_s); \quad Q_r = (\bar{x}_r - \bar{x}_s)' \Sigma_r^{-1} (\bar{x}_r - \bar{x}_s), \quad (10)$$

where the  $J \times J$  weighting matrices  $\Sigma_s$  and  $\Sigma_r$  (non-singular) are given by (3) and (6). Four of the variances have less transparent expressions and are shown only as concepts.

**Table 1.** Mean and variance of  $f$ ,  $g$  and  $f \times g$ . The quantities  $Q_r$  and  $Q_s$  are given in (10).

Factor	mean in $s$	mean in $r$	variance in $s$	variance in $r$
$f$	1	$1 + Q_s$	$Q_s$	$\text{var}_r(f)$
$g$	$1 + Q_r$	1	$\text{var}_s(g)$	$Q_r$
$f \times g$	1	1	$\text{var}_s(f \times g)$	$\text{var}_r(f \times g)$

The properties in Table 1, used in later sections, follow from the definitions in (8) and (9) by standard matrix and vector manipulations, using also  $\bar{x}_s' \Sigma_s^{-1} x_k = \bar{x}_r' \Sigma_r^{-1} x_k = 1$  for all  $k$ , and  $\bar{x}_r' \Sigma_r^{-1} \bar{x}_s = \bar{x}_s' \Sigma_s^{-1} \bar{x}_r = 1$ ; these follow from (1).

By Table 1,  $\bar{f}_r = 1 + Q_s \geq 1 = \bar{f}_s$ . Equality holds only for  $Q_s = 0$ , implying  $\bar{x}_r = \bar{x}_s$ . In general  $f_k \times g_k \neq 1$  for any particular unit  $k$ , but Table 1 shows that the mean of the products  $f_k \times g_k$  is 1, over  $s$  as well as over  $r$ . This interesting property says that one factor is the inverse of the other, in a generalized sense. In

the group vector case the inverse relationship holds in an exact numerical sense,  $f_k g_k = 1$  for every  $k$ .

The covariances are

$$\text{cov}_s(f, g) = \text{mean}_s(f \times g) - \bar{f}_s \bar{g}_s = 1 - 1 \times (1 + Q_r) = -Q_r < 0, \quad (11)$$

$$\text{cov}_r(f, g) = \text{mean}_r(f \times g) - \bar{f}_r \bar{g}_r = 1 - (1 + Q_s) \times 1 = -Q_s < 0 \quad (12)$$

Hence,  $f_k$  and  $g_k$  are negatively correlated, over  $s$  as over  $r$ . More specifically, the coefficient of correlation over  $s$  is usually large negative, not far from  $-1$ . This claim is justified by an approximation shown in the Appendix, whereby

$$\text{corr}_s(f, g) \approx -1/(1 + Q_s). \quad (13)$$

The right-hand side is greater than  $-1$ , but not far from  $-1$ , because compared with  $1$ ,  $Q_s$  is small positive. The approximation in (13) may not be highly accurate for all outcomes  $r$ , given  $s$ , but a large negative correlation is indicated.

The covariances with the auxiliary vector are

$$\text{cov}_s(f, x) = \sum_{k \in s} d_k (f_k - 1)(x_k - \bar{x}_s) / \sum_{k \in s} d_k = (\bar{x}_r - \bar{x}_s), \quad (14)$$

$$\text{cov}_r(g, x) = \sum_{k \in r} d_k (g_k - 1)(x_k - \bar{x}_r) / \sum_{k \in r} d_k = -(\bar{x}_r - \bar{x}_s). \quad (15)$$

It is interesting to note that  $\text{cov}_s(f, x) = -\text{cov}_r(g, x)$ .

