DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 709–734, http://dx.doi.org/10.1515/JOS-2017-0033

Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance

Carl-Erik Särndal[1] *and Peter Lundquist*[2]

One objective of adaptive data collection is to secure a better balanced survey response. Methods exist for this purpose, including balancing with respect to selected auxiliary variables. Such variables are also used at the estimation stage for (calibrated) nonresponse weighting adjustment.

Earlier research has shown that the use of auxiliary information at the estimation stage can reduce bias, perhaps considerably, but without eliminating it. The question is: would it have contributed further to bias reduction if, prior to estimation, that information had also been used in data collection, to secure a more balanced set of respondents? If the answer is yes, there is clear incentive, from the point of view of better accuracy in the estimates, to practice adaptive survey design, otherwise perhaps not.

A key question is how the regression relationship between the survey variable and the auxiliary vector presents itself in the sample as opposed to the response. Strength in the relationship is helpful but is not the only consideration. The dilemma with nonresponse is one of inconsistent regression: a regression model appropriate for the sample often fails for the responding subset, because nonresponse is selective, non-random.

In this article, we examine how nonresponse bias in survey estimates depends on regression inconsistency, both seen as functions of response imbalance. As a measure of bias we use the deviation of the calibration adjusted estimator from the unbiased estimate under full response. We study how the deviation and the regression inconsistency depend on the imbalance. We observe in empirical work that both can be reduced, to a degree, by efforts to reduce imbalance by an adaptive data collection.

Key words: Accuracy; adaptive data collection; auxiliary variables; balanced response; responsive design.

1. Introduction

1.1. Responsive Design

Responsive design for household and other surveys is a concept due to Groves and Heeringa (2006) aiming at active control of survey errors and costs, mainly in the planning and data collection phases. A number of contributions to the literature have followed in this vein.

In view of the high survey nonresponse in many surveys today, responsive or adaptive data collection has been promoted as a possibility to obtain, in a cost efficient way, a final

[1] Carl-Erik Särndal, Statistics Sweden, Klostergatan 23, 701 89 Örebro. Email: carl.sarndal@telia.com
[2] Peter Lundquist, Statistics Sweden, Karlavägen 100, 104 51, Stockholm, Sweden. Email: peter.lundquist@scb.se
Acknowledgments: The authors thank two referees and an Associate Editor, all anonymous, for helpful comments.

set of respondents with improved chances for accurate – less biased – survey estimates. Techniques have been suggested and examined: case prioritization, stopping rules, balancing, and so on. The terms representativeness and balance are now often used in discussing quality of the set of responding units.

Representativeness and other indicators are discussed in Schouten et al. (2009) and Schouten et al. (2011). One measure is known as the R-indicator.

Related to "representativeness" is "balance". One may describe their relationship by saying that representativeness is a property of a realized response set, whereas balancing is an activity during data collection that tries to deliver in the end a well representative response.

With a long history in statistics, balance has been used in somewhat different meanings, and well before its application to survey nonresponse. Balancing with respect to chosen auxiliary variables rests on an idea of equality, or closeness, of means in a smaller set of units with corresponding means in a larger set that contains the smaller one. Examples are Deville and Tillé (2004), for balancing a probability sample with the cube method, and Legg and Yu (2010), for sample set restriction procedures.

In a context of survey nonresponse, balancing the data collection aims at equality or near-equality of auxiliary variable means for the set of respondents with corresponding means for the selected probability sample, or for the population. To this end, a measure of imbalance is continuously monitored, and methods exist to reduce it in a data collection extending over a period of time, days or weeks, during which contact attempts, some of them unfruitful, are made with the units in the selected probability sample. The imbalance statistic was used in this manner in Lundquist and Särndal (2013), Särndal and Lundquist (2014). In practice, the imbalance can be reduced, but not eliminated.

Reduced imbalance, for a given auxiliary vector $\mathbf{x}$, may well be obtained in data collection, but for what good? Will it contribute significantly to improved accuracy in the estimates for the survey variables y? Even if not used in data collection, the auxiliary vector $\mathbf{x}$ would serve to compute calibrated adjustment weights at the estimation stage. Is there then an added advantage in using also that vector in data collection to get better balance? If yes, there is combined usage of the auxiliary variables: They are re-utilized at the estimation stage, which is perfectly legitimate, even recommendable. The issue here is the one we refer to as *single usage* versus *combined usage* of the auxiliary vector, that is, use at the estimation stage only as opposed to use of the vector first in data collection to get reduced imbalance, then again at the estimation stage for calibrated weighting.

Whether or not better balanced response will improve the accuracy of survey estimates is primarily a question of statistical inference. It poses a challenge to the field of adaptive or responsive survey design. Is such design worth it, from the point of view of improved accuracy and reduced nonresponse bias?

1.2. Auxiliary Variables, Their Role in Data Collection and in Estimation

Statistical agencies in some countries can count on a vast supply of auxiliary variables to choose from, known for the full probability sample or for the population, but such richness does not entirely resolve the nonresponse bias problem.

It is commonly understood that the selected auxiliary variables should (a) well explain the survey variables y, or at least the most important among those, and/or (b) well explain the response, that is, the 0/1 indicator of nonresponse/response. Those are guidelines for a "post data collection activity".

Thus, a vast literature treats the nonresponse problem essentially as one of statistical inference with a fixed unchanging set of respondents: Try to get the best estimation – notably least possible nonresponse bias – with the realized response, the one that a terminated data collection happened to give. The auxiliary variables play a central role, but improving the data collection phase is not an issue. These attempts often involve a use of estimated response probabilities for a response that may be assumed non-ignorable. Recent reviews and discussions of the field of weighting adjustment for nonresponse are Brick (2013), Matei and Ranalli (2015), Haziza and Lesage (2016), and Tourangeau et al. (2017).

A merit of the responsive design movement is that it has put the data collection in focus. The data collection is undeniably an integral part of the inference process, particularly important in times of high survey nonresponse. Several recent articles seek evidence, theoretical and empirical, on possible favorable effects of a response that is made to be more representative or better balanced, for example, Schouten et al. (2016), Schouten et al. (2013), Lundquist and Särndal (2013), Särndal et al. (2016). These articles point to some, although not strongly pronounced, favorable effect of balancing. Articles that treat questions of a related nature are Vartivarian and Little (2002) and Little and Vartivarian (2005).

The selection of auxiliary variables, when there are many to choose from, is discussed, for example, in Schouten (2015). It is an important issue, because even "picking many" auxiliary variables gives at best a partial correction for nonresponse bias. A random pick from a large set of available auxiliary variables may even be justified, or not markedly worse, than a more directed choice.

An adaptive data collection operates on a designated auxiliary vector $\mathbf{x}$ with values known for the sample. Särndal and Lundquist (2014) found that the estimator formed by calibration on a well-motivated $\mathbf{x}$-vector will deviate less (from the unbiased, full response estimator) if the response has been tailored to have low imbalance (as measured by an appropriate statistic for that concept). We tend to describe the improvement as "undeniable, but modest rather than strong".

We can thus expect balancing to give a certain reduction of nonresponse bias. Although important, it is an incomplete cure. Adaptive survey design can be attractive for reasons that are practical more so than theoretical, controlling survey cost for example. But in a statistical inference perspective, a certain disappointment may be felt by some with the progress to date. Why does balancing not reduce the bias much more decisively, to near zero levels? Some of the reasons why this does not happen are indicated in the following.

1.3. Regression Inconsistency

In the estimation phase, calibrated weights are computed on the chosen calibration vector $\mathbf{x}$ and applied to the units in the responding subset r from the probability sample s. The resulting estimator will still have remaining bias. The deviation of the calibration adjusted estimator from the unbiased estimator under full response is a focal point in this article.

The size of this deviation is intimately connected to the nature and the quality of the regression relationship between the study variable y and the auxiliary vector $\mathbf{x}$. If an assumed regression relationship between $\mathbf{x}$ and y would hold perfectly, the deviation is zero. This ideal is never met in the real world. Considerable deviation and bias may remain after calibrated weight adjustment.

The nature of the relationship between $\mathbf{x}$ and y must thus be questioned. The assumed model, linear or nonlinear, for the relationship may be incorrect, that is, the mathematical form of the model may not be right. Furthermore, the explanatory x-variables posted in the model may not be powerful enough.

Although important, these features are often only a minor part of the bias problem. More serious is that the model fitted on the response – which is what we have to work with – does not correctly portray the relationship existing in the sample that nonresponse prevented us from fully observing. A regression model holding for the sample cannot be estimated correctly on the response realized from the sample, because nonresponse is selective, informative.

Hence bias is often more a selection problem – non-randomness, or non-ignorability, of the nonresponse – than one of misspecified form of the model or one of omitted predictor variables.

Bias due to survey nonresponse is therefore connected with the selection problem in regression theory. The rich literature following the seminal article by Heckman (1979) proposes methods to diminish or reduce the problem of biased estimation of regression relationship. Many articles deal with extensions of "Heckman methodology"; a review is, for example, Vella (1998). To apply these methods would again amount to a "post data collection activity".

In the survey statistics literature, non-ignorable nonresponse has also been extensively studied. Frequently, a model assumed for the sample is not well estimated by the response. The estimated model parameters are wrong; consequently, so are the predictions of the missing y-values. The regression estimated on the response is inconsistent. In this article, we identify a measure of *regression inconsistency*. Alternatively it might be called "lack-of-fit measure", but "inconsistency" seems more appropriate, since the root of the problem is selection bias rather than omitted predictor variables.

We are then able to write the deviation of the calibration adjusted estimate as a sum of a regression inconsistency term and a residual that is zero under perfect balance, that is, when $\mathbf{x}$-vector means in the response set and in the full sample fully agree.

From earlier work we know that the deviation of the calibration estimator from unbiasedness is reduced to some extent by a reduced response imbalance. Here we ask: will reduced imbalance also help to overcome or reduce the regression inconsistency? Hence, we study the *deviation* and the part of it that is due to *regression inconsistency*, viewing both as functions of the degree of response *imbalance*.

The contents are arranged as follows: Section 2 introduces the notation. Section 3 deals with estimation for a survey variable observed for only the responding subset. The auxiliary vector and its relation to survey variable y are then in focus. The calibration estimator is reviewed. Its deviation from the unbiased estimate (requiring full response) is decomposed into a regression inconsistency term and a residual. Section 4 discusses response imbalance and response propensity, and their effect on the regression

inconsistency. The relation between imbalance and regression inconsistency is analyzed. Section 5 gives the background for the empirical work. Several hypotheses are stated for empirical testing. This work relies on data from Statistics Sweden's Living Conditions Survey and its Labour Force Survey; these are described. For these two surveys, Section 6 presents empirical analysis of the deviation and the regression inconsistency. Section 7 dwells on the variance properties of the deviation in probability sampling from a finite population. A variance estimator is derived. Section 8 asks: Is the nonresponse bias important enough to worry about? Significance testing of the deviation is carried out. A discussion concludes the article in Section 9.

It should be noted that the empirical work in this article is carried out with "pseudo y-variables", known for the full probability sample, rather than with "real y-variables", which are available under nonresponse only for the responding subset. By the issues raised in this article, we wish to communicate an awareness of the effects of nonresponse on quantities of vital interest in a survey. The techniques that we outline can be helpful for methodological evaluation of nonresponse and its effects on important surveys in a national statistical institute.

2. Notation

We consider a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$; let s be a probability sample from U, drawn so that unit k has the known inclusion probability π_k and the sampling weight $d_k = 1/\pi_k$. The response set r, where $r \subset s$, is the set of units for which the value y_k of the survey variable y (continuous or categorical) is observed. The target of estimation is the total

$$Y = \sum_U y_k = N\bar{y}_U.$$

For example, $\sum_A y_k$, in displayed equations $\sum_A y_k$, is used for $\sum_{k\in A} y_k$ or $\sum_{k\in A} y_k$. Summation over a set of units $k \in A \subseteq U$ is written compactly as $\sum_A$. For example, $\sum_A y_k$ is used for $\sum_{k\in A} y_k$ or $\sum_{k\in A} y_k$.