The fit of a linear regression with intercept of  $g_k$  on  $f_k$  over  $k \in s$  gives the slope coefficient  $b = \text{cov}_s(f, g) / \text{var}_s(f) = -Q_r / Q_s$  and the intercept  $a = \bar{g}_s - b \bar{f}_s = 1 + Q_r + Q_r / Q_s$ . The predicted  $g_k$ -value from this linear fit is  $\hat{g}_k = a + b f_k$ , so for every  $k \in s$  we have the equation

$$(\hat{g}_k - 1) / Q_r + (f_k - 1) / Q_s = 1. \quad (16)$$

### 3.2. Empirical Illustration of the Relationship

Figure 1 illustrates the relationship between the  $f$ - and  $g$ -factors in a specific experiment. From a data set collected in an Estonian household survey a simple random sample  $s$  of 700 households (HH) was drawn and then kept fixed. A number of characteristics of each household and head of household (HD) were recorded. Response probabilities  $\phi_k$  (where  $k$  designates a household) were then computed for  $k \in s$  by the model

$$\text{logit}(\phi) = 5 - 4 \times \text{HD sex} + 2 \times \text{HD employment status} - 0.0004 \times \text{HH income}.$$

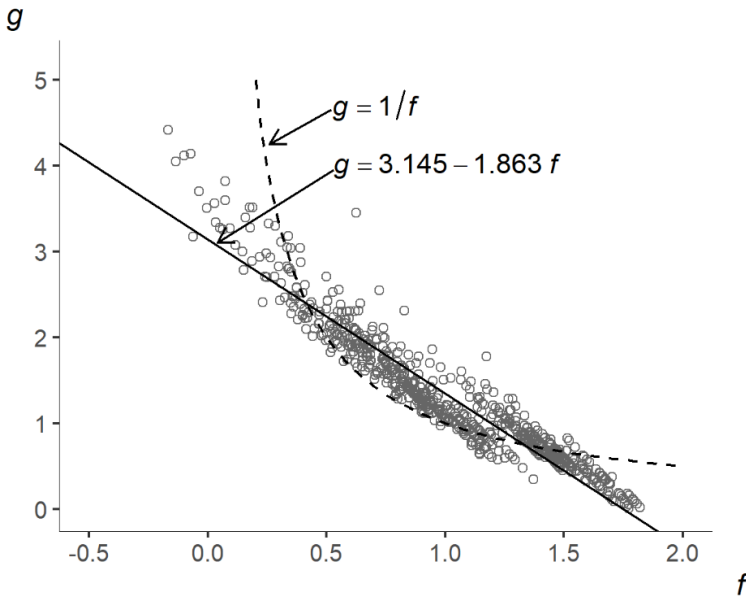
Here, *HD sex* (1 for woman, 0 for man) and *HD employment status* (1 for employed, 0 for unemployed) are dichotomous; *HH income* is continuous. The model deliberately assigns lower response probability to high income households where the head is unemployed female. One single response set  $r$ , with response rate  $P = 60\%$ , was realized by giving household  $k$  the response probability  $\phi_k$ . Given that set  $r$ , computations were carried out with the vector

$x = (\text{HD education, HD sex, HH size, HH children, HD employment status, HH expenditure})$ .



Here, *HD education*, with 3 exhaustive categories, was coded as (1,0,0), (0,1,0) and (0,0,1). The variables *HH size* and *HH children* (the number of children in household) are discrete univariate; *HH expenditure* is continuous. The dichotomous *HD sex* and *HD employment status* are as explained earlier. This  $x$  is not a group vector, so the inverse relationship  $g_k = 1/f_k$  will not hold with exactness for all  $k$ , but it does so to the degree of approximation that Figure 1 illustrates. The dimension of  $x$  is 8: The first variable occupies 3 positions, the other 5 variables one position each. The response set  $r$  has considerable imbalance;  $IMB = 0.055$ , computed by (17) below.

The  $f$ - and  $g$ -factors were computed on the realized  $r$  and  $s$ . The 700 points  $(f_k, g_k)$  for  $k \in s$  are plotted (as hollow small circles) in Figure 1. The figure illustrates that  $f_k$  can be negative for a small number of units  $k \in s$ . In the figure, none of the points with  $f_k < 0$  belong to  $r$ . Consequently, the linear approximation of  $g_k$  through  $f_k$  works quite well in the response set  $r$ . The solid line is the linear regression line  $g = a + bf$ , with  $a = 1 + Q_r + Q_r/Q_s = 3.145$  and  $b = -Q_r/Q_s = -1.863$ . The dashed curve is  $g = 1/f$ . We verified empirically, for the group vector  $x = (\text{HD education} \times \text{HD employment status})$ , with  $3 \times 2 = 6$  groups, that  $g_k = 1/f_k$  holds exactly for all  $k$ , as it should.