The (design weighted) response rate in the sample s is

$$P = \sum_r d_k \Big/ \sum_s d_k. \tag{1}$$

In an equal probability sampling design, where d_k is constant, then $P = m/n$, the ratio of the respective sizes of r and s. The data collection is viewed as a dynamic process; interventions may take place; the rate P is evolving during the period which may last several weeks.

Access to auxiliary variables is indispensable (a) for a monitored data collection, (b) for bias adjustment at the estimation stage. Denote by $\mathbf{x}$ an auxiliary vector with value $\mathbf{x}_k$ known for $k \in s$ or $k \in U$. For some variables x in the vector $\mathbf{x}$, the values x_k come from available (population) registers. The role of paradata should be emphasized, namely, when x_k, known for the sample units alone, may reflect some aspect of how the data collection is done – the interviewing for example – for the given sample s, or some property readily observed for the sample units but only for those. The x-variables in this article are primarily categorical, as is often the case in statistical agencies.

Means (sample mean, response mean) of y and of $\mathbf{x}$ are important in the following; they are:

$$\bar{y}_s = \sum_s d_k y_k \Big/ \sum_s d_k; \quad \bar{y}_r = \sum_r d_k y_k \Big/ \sum_r d_k;$$
$$\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \Big/ \sum_s d_k; \quad \bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \Big/ \sum_r d_k. \tag{2}$$

Of these, $\bar{y}_s$ is conceptually defined but not computable in practice; y_k is missing for $k \in s - r$. We assume that $\mathbf{x}$ satisfies an "$\mathbf{x}$-vector condition", invoked in several of the derivations: One can identify a vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \quad \text{for all } k. \tag{3}$$

It brings mathematical convenience at no significant loss of generality. For example, when $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ is a *group vector*, one where the single "1" points out membership in one of a number of exclusive and exhaustive categories, then $\boldsymbol{\mu} = (1, \ldots, 1, \ldots, 1)'$ satisfies the requirement. But more often $\mathbf{x}$ is not a group vector, as in the simple example when $\mathbf{x}$, with dimension four, codes three exhaustive categories of "education", and "gender" is coded as a univariate 0/1 variable, then $\boldsymbol{\mu} = (1, 1, 1, 0)'$ satisfies the requirement.

3. Estimation for a Survey Variable Observed for Only the Responding Subset

3.1. Estimator by Calibration, or Alternatively by Regression/Prediction

To estimate the total $Y = \sum_U y_k = N \bar{y}_U$ we make use of the values $\mathbf{x}_k$, known for $k \in s$, on a chosen auxiliary vector $\mathbf{x}$. This gives the *calibration estimator*

$$\hat{Y}_{CAL} = (1/P) \sum_r d_k g_k y_k,$$

where

$$g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k; \quad \boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_r d_k \tag{4}$$

and $\bar{\mathbf{x}}_s$ is given in (2). The name reflects the *calibration property* of the weights $d_k g_k$ with respect to the vector $\mathbf{x}$,

$$(1/P) \sum_r d_k g_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k.$$

The Horvitz-Thompson estimator on the right hand side of the equation is unbiased for the population x-total $\sum_U \mathbf{x}_k$. Therefore the calibration feature of the weights $d_k g_k$ tends to reduce the bias that would affect less attractive alternatives, such as the naive expansion of the respondent y-mean,

$$\hat{Y}_{EXP} = (1/P) \sum_r d_k y_k = \hat{N} \bar{y}_r$$

with $\hat{N} = \sum_s d_k$. The calibration argument makes no explicit reference to a capacity of $\mathbf{x}$ to "explain" y; in a purely technical sense, $\mathbf{x}$ is just a vector that one has decided to calibrate weights on.

But in an alternative derivation of $\hat{Y}_{CAL}$, the regression/prediction approach, the relationship between $\mathbf{x}$ and y is central: We fit the regression of the survey variable y on the auxiliary $\mathbf{x}$ for the response r. This "response fit" is feasible because y_k is available for $k \in r$. It yields a regression vector $\mathbf{b}_r$ and predictions $\hat{y}_k$ extrapolated to the whole sample s because $\mathbf{x}_k$ is available for $k \in s$:

$$\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_r d_k \mathbf{x}_k y_k; \quad \hat{y}_k = \mathbf{x}_k' \mathbf{b}_r; \quad k \in s \tag{5}$$

Then we have $\sum_s d_k \hat{y}_k = \hat{N} \bar{\mathbf{x}}_s' \mathbf{b}_r = \hat{Y}_{CAL}$. Yet another construction with the same end result is "observations y_k for the response r, predictions $\hat{y}_k = \mathbf{x}_k' \mathbf{b}_r$ for the nonresponse $s - r$":

$$\sum_r d_k y_k + \sum_{s-r} d_k \hat{y}_k = \sum_s d_k \hat{y}_k = \hat{N}\, \bar{\mathbf{x}}_s' \mathbf{b}_r = \hat{Y}_{CAL}.$$

This follows from the $\mathbf{x}$-vector condition (3).

A critical analyst may note, quite appropriately, that a weakness of the regression/prediction approach is that a regression fitted on the response may not well represent a regression holding in the sample. This is a central issue for this article.

In theoretical study, although not in practice, we can confront the biased estimators $\hat{Y}_{CAL}$ and $\hat{Y}_{EXP}$ with the unbiased Horvitz-Thompson estimator of Y, requiring full response and given by

$$\hat{Y}_{FUL} = \sum_s d_k y_k = \hat{N} \bar{y}_s.$$

We refer to those estimator types as CAL, EXP and FUL. Our main interest is in the difference between CAL and FUL. Can calibration on $\mathbf{x}$ "fill the gap"?