**Figure 1.** Relationship between  $f$ - and  $g$ -factors for a sample of size 700. Each circle represents a sample element.

#### 4. Achieving Low Imbalance in the Data Collection

The incidences  $f_k$  are important for the data collection. They are used for creating a well balanced response set. The response  $r$  is called perfectly balanced with respect to the vector  $x$  if  $\bar{x}_r = \bar{x}_s$  (Särndal, 2011). It follows from (2)

that the equality in means is achieved if  $f_k = 1$  for all  $k$ . The equality  $\bar{x}_r = \bar{x}_s$  also holds if  $g_k = 1$  for all  $k$ , as seen from (5). To get a perfectly balanced response  $r$  is a distant possibility in a survey data collection, especially for a long  $x$  vector. We can strive to come close. But ordinarily, a perfect balance is not achieved. Since  $\bar{x}_r - \bar{x}_s$  is a vector, a scalar measure of the difference is created, called the *imbalance* of the response  $r$  with respect to the vector  $x$  for the given sample  $s$ ,

$$\text{IMB}(r, x|s) = P^2 Q_s, \quad (17)$$

where  $P$  is the response rate and  $Q_s$  is given in (10) (Särndal, 2011; Lundquist and Särndal, 2013; Särndal and Lundquist, 2014). Although  $\text{IMB}(r, x|s)$  is more descriptive, we shall use for simplicity the notation  $\text{IMB}$ . For any  $r, s$  and vector  $x$ ,  $0 \leq \text{IMB} \leq P(1 - P) \leq 0.25$ . For most survey data,  $\text{IMB}$  does not come close to the upper bound  $P(1 - P)$ ; typical values are in the range 0.03 to 0.06.

A measure related to  $\text{IMB}$  is the  $R$ -indicator, with  $R$  for “representativeness” (Bethlehem et al., 2011). It is different in its background, which is estimation of response probabilities assumed to exist for all population units.

The incidences  $f_k$ , computable for all  $k \in s$ , are tools for an adaptive data collection aiming at an ultimate response set  $r$  with low imbalance. A property making this possible is that the variance (computed over  $s$ ) of the (estimated) propensities  $P_k = Pf_k$  is equal to the imbalance,  $\text{IMB} = P^2 Q_s$  (see Table 1). The  $P_k$  can be computed continuously during an ongoing data collection period. Therefore, an avenue to low imbalance in the final response  $r$  is to manage the data collection to achieve in the end a low variance of  $P_k$ , and therefore low  $\text{IMB}$ . There may be several ways to accomplish this. One is the threshold method proposed in Särndal and Lundquist (2014), which we now describe.

The data collection, which may last several days or weeks, is seen as a dynamic process where inspections and change of protocol may take place, at specified points. For example, one may decide, at a certain point, to focus the continued data collection on specific types of units, say those that are so far underrepresented.

In the threshold method, the propensities  $P_k = Pf_k$  are computed for  $k \in s$  at several points, say four to six, in the data collection period, and with a “monitoring vector”  $x$  designated for this purpose.

At the first inspection point, units with propensity greater than a fixed threshold, say 0.60, are set aside and not further contacted during the period. Contact attempts continue with the remaining non-responding sample units; as a result more units join the response set. At the second inspection point,  $P_k$  is computed again for all  $k \in s$ , and some more units, those with the new propensity  $P_k$  greater than 0.60, join those already set aside. This pattern is repeated at each remaining inspection point; at each of these some more units are set aside. Non-responding units remaining at the last inspection point are pursued until the very end of the data collection period. By the mechanics of this procedure, the variability of the propensities - and therefore the imbalance  $\text{IMB}$  - is more and more reduced. In the end, the imbalance  $\text{IMB}$  can be quite low. Alternative adaptive designs can be constructed with a similar objective.

## 5. The Estimation Stage

After a completed data collection, it remains to produce estimates of important finite population parameters, such as the population total  $Y = \sum_{k \in U} y_k$ , using the values  $y_k$  available for  $k \in r$ . The estimates are design biased, more or less.

If individual response probabilities  $\phi_k$  were known, then  $\hat{Y}_{2ph} = \sum_{k \in r} d_k \phi_k^{-1} y_k$  would be unbiased for the total  $Y = \sum_{k \in U} y_k$ . This claim derives from design-based theory for two-phase selection: First a probability sample  $s$  from  $U$ , then a response  $r$  from the given  $s$ . Since  $\phi_k$  is unknown,  $\hat{Y}_{2ph}$  should be adjusted. Brick (2013) reviews three types of weighing adjustment procedures in surveys with nonresponse. In the first of these, the unknown individual response probabilities  $\phi_k$  in  $\hat{Y}_{2ph}$  are replaced by estimates  $\hat{\phi}_k$ . This results in

$$\hat{Y}_{ADJ} = \sum_{k \in r} d_k \hat{\phi}_k^{-1} y_k, \quad (18)$$

also referred to as “quasi-randomization” estimators. Access to suitable auxiliary variables and the choice of the model for the response mechanism play an important role in (18).

Brick’s (2013) second type is the weighting class estimator. It is a special case of (18), where  $\hat{\phi}_k^{-1}$  is equal to the inverse of a group response rate. That is, if the sample  $s$  is divided into  $J$  mutually exclusive and exhaustive subgroups  $s_j$  with  $r_j$  as the responding subset of  $s_j$ ,  $j = 1, \dots, J$ , then  $\hat{\phi}_k^{-1} = \sum_{k \in s_j} d_k / \sum_{k \in r_j} d_k$ , common to all units  $k$  in a group.

The third weighting adjustment estimator in Brick’s (2013) review is the calibration estimator. It differs in its construction from (18) but is still unmistakably design-based in its orientation. All three weighting adjustment procedures are imperfect under nonresponse because they fail to meet the design-based criterion of unbiased estimation.

Here, we distinguish three arguments for constructing an estimator for  $Y = \sum_{k \in U} y_k$ . They are: Weighting by inverse incidence (Section 5.1), calibration estimation (Section 5.2) and estimation by explicit modelling/prediction (Section 5.3).

### 5.1. Weighting by Inverse Incidence

Weighting by inverse incidence does not require any response model. It reflects the intuitive idea that units in  $r$  with low incidence get relatively higher weight, and vice versa.

The incidence factor  $f_k$  is given in (4), the inverse incidence factor  $g_k$  in (7). Now put

$$P_k = P f_k; \quad v_k = P^{-1} g_k, \quad (19)$$

where  $P$  is the overall response rate.  $P_k$  and  $v_k$  are each other’s inverse, in that the mean of their product,  $P_k v_k = f_k g_k$ , is equal to one, over  $r$  and over  $s$  (see Table 1). The inverse incidence weighting estimator of  $Y = \sum_{k \in U} y_k$  is then given by

$$\hat{Y}_{WEI} = \sum_{k \in r} d_k v_k y_k. \quad (20)$$

This is weighting adjustment as in (18), if we let  $\hat{\phi}_k^{-1} = v_k$ . Moreover,  $P_k$  is reminiscent of a second phase inclusion probability for unit  $k$ , that is, in “drawing” the response set  $r$  from  $s$ . The sample mean of  $P_k$  is  $\bar{P}_s = \sum_{k \in s} d_k P_k / \sum_{k \in s} d_k = P \bar{f}_s = P$ , the overall response rate.

The weighting in (20) is motivated purely by inverse incidence, based on a given  $x$ -vector, with no particular variable  $y$  in mind. The same weights are applied to all variables  $y$ , whatever their special characteristics. This is appealing in surveys where many  $y$ -variables require estimation, none of them deemed to be truly more important or different in nature. Implicit in the inverse incidence weighting is a relationship between the 0/1 indicator of the response and the auxiliary vector  $x$  that determines the incidence  $f_k$ .