3.2. *Deviation of the Calibration Estimator from the Unbiased Estimator Requiring Full Response*

The linear regression fit of y on $\mathbf{x}$ is not feasible for the sample s, because that would require y_k for all of s. Conceptually, this "sample fit" is, however, important, with coefficient vector and residuals given by

$$\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_s d_k \mathbf{x}_k y_k; \quad e_k = y_k - \mathbf{x}_k' \mathbf{b}_s, \quad k \in s. \tag{6}$$

Although the CAL estimator gains strength from the calibrated weighting, it is not unbiased, but ordinarily less biased, often much less so, than the naive EXP estimator. The bias is reflected in the CAL estimator's deviation from the unbiased FUL. The (scaled) deviation is $(\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{N} = \Delta_r$ with

$$\Delta_r = \bar{\mathbf{x}}_s' \mathbf{b}_r - \bar{y}_s = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s. \tag{7}$$

Note that $\bar{y}_s = \bar{\mathbf{x}}_s' \mathbf{b}_s$ by (3). We seek to find out more about size and other properties of Δ_r when the response is managed, in data collection for the given sample, to make it better

balanced. That is, s is fixed, r is manipulated. We call Δ_r simply "the deviation", short for "the deviation (normed by $\hat{N}$) of the CAL estimator, built on a specified calibration vector $\mathbf{x}$, from the unbiased FUL estimator". As the expression for Δ_r tells, the deviation is caused by the difference between two linear regressions, more specifically the vector difference $\mathbf{b}_r - \mathbf{b}_s$. The deviation Δ_r is unknown in practice, can practically never be assumed negligible, and can be large.

3.3. Regression Inconsistency as an Important Part of the Deviation

The sample mean of the sample fit residuals $e_k = y_k - \mathbf{x}_k'\mathbf{b}_s$ given in (6) is zero:

$$\bar{e}_s = \sum_s d_k e_k \Big/ \sum_s d_k = \bar{y}_s - \bar{\mathbf{x}}_s'\mathbf{b}_s = 0.$$

The response mean of those same residuals is non-zero:

$$\bar{e}_r = \sum_r d_k e_k \Big/ \sum_r d_k = \bar{y}_r - \bar{\mathbf{x}}_r'\mathbf{b}_s \neq 0. \tag{8}$$

That the residual mean $\bar{e}_r$ is non-zero is an expression of the selective, non-random nature of the response r. It also expresses that the regression model for the sample does not hold for the response. The sample regression (6) is *inconsistent* for the response, and $\bar{e}_r$ serves to measure this.

Noting that $\bar{y}_r = \bar{\mathbf{x}}_r'\mathbf{b}_r$, a consequence of (3), we can write the regression inconsistency as

$$\bar{e}_r = \sum_r d_k e_k \Big/ \sum_r d_k = (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_r, \tag{9}$$

and the deviation Δ_r in (7) as the sum of inconsistency and a residual,

$$\Delta_r = \bar{e}_r + u_r \tag{10}$$

where $u_r = -(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s)$.

One reason for calling u_r "residual" is that $u_r = 0$ if the response is balanced to have $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$; otherwise, u_r is likely to be a minor, although non-negligible, part of the deviation Δ_r.

None of Δ_r, $\bar{e}_r$ and u_r is computable in a real survey, because all require y_k for the entire sample s. We can study the three terms and their relative importance, theoretically and empirically, by choosing y to be a variable with values known for the whole sample, a "pseudo y-variable". Such studies can be carried out in statistical agencies as an important way to get insight into the effect of nonresponse on their most important surveys. Here we use data from two Statistics Sweden surveys that employ a number of register variables known for the sample s. We choose some of those to be x-variables (that is, auxiliary) and designate a couple of them to play the role of y-variables (that is, they are pseudo y-variables).

The linear regressions (5) and (6) suggest implicitly that y is a continuous rather than a categorical variable; in the latter case, nonlinear regression, logistic or other, would be favoured as "more realistic modeling". However, important formulas in this Section 3 and

later, which contain quantities such as $\bar{\mathbf{x}}_s'\mathbf{b}_s = \bar{y}_s$, $\bar{\mathbf{x}}_r'\mathbf{b}_r = \bar{y}_r$, $\bar{\mathbf{x}}_s'\mathbf{b}_r$ and $\bar{\mathbf{x}}_r'\mathbf{b}_s$, continue to be well defined and computable also when y_k is dichotomous 0/1. The empirical Sections 6 to 8 use these formulas both for continuous and categorical y_k.

One should note that trust in an assumed model (linear or nonlinear regression) for the relationship between y and $\mathbf{x}$ risks to be misleading. If m is the simple linear model stating that $E_m(\varepsilon_k|\mathbf{x}_k) = 0$ for $k \in s$, where $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, then $E_m(\mathbf{b}_r|r,s) = E_m(\mathbf{b}_s|s) = \boldsymbol{\beta}$ and therefore $E_m(\Delta_r|r,s) = E_m(\bar{e}_r|r,s) = 0$. But expecting Δ_r and $\bar{e}_r$ to be zero is misleading, because zero expected residual ε_k is unrealistic for the nonresponse; the model with a $\boldsymbol{\beta}$ common to the response and the sample fails under selective nonresponse. Non-linear modeling attempts would have the same weakness.

The adaptive data collection objective to make $\bar{\mathbf{x}}_r$ close to $\bar{\mathbf{x}}_s$ is attractive, but it is not evident that this would reduce the difference $\mathbf{b}_r - \mathbf{b}_s$ present both in the deviation Δ_r and the inconsistency $\bar{e}_r$. We set out to find out more about this question.

4. Response Imbalance, Response Propensity, and Their Effect on the Regression Inconsistency

4.1. The Response Imbalance Statistic

Response sets r differ in regard to their representativeness, or balance, with respect to a specified vector $\mathbf{x}$. The response r has *perfect balance* if $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. When $\bar{\mathbf{x}}_r \neq \bar{\mathbf{x}}_s$, the response r is imbalanced, more or less. Both means are computable, and so is their difference $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$. We measure the *imbalance* of r given s, with respect to $\mathbf{x}$, by the scalar quantity

$$IMB = P^2 Q_s; \quad Q_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s); \quad \boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_s d_k. \tag{11}$$

The weighting matrix $\boldsymbol{\Sigma}_s$ is computable and assumed non-singular. One can show that $0 \leq IMB \leq P(1-P)$ whatever r, s and $\mathbf{x}$. The imbalance IMB addresses simultaneously all the auxiliary x-variables in $\mathbf{x}$. This is in contrast to a nonresponse analysis often done in practice one x-variable at a time. But IMB does not involve any of the perhaps numerous survey variables y. Reducing IMB to zero does *not eliminate* bias for any one y-variable, but may *reduce* bias for a majority of the y-variables.