## 5.2. Calibration Estimation

A well-known weighting adjustment estimator is the calibration estimator. Weighting is based on  $x$  with implicit  $y$ -to- $x$  relationship. Still, all  $y$ -variables are typically given the same weighting. For comparability reasons, we consider calibration up to  $s$ . Weight factors  $u_k$  are calibrated “from  $r$  up to  $s$ ”, to satisfy the calibration equation

$$\sum_{k \in r} d_k u_k x_k = \sum_{k \in s} d_k x_k. \quad (21)$$

The resulting calibration estimator is then

$$\hat{Y}_{\text{CAL}} = \sum_{k \in r} d_k u_k y_k. \quad (22)$$

If we choose  $u_k$  to be linear in  $x_k$ ,  $u_k = \lambda' x_k$ , it follows from the derivation in Section 2.3 that  $u_k = P^{-1} g_k$ , where  $g_k$  is the inverse incidence given in (7). Then, (22) is the linear calibration estimator,  $\hat{Y}_{\text{CALlin}}$ , which we can express in several ways:

$$\hat{Y}_{\text{CALlin}} = \sum_{k \in r} d_k v_k y_k = P^{-1} \sum_{k \in r} d_k g_k y_k = \sum_{k \in s} d_k \hat{y}_k = \hat{N} \bar{x}'_s b_r, \quad (23)$$

where  $\hat{y}_k = x'_k b_r$  and  $b_r$  is the regression coefficient vector in a linear regression fit of  $y$  on  $x$  over  $r$ ,

$$b_r = (\sum_{k \in r} d_k x_k x'_k)^{-1} \sum_{k \in r} d_k x_k y_k. \quad (24)$$

Hence, the inverse incidence weighting estimator  $\hat{Y}_{\text{WEI}}$  in (20) has a double identity: It is at the same time a (linear) calibration estimator.

The purely mechanical aspect of the calibration approach is to deliver weights to satisfy (21) – which has an unbiased Horvitz-Thompson estimator on the right hand side – and to apply these weights in the estimation. But the purpose is also to explain the  $y$ -variable through the auxiliary vector  $x$ . The calibration approach is thus double-natured: The weighting aspect is combined with implicit relationship  $y$ -to- $x$ . This can be seen when we examine the deviation of  $\hat{Y}_{\text{CALlin}}$  from the unbiased estimator requiring full response,  $\hat{Y}_{\text{FUL}} = \sum_{k \in s} d_k y_k$ . This deviation can be written as

$$\hat{Y}_{\text{CALlin}} - \hat{Y}_{\text{FUL}} = - \sum_{k \in s} d_k e_k \quad (25)$$

with the residual  $e_k = y_k - x'_k b_r$ , where  $b_r$  is the regression vector given in (24). If the model fits well in the response set, the residuals are small, and  $\hat{Y}_{\text{CALin}}$  based on the response is close to the unbiased  $\hat{Y}_{\text{FUL}}$ .

Calibration estimators have been extensively studied for the last 20 years. One direction is to use information both in the sample and population levels. Another direction is to use non-linear forms of calibration. Some references are Deville (1998), Deville and Särndal (1992), Folsom and Singh (2000), Estevao and Särndal (2000), Montanari and Ranalli (2003, 2005, 2012), Särndal and Lundström (2005), Chang and Kott (2008), Kott and Chang (2010), Kott and Liao (2012).

### 5.3. Estimation by Explicit Modelling/Prediction

The modelling/prediction approach is based on replacing missing  $y$ -values by the best possible substitutes that statistical theory can offer. This argument is, on surface at least, very different from both incidence weighting and calibration weighting. Its importance is illustrated by Little's (2013) discussion of Brick (2013).

This approach focuses directly on one  $y$ -variable at a time. From an explicitly formulated (linear or non/linear) model for the  $y$ -to- $x$  relationship, and a fit of that model based on  $(y_k, x_k)$  for  $k \in r$ , predicted values are obtained for the non-observed  $y_k$ , using the values  $x_k$  known for  $k \in s - r$ . Observed  $y_k$  together with predictions  $\hat{y}_k$  are used to build the estimator of the population  $y$ -total,

$$\hat{Y}_{\text{PRED}} = \sum_{k \in r} d_k y_k + \sum_{k \in s-r} d_k \hat{y}_k. \quad (28)$$

Examination of the design-based behaviour of  $\hat{Y}_{\text{PRED}}$  has shown that strong regression relationship holds good prospects for a considerable reduction of the (design-based) nonresponse bias. Early references are Bethlehem (1988) and Cassel et al. (1983).

A variety of models and methods can be entertained to get the predicted values  $\hat{y}_k$ . A simple application is by ordinary linear regression fit of  $y$  on  $x$ , resulting in the regression vector  $b_r$  in (24) and predicted values  $\hat{y}_k = x'_k b_r$  for  $k \in s$ . Note that  $\sum_{k \in r} d_k (y_k - \hat{y}_k) = 0$  because of (1). Then

$$\hat{Y}_{\text{PREDlin}} = \sum_{k \in r} d_k y_k + \sum_{k \in s-r} d_k \hat{y}_k = \sum_{k \in s} d_k \hat{y}_k = \hat{Y}_{\text{CALin}} = \hat{Y}_{\text{WEI}}. \quad (29)$$

It can also be seen as a result of the linear generalized regression (GREG) construction;

$$\hat{Y}_{\text{GREG}} = \sum_{k \in s} d_k \hat{y}_k + \sum_{k \in r} d_k (y_k - \hat{y}_k) = \sum_{k \in s} d_k \hat{y}_k. \quad (30)$$

Hence the inverse incidence weighting estimator  $\hat{Y}_{\text{WEI}}$  in (20) has multiple identities: It is at the same time (a) a calibration estimator, (b) a prediction estimator, and (c) a GREG estimator. It is important to note that this equivalence happens under the linear formulation, and under the  $x$ -vector condition in (1).

We can link the bias to the tendency of nonresponse to misrepresent the regression relationship: Denote by  $b_s = (\sum_{k \in s} d_k x_k x'_k)^{-1} \sum_{k \in s} d_k x_k y_k$  the regression coefficient vector in the linear fit of  $y$  on  $x$  over  $s$ . Then, by (1),  $\hat{N} \bar{x}'_s b_s = \hat{N} \bar{y}_s = \sum_{k \in s} d_k y_k = \hat{Y}_{\text{FUL}}$  and the deviation from the unbiased estimation can be written as

$$\hat{Y}_{\text{PREDlin}} - \hat{Y}_{\text{FUL}} = \hat{Y}_{\text{CALin}} - \hat{Y}_{\text{FUL}} = \hat{N} \bar{x}'_s (b_r - b_s), \quad (31)$$

where  $b_r$  is given in (24). As is well known from regression theory, the selection effect is likely to distort an estimated regression relationship, that is, to make the regression vectors  $b_r$  and  $b_s$  differ considerably, and thus  $\hat{Y}_{\text{PREDlin}}$  to differ from the unbiased  $\hat{Y}_{\text{FUL}}$ . Särndal et al. (2016) evaluate the deviation  $\Delta = (\hat{Y}_{\text{CALlin}} - \hat{Y}_{\text{FUL}})/\hat{N} = \bar{x}_s'(b_r - b_s)$  under certain assumptions, and find potential for improved accuracy under adaptive design. Expressions for the design-based bias have been derived for some types of regression-based estimators (Fuller et al. 1994, Särndal and Lundström 2005, Brick and Jones 2008).

In the model-based version of the modelling/prediction approach, the sampling design and the sampling weights  $d_k$  may not enter at all. A comprehensive coverage is found in books such as Valliant et al. (2000) and Chambers et al. (2012). Other recent contributions are Breidt and Opsomer (2000), Breidt et al. (2005), Little (1986).

## 6. Conclusion

We have examined a survey setting where nonresponse is occurring in a probability sample from the finite population. We emphasized an integrated view, in which the data collection and the estimation stage can benefit from each other, and support each other, in making inference about the population.

We have assumed that an appropriate auxiliary vector was formulated, from the available supply of auxiliary variables, categorical or continuous. We discussed the auxiliary vector's important role in forming a bridge between a realized set of respondents and the full probability sample. To that end, we formulated the concepts of *incidence* and *inverse incidence* of the sample units. A realized response set can be described by the (computable) incidences of the sample units; vice versa, the drawn sample can be described by the (also computable) inverse incidences of the responding units.