Adaptive data collection can grant a degree of closeness of $\bar{\mathbf{x}}_r$ to $\bar{\mathbf{x}}_s$, a relatively low IMB. One approach is the threshold method used for the empirical work in later sections, see Särndal and Lundquist (2014). To achieve perfect balance, $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, is unlikely in practice.

How does an effort to reduce IMB to some degree in data collection affect the quantities Δ_r, $\bar{e}_r$ and u_r in (10) and the relation between them? Is the deviation Δ_r and the regression inconsistency $\bar{e}_r$ significantly reduced by adaptive data collection? We shall examine these questions. The part of the CAL estimator deviation Δ_r attributable to regression inconsistency is $\bar{e}_r/\Delta_r = 1 - (u_r/\Delta_r)$. We shall examine this ratio, which depends on the survey variable y, on the choice of variables for the auxiliary vector $\mathbf{x}$ and on the degree of imbalance IMB.

4.2. Imbalance Interpreted Through Response Propensity

The response imbalance *IMB* is tied to *response propensity* in the following way: For $k \in s$, define

$$I_k = i_k/P; \quad f_k = \bar{\mathbf{x}}_r' \mathbf{\Sigma}_s^{-1} \mathbf{x}_k \tag{12}$$

where $i_k = 1$ for responding unit ($k \in r$) and $i_k = 0$ for nonresponding unit ($k \in s - r$), $\mathbf{\Sigma}_s$ is given in (11), and P in (1). Their relationship is that f_k is the predicted value of I_k, which is a response indicator normed to have mean 1 over s. That is, $f_k = \hat{I}_k$, the predicted value for unit k from the linear regression of I_k on $\mathbf{x}_k$ over s. (See proof in Appendix 1, part (a).) Despite certain resemblance, f_k must not be confused with the weight factor g_k in the CAL estimator (4).

The quantity Pf_k is a measure of response propensity for unit k characterized by the auxiliary value $\mathbf{x}_k$. The mean of the propensities Pf_k is the sample response rate: $P\bar{f}_s = P$. The relationship with imbalance lies in the sample variance: $P^2 S_{fs}^2 = IMB$ given in (11). (To see this, use that $\bar{f}_s = \sum_s d_k f_k / \sum_s d_k = 1$; $S_{fs}^2 = \sum_s d_k (f_k - \bar{f}_s)^2 / \sum_s d_k = Q_s$, as shown in Appendix 1, part b.) The connection between *IMB* and the variance of the propensities motivates methods to produce a final response r with low *IMB*: In the threshold method explained later, the values Pf_k are computed and examined at given points in the data collection period; by appropriate interventions, their variance – and therefore *IMB* – is reduced.

4.3. Regression Inconsistency Analyzed in Terms of Imbalance

We examine now the effect, if any, of a reduced imbalance *IMB* on the regression inconsistency (9), which we write as $\bar{e}_r = T_1 - T_2$, where

$$T_1 = \sum_s d_k (I_k - 1) y_k \Big/ \sum_s d_k = \bar{y}_r - \bar{y}_s;$$

$$T_2 = \sum_s d_k (f_k - 1) y_k \Big/ \sum_s d_k = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \mathbf{b}_s.$$

with I_k and f_k defined in (12).

In the ideal case of full response, $r = s$, we would have $T_1 = T_2 = \bar{e}_r = \Delta_r = 0$. Under nonresponse, but still a perfectly balanced response, so that $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, then $T_2 = 0$ but $T_1 \neq 0$ so $\Delta_r = \bar{e}_r = \bar{y}_r - \bar{y}_s \neq 0$. In practice, we will not achieve perfect balance but can work in that direction.

An adaptive data collection that brings a response with very low imbalance will likely reduce T_2 considerably and possibly T_1 to some extent, resulting in a reduction, possibly modest, of the deviation Δ_r and of the regression inconsistency $\bar{e}_r = T_1 - T_2$. This is illustrated in the empirical work in later sections.

The term $T_1 = \bar{y}_r - \bar{y}_s$ is stated in the y-variable alone. Seemingly it has nothing to do with the auxiliary vector $\mathbf{x}$, but this is not quite true. The $\mathbf{x}$-vector is instrumental in bringing an ultimate response r with, preferably, a low imbalance *IMB*, and $\bar{y}_r$ may come closer to $\bar{y}_s$.

The correlation between I_k and f_k is positive and given by

$$corr(I,f) = \frac{S_{If\,s}}{S_{I\,s}S_{f\,s}} = \left(\frac{IMB}{P(1-P)}\right)^{1/2}, \tag{13}$$

where $S_{I\,s}$ and $S_{f\,s}$ are the respective standard deviations of I_k and f_k, and $S_{If\,s}$ is their covariance, all three over s. (The proof of (13) is given in Appendix 1, part (b).)

For a chosen vector $\mathbf{x}$, we have methods to reduce *IMB* and $corr(I,f)$ during data collection. The objective is not high but low $corr(I,f)$, because it would mean that the response coded by I_k is nearly unrelated to the auxiliary vector. This is advantageous, although short of the ideal where I_k would be unrelated to the y-variable itself: That would eliminate the nonresponse bias. But the objective takes a step in the right direction.

One of the techniques in adaptive data collection is *prioritization*, namely of units observed to have distinctly low response propensity, at some point in the data collection period. This motivates spending extra effort on getting response from precisely those units, that is, to boost their low response propensity Pf_k towards an average response propensity, thereby reducing the propensity variance $P^2S_{fs}^2 = IMB$ and the correlation $corr(I,f)$.

Even without any attempt in data collection to balance the response, the correlation is modest in most surveys, typically $corr(I,f) < 0.3$. It can be considerably reduced by a reduced imbalance, as Table 1 in Section 6 illustrates. To get $corr(I,f) = 0$ is tantamount to "random response with respect to the auxiliary vector $\mathbf{x}$" (although not with respect to the survey variable y itself).