As we showed, the incidences are used in an adaptive data collection to realize a final response set with low imbalance. The inverse incidences are used at the estimation stage, for building a weighted estimator. It is one that does not use any assumptions about a probabilistic response mechanism. We pointed out that it coincides, in the special case of a "linear formulation", with estimators derived by other approaches: Calibration, modelling/prediction and GREG. These approaches have branched out in their own directions and have generated a stream of literature that we do not review here.

To a considerable degree, this article has dealt with concepts and principles. This has left unanswered a number of other important aspects. Among these is the question whether a reduced imbalance in the ultimate response set will lead to reduced bias in the estimates, over and beyond what (weighting) adjustment alone can accomplish at the estimation stage. There is some positive evidence in this direction in the recent literature. A relationship between auxiliary vector  $x$  and survey variable  $y$  is implicitly assumed; one can say that balancing the survey response gives some added protection against large nonresponse bias. Recent articles in this direction are Schouten et al. (2016) and Särndal et al. (2016). Also, Tourangeau et al. (2017) confirm that a bias reduction, although perhaps marginal, can be realized by balancing, and these authors claim that further

improvement may be possible, through alternative and better adaptive designs. These and other recent contributions to the literature underline the need for an integrated view, one where data collection and estimation are considered together; in this article, we have also taken a step in that direction.

## **Acknowledgements**

This work was partly supported by the Institutional Research Funding IUT34-5 of Estonia.

## REFERENCES

- BEAUMONT, J. F., (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31, pp. 227–231.
- BETHLEHEM, J. G., (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, pp. 251–260.
- BETHLEHEM, J. G., COBBEN, F., SCHOUTEN, B., (2011). *Handbook on Nonresponse in Household Surveys*. New York: Wiley.
- BREIDT, F. J., CLAESKENS, G., OPSOMER, J. D., (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92 (4), pp. 831–846.
- BREIDT, F. J., OPSOMER, J. D., (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28 (4), pp. 1026–1053.
- BRICK, J. M., (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29 (3), pp. 329–353.
- BRICK, J. M., JONES, M. E., (2008). Propensity to respond and nonresponse bias. *Metron*, 66 (1), pp. 51–73.
- CASSEL, C., SÄRNDAL, C. E., WRETMAN, J., (1983). Some uses of strategical models in connection with the nonresponse problem. In: *Incomplete Data in Sample Surveys*, ed. By W. G. Madow and I. Olkin, Vol. 3, New York: Academic Press.
- CHAMBERS, R. L., CLARK, R. G., (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press.
- CHANG, T., KOTT, P. E., (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95 (3), pp. 555–571.
- DEVILLE, J. C., (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. Paper presented at Congrès de l'ACFAS, Sherbrooke, Québec.
- DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration estimation in survey sampling. *Journal of the American Statistical Association*, 87 (418), pp. 375–382.
- ESTEVAO, V., SÄRNDAL, C. E., (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, pp. 379–399.
- FOLSOM, R. E., SINGH, A. C., (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse and poststratification. *Proceedings, Section of Survey Research Methods, American Statistical Association, Washington DC*, pp. 598–603.



- FULLER, W. A., LOUGHIN, M. M., BAKER, H. D., (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey. *Survey Methodology*, 20, pp. 75–85.
- KAMINSKA, O., (2013). Unit nonresponse and weighting adjustments: a critical review: discussion. *Journal of Official Statistics*, 29, pp. 355–358.
- KOTT, P. S., CHANG, T., (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105 (491), pp. 1265–1276.
- KOTT, P. S., LIAO, P., (2012). Providing double protection for unit nonresponse with a nonlinear calibration weighting. *Survey Research Methods*, 6 (2), pp. 105–111.
- LITTLE, R. J. A., (2013). Discussion. *Journal of Official Statistics*, 29, pp. 363–366.
- LITTLE, R. J. A., (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54 (2), pp. 139–157.
- LUNDQUIST, P., SÄRNDAL, C.-E., (2013). Aspects of responsive design. With applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, pp. 557–582.
- MONTANARI, G. E., RANALLI, M. G., (2003). Nonparametric methods in survey sampling. In M. Vinci, P. Monari, S. Mignani, A. Montanari (eds.), *New Developments in Classification and Data Analysis*. Berlin: Springer.
- MONTANARI, G. E., RANALLI, M. G., (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, pp. 1429–1442.
- MONTANARI, G. F., RANALLI, M. G., (2012). Calibration inspired by semiparametric regression as a treatment for nonresponse. *Journal of Official Statistics*, 28, pp. 239–277.
- OLSON, K., GROVES, R. M., (2012). An Examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, 28, pp. 29–51.
- POLITZ, A., SIMMONS, W., (1949). An attempt to get “Not at Homes” into the sample without callbacks. *Journal of the American Statistical Association*, 44 (245), pp. 9–31.
- 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.
- SÄRNDAL, C.-E., LUMISTE, K., TRAAT, I., (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? *Survey Methodology*, 42 (2), pp. 219–238.
- SÄRNDAL, C.-E., LUNDSTRÖM, S., (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

- SÄRNDAL, C.-E., LUNDQUIST, P., (2014). Accuracy in estimation with nonresponse: a function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2, pp. 361–387.
- SCHOUTEN, B., COBBEN, F., BETHLEHEM, J., (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, pp. 101–113.
- SCHOUTEN, B., COBBEN, F., LUNDQUIST, P., WAGNER, J., (2016). Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society, Series A*, 179 (3), pp. 727–748.
- SCHOUTEN, B., SHLOMO, N., SKINNER, C., (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, pp. 1–24.
- TOURANGEAU, R., BRICK, J. M., LOHR, S., LI, J., (2017). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society, Series A*, 180 (1), pp. 201–223.
- VALLIANT, R., DORFMAN, A. H., ROYALL, R. M., (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons Inc.

## APPENDIX

**Proof that the incidence factors  $f_k$  in (4) have minimal variance subject to (2):**

Using the Lagrange multiplier method, we seek the minimum of

$$\sum_{k \in S} d_k (f_k - \bar{f}_s)^2 - 2\lambda' (\sum_{k \in S} d_k f_k x_k - (\sum_{k \in S} d_k) \bar{x}_r). \quad (32)$$

Setting the derivative with respect to  $f_k$  equal to zero gives

$$2d_k (f_k - \bar{f}_s) - 2d_k \lambda' x_k = 0; \quad f_k - \bar{f}_s = \lambda' x_k. \quad (33)$$

Determine  $\lambda$  from the condition in (2):  $\lambda' = \bar{x}_r' \Sigma_s^{-1} - \bar{f}_s \bar{x}_s' \Sigma_s^{-1}$ . Post-multiply by  $x_k$  and note that  $\bar{x}_s' \Sigma_s^{-1} x_k = 1$  by (1). This gives  $\lambda' x_k = \bar{x}_r' \Sigma_s^{-1} x_k - \bar{f}_s$  and  $f_k = \bar{f}_s + \lambda' x_k = \bar{x}_r' \Sigma_s^{-1} x_k$ , as given in (4).

### Derivation of the approximation in (13):

By definition,  $\text{corr}_s(f, g) = \text{cov}_s(f, g) / (\text{var}_s(f) \text{var}_s(g))^{1/2}$ . First use  $\text{var}_s(g) / \bar{g}_s^2 \approx \text{var}_r(g) / \bar{g}_r^2$ , assuming that the coefficient of variation of  $g$  (standard deviation divided by mean) is roughly the same over  $r$  as over  $s$ . Then by Table 1,  $\text{var}_s(g) \approx Q_r (1 + Q_r)^2$ , and  $\text{var}_s(f) = Q_s$ . Both  $Q_r$  and  $Q_s$  are small compared to 1 and not greatly different, so  $(1 + Q_r) / (1 + Q_s) = 1 + \delta$  for some small  $\delta$ . Then

$$\text{corr}_s(f, g) = \frac{-Q_r}{(Q_s Q_r)^{1/2} (1 + Q_r)} = -\frac{1}{1 + Q_s} h(\delta), \quad (34)$$

where  $h(\delta) = (1 + (1 + Q_s^{-1})\delta)^{1/2} / (1 + \delta)$ . Now, for small  $\delta$ ,  $h(\delta) \approx 1$ . The formula in (13) follows.