4.4. Insensitivity to the Study Variable

The deviation Δ_r and the regression inconsistency $\bar{e}_r$, given in (10), depend on the survey variable y. So does the ratio $\bar{e}_r/\Delta_r$, but it is expected to be quite insensitive to y. To see this, express Δ_r, $\bar{e}_r$ and u_r as weighted y-means over the sample s: For $k \in s$, let

$$G_k = I_k g_k - 1; \quad F_k = I_k - f_k, \tag{14}$$

where I_k, f_k and g_k are given in (12) and (4). Then

$$\begin{aligned} \Delta_r &= \sum_s d_k G_k y_k \Big/ \sum_s d_k; \quad \bar{e}_r = \sum_s d_k F_k y_k \Big/ \sum_s d_k; \\ u_r &= \sum_s d_k (G_k - F_k) y_k \Big/ \sum_s d_k \end{aligned} \tag{15}$$

These three equations hold also with y_k replaced by $e_k = y_k - \mathbf{x}_k'\mathbf{b}_s$ because $\sum_s d_k G_k \mathbf{x}_k' = \mathbf{0}'$, $\sum_s d_k F_k \mathbf{x}_k' = \mathbf{0}'$. The correlation between F_k and G_k is high positive. A derivation (see Appendix 1, part (c)) shows that

$$corr(F,G) = \frac{S_{FGs}}{S_{Fs}S_{Gs}} = \frac{1-P}{\{(1-P-PQ_s)(1-P+Q_r)\}^{1/2}} \tag{16}$$

where S_{Fs} and S_{Gs} are the respective standard deviations of F_k and G_k, and S_{FGs} their covariance, all three over s, Q_s is given in (11), and Q_r is the "dual quadratic form",

$$Q_r = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_r^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s) \tag{17}$$

with weighting matrix $\mathbf{\Sigma}_r$ given in (4). The correlation (16) is high positive – greater than 0.95 for the data sets we worked with – because for a nonresponse $1 - P$ of 30% or more, the term $1 - P$ dominates $PQ_s \geq 0$ and $Q_r \geq 0$. Typically, Q_s and Q_r are less than 0.10, Q_r usually the somewhat larger. Now let us approximate the ratio $\bar{e}_r/\Delta_r$ as

$$\frac{\bar{e}_r}{\Delta_r} = \frac{\sum_s d_k F_k e_k}{\sum_s d_k G_k e_k} \approx \frac{\sum_s d_k \hat{F}_k e_k}{\sum_s d_k G_k e_k}$$

where F_k is replaced by its predicted value, $\hat{F}_k$, from the linear regression fitted over s of F_k on G_k, quite accurate in view of the high correlation (16). The slope coefficient (see Appendix 1, part (c)) is $b_{FG} = S_{FGs}/S^2_{Gs} = (1 - P)/(1 - P + Q_r)$, slightly less than 1 and independent of the y-variable. The intercept is zero because F_k and G_k have mean zero over s. The predicted values for $k \in s$ are $\hat{F}_k = b_{FG} G_k$. Then, independently of the y-variable,

$$\bar{e}_r/\Delta_r \approx b_{FG} = (1 - P)/(1 - P + Q_r) \leq 1 \tag{18}$$

Although (18) is an approximation with non-negligible error, it suggests a ratio $\bar{e}_r/\Delta_r$ somewhat less than one and little sensitive to the y-variable. We illustrate this empirically in Section 6. If we compare two $\mathbf{x}$-vectors, for fixed r and s, then (18) suggests that the one with the larger Q_r will give the smaller ratio $\bar{e}_r/\Delta_r$. Larger Q_r and smaller ratio $\bar{e}_r/\Delta_r$ ordinarily happen when a given vector $\mathbf{x}$ is extended by adding more x-variables to it.

5. Background of the Empirical Work

5.1. Hypotheses to Be Tested

The results in Sections 3 and 4 generate several hypotheses for empirical testing. The objective is to examine how a decreased imbalance *IMB* may impact on quantities such as Δ_r, $\bar{e}_r$ and u_r. As a part of this we can contrast *the single usage* of the auxiliary vector (only at the estimation stage, for calibrated weighting) with *the combined usage* (where that same vector serves first in data collection to get reduced imbalance, then re-used at the estimation stage for calibrated weighting). Here we get some insight into the question: Are efforts in adaptive data collection to reduce imbalance worth it, from an improved accuracy standpoint? The data used for this empirical work come from two important Statistics Sweden surveys.

When *IMB* is made to decrease, we expect to see these effects (where "decrease" and "smaller than" should be understood as "in absolute value", since Δ_r, $\bar{e}_r$ and u_r can have either sign):

(1) All of Δ_r, $\bar{e}_r$ and u_r will go decreasing;
(2) The most pronounced decrease will occur in the residual term u_r known to be zero when *IMB* is zero;
(3) Both the regression inconsistency $\bar{e}_r$ and the deviation Δ_r are likely to decrease to some extent, but remain distinctly non-zero even if *IMB* is reduced to near zero;
(4) The part of the deviation due to inconsistency, $\bar{e}_r/\Delta_r$, is expected to remain high, somewhat less than one, for all levels of *IMB*, to approach 1 when *IMB* approaches zero, and to be rather insensitive to the survey variable y;

(5) Although $\bar{e}_r/\Delta_r < 1$ does not necessarily hold, it is likely in a majority of cases.

Point (3) says that efforts to reduce *IMB* to low levels is an incomplete cure for nonresponse bias and for regression inconsistency. There may be positive effect, but not complete remedy.

5.2. Preparing the Data Sets for the Empirical Work

We use data from two important Statistics Sweden surveys, the Living Conditions Survey in its 2009 edition, and the Labour Force Survey in its 2012 edition. Facts about these surveys are given in Appendix 2.

From the first of these sources, we prepared the data set called LCS2009 of size 8,220. It is a subset of the entire 2009 Living Conditions Survey sample.

From the monthly Labour Force Surveys in 2012, we prepared the data set called LFS2012 of size 32,265. It is composed of parts of the data from the twelve monthly surveys in 2012. Essentially, we combined the twelve wave-one samples to arrive at LCS2012; see Appendix 2 for further detail.

Both LCS2009 and LFS2012 contain a number of variables for every unit, some used as x-variables and two used as pseudo y-variables. For the exercise here, we treat both data sets as simple random samples, so the sampling weight d_k is constant in the formulas used.

For both LCS2009 and LFS2012, we specified five alternative response sets representing different levels of imbalance *IMB*. The first of these, denoted r_{Act}, is the actual response recorded in the survey data collection. The other four are experimental response sets, generated with the threshold method, described in Subsection 5.3 and in more detail in Särndal and Lundquist (2014). This permits us to see the effect of imbalance on quantities of interest defined in the preceding theory sections.

For the computations, an $\mathbf{x}$-vector is specified for each of the two data sets. The $\mathbf{x}$-vector serves to obtain the four generated response sets, through the threshold method described below. For those four response sets, the same $\mathbf{x}$-vector also serves to compute quantities relevant for the estimation stage, the deviation Δ_r, the regression inconsistency $\bar{e}_r$ and others. This is the *combined usage* of the x-vector referred to in Subsection 5.1, namely, both in data collection and in estimation. In contrast to this, with the actual response set, this $\mathbf{x}$-vector is used in estimation but not in data collection; it is the *single usage* of the vector.

In LCS2009, the first response set, denoted r_{Act}, is composed of those 67.9% of the 8,220 persons who actually responded in the 2009 data collection. The $\mathbf{x}$-vector is specified as

$$\mathbf{x} = \mathbf{x}_{\text{dm14}} = ((Educ \times Owner \times Origin), Phone, Age, Civil, Gender) \qquad (19)$$

The vector $\mathbf{x}_{\text{dm14}}$ consists entirely of categorical variables, chosen from the rich supply of potential x-variables. The same vector $\mathbf{x}_{\text{dm14}}$ was used to compute quantities of interest such as *IMB*, Δ_r, $\bar{e}_r$ and u_r.

In $\mathbf{x}_{\text{dm14}}$, *Educ, Owner* and *Origin* are binary (valued 1 or 0) and crossed, accounting for $2^3 = 8$ vector positions, *Phone*, *Civil*, and *Gender* are binary; *Age* occupies three vector positions for four exhaustive categories. This gives $\mathbf{x}_{\text{dm14}}$ the dimension

$2^3 + 1 + 3 + 1 + 1 = 14$ and 256 possible values, representing that many characteristics of the 8,220 sample units. *Educ* is 1 for person with high education, 0 otherwise; *Owner* is 1 for a person who owns his place of residence, 0 otherwise; *Origin* is 1 for a person born in Sweden, 0 otherwise; *Phone* is 1 for a person with phone number accessible at the start of the data collection, 0 otherwise; *Civil* is 1 for married or widower, 0 otherwise; *Gender* is 1 for male, 0 otherwise. *Age*, with four exhaustive age brackets, 24 and under, 25–64, 65–74, 75 and over is coded as (1,0,0), (0,1,0), (0,0,1), or (0,0,0).

The four experimental response sets, r_{A65}, r_{A60}, r_{A55} and r_{A50}, were generated from r_{Act} by the threshold method, so as to have successively lower imbalance *IMB*. The notation refers to the different thresholds applied in the method, 65%, 60%, 55%, and 50%. The intervention points were attempts 3, 6 and 9 of the ordinary data collection, the end of the ordinary data collection, and attempt 4 of the follow-up. Two pseudo *y*-variables are specified in LCS2009: *Income*, available from the Swedish tax register, and *Employed* (binary; 1 for employed person, 0 otherwise). The coefficient of determination (R-square) for *Income* regressed on $\mathbf{x}_{\text{dm14}}$ is $R^2 = 0.28$, thus not exceptionally high.

In LFS2012, the first response set, r_{Act}, is composed of those 70.6% of the 32,265 persons who had actually responded in the LFS survey. The **x**-vector was chosen as

$$\mathbf{x} = \mathbf{x}_{\text{dm16}} = ((Age \times Educ \times Owner), Origin, Urban, Civil, Gender) \qquad (20)$$

The vector $\mathbf{x}_{\text{dm16}}$ has dimension $(3 \times 2 \times 2) + 1 + 1 + 1 + 1 = 16$; the vector codes 192 different characteristics of the 32,265 persons. Here *Age* is defined by three exhaustive brackets, 34 and under, 35–64, 65 and over, rather than four as in LCS2009. This is because the upper age limit of the LFS target population is 74, whereas LCS has no age limit. *Urban* is 1 for big city dweller, 0 otherwise. The other variables in $\mathbf{x}_{\text{dm16}}$ are as in $\mathbf{x}_{\text{dm14}}$ defined above for LCS2009. The pseudo *y*-variables are *Income* and *Employed*, as earlier defined. The coefficient of determination (R-square) for *Income* regressed on $\mathbf{x}_{\text{dm16}}$ is $R^2 = 0.27$.

The four experimental response sets, r_{A68}, r_{A65}, r_{A63} or r_{A60} were generated from r_{Act} by the threshold method to have successively lower imbalance *IMB*. The notation refers to the thresholds used, 68%, 65%, 63%, and 60%. We used eight intervention points: Call attempts 3, 5, 7, 9, 12, 15, 18, and 22. The two pseudo y-variables specified in LFS2012 are *Income* and *Employed* (binary; 1 for employed person, 0 otherwise). To facilitate comparison, these *y*-variables are the same as for LCS2009.

It should be emphasized that the choice of **x**-vector, (19) and (20) respectively, was not made to meet any criterion of theoretical optimality, if one exists. The vectors simply represent reasonable choices of auxiliary variables for the two surveys; they are well adapted to the computations in this article; most of the variables are used by Statistics Sweden methodologists in weight calibration for the two surveys. Other choices of **x**-vector could certainly be entertained.

The steps of the threshold method are outlined in the following Subsection 5.3. The reader more interested in empirical results may proceed to Section 6.

5.3. The Steps of the Threshold Method

Briefly, the steps in the threshold method are as follows: The method operates on the WinDATI file of "events data", that is, the series of all call attempts stored for every unit in the sample. WinDATI is Statistics Sweden's telephone data collection system. Specified at the outset are (i) the intervention points, say just after attempts 3, 6, 9, . . .; (ii) the threshold value, say 60% and (iii) the monitoring vector, that is, the **x**-vector for computing the response propensity Pf_k, with P and f_k given in (1) and (12). (In this article, the same vector **x** serves also at the estimation stage for calibrated weights computation.) At each intervention point, Pf_k is computed for $k \in s$ based on the specified vector **x**. Units with Pf_k greater than the specified threshold are set aside; it is pretended that these high propensity units are not subject to any further call attempts, thus we drop them from the actual response r_{Act} and pretend that the call attempts continue with the remaining lower propensity units. Specifically, at the first intervention point, Pf_k is computed (on the response set present at that moment) for all sample units $k \in s$; those with Pf_k greater than the threshold are set aside at that point; call attempts continue with the rest. At the second intervention point, Pf_k is recomputed (but on the somewhat larger response set then at hand) for all $k \in s$, including the units set aside at the first point. (Those latter units may have their propensity somewhat changed, but they remain aside.) Among the units still in contention, those with new propensity Pf_k greater than the same fixed threshold are set aside, and so on at the remaining points. Thus at every intervention point, some more units leave the actual response r_{Act}. (In this experiment we must drop units from the actual response; we are not in a position to add any new incoming respondents.) Units remaining at the last intervention point are pursued until the very end of the data collection period. A feature of this construction is that the variability of the propensities becomes more and more reduced, and since *IMB* is the variance of the propensities, the resulting final response gets to have lowered imbalance *IMB*. The effect of the chosen threshold is that lower threshold accentuates the decrease in *IMB*.

6. Descriptive Analysis for the Living Conditions Survey and the Labour Force Survey

Tables 1, 2, and 3 present results computed with formulas in Sections 3 and 4 on the data sets LCS2009 and LFS2012, each considered as a simple random sample s, of size $n = 8{,}220$ for the first, $n = 32{,}265$ for the second; all design weights d_k are equal. The

Table 1. Imbalance IMB (multiplied by 10^2) and corr = corr(I,f) between response indicator and response propensity, computed on LCS2009 and on LFS2012, for five response sets.

LCS2009, vector $\mathbf{x}_{\text{dm14}}$				LFS2012, vector $\mathbf{x}_{\text{dm16}}$			
Resp.set	Rate P	*IMB*	*corr*	Resp.set	Rate P	*IMB*	*corr*
$\boldsymbol{r_{Act}}$	0.679	1.878	0.294	$\boldsymbol{r_{Act}}$	0.706	0.788	0.195
$\boldsymbol{r_{A65}}$	0.633	1.120	0.220	$\boldsymbol{r_{A68}}$	0.674	0.251	0.107
$\boldsymbol{r_{A60}}$	0.606	0.890	0.193	$\boldsymbol{r_{A65}}$	0.659	0.158	0.084
$\boldsymbol{r_{A55}}$	0.569	0.648	0.163	$\boldsymbol{r_{A63}}$	0.648	0.128	0.075
$\boldsymbol{r_{A50}}$	0.533	0.426	0.131	$\boldsymbol{r_{A60}}$	0.625	0.061	0.051

Table 2. Analysis, LCS2009 data set, vector $\mathbf{x}_{\text{dm14}}$*, five response sets (actual, and four generated): IMB (multiplied by* 10^2*); deviation* Δ_r*, regression inconsistency* $\bar{e}_r$*, residual* u_r *(all three multiplied by* 10^{-3} *for Income, by* 10^2 *for Employed), and ratio* $\bar{e}_r/\Delta_r$*.*

		Income				Employed			
Resp.set	*IMB*	Δ_r	$\bar{e}_r$	u_r	$\bar{e}_r/\Delta_r$	Δ_r	$\bar{e}_r$	u_r	$\bar{e}_r/\Delta_r$
r_{Act}	1.878	7.46	5.00	2.46	0.670	2.08	1.47	0.61	0.707
r_{A65}	1.120	7.51	5.46	2.05	0.727	2.01	1.47	0.54	0.729
r_{A60}	0.890	6.99	5.10	1.89	0.730	1.92	1.40	0.52	0.732
r_{A55}	0.648	6.16	4.38	1.78	0.711	1.97	1.51	0.46	0.768
r_{A50}	0.426	5.20	3.75	1.45	0.721	1.75	1.33	0.42	0.761

rows refer to the five response sets r. For LCS2009 they are r_{Act} (actual), and r_{A65}, r_{A60}, r_{A55} and r_{A50} (generated), for LFS2012 they are r_{Act} (actual) and r_{A68}, r_{A65}, r_{A63} and r_{A60} (generated). The actual response, $r = r_{Act}$, is 67.9% of the sample s for LCS2009, compared with 70.6% for LFS2012. The generated sets grow successively smaller, because the threshold method, as applied here, proceeds by dropping high propensity units from the actual response. Hence, a drop in the response rate, $P = m/n$, where m is the size of r.

The **x**-vector for computing Tables 1 to 3 is $\mathbf{x}_{dm14}$ given in (19) for LCS2009, $\mathbf{x}_{dm16}$ given in (20) for LFS2012. We cannot fully compare LCS2009 with LFS2012, because the x-vectors are not identical and the designs are different.

In Table 1, the imbalance *IMB* and the correlation $corr(I,f)$ are computed by (11) and (13), respectively. The results confirm the anticipation that the sequence of five response sets (the five rows) will bring a drop in *IMB* and in $corr(I,f)$. But there is a pronounced difference between LCS2009 with LFS2012. For the former, *IMB* is still fairly high in the bottom line, whereas for the latter, it is very low. This may hint at a data collection in some ways better adapted, or more accomplished, in the Labour Force Survey than in the Living Conditions Survey, but there is no concrete evidence to this effect; in any case, the **x**-vectors are not fully identical.

Table 2 presents results for the LCS2009 data set, viewed as a sample s of size 8,220, for the two pseudo y-variables, *Income* (continuous) and *Employed* (categorical, 1 or 0). Of principal interest is to track the development of the deviation Δ_r, the regression inconsistency $\bar{e}_r$, and the ratio $\bar{e}_r/\Delta_r$ as the imbalance *IMB* is reduced over the five response sets.

Table 3. Analysis, LFS2012 data set, vector $\mathbf{x}_{\text{dm16}}$*, five response sets (actual, and four generated): IMB (multiplied by* 10^2*); deviation* Δ_r*, regression inconsistency* $\bar{e}_r$*, residual* u_r *(all three multiplied by* 10^{-3} *for Income, by* 10^2 *for Employed), and ratio* $\bar{e}_r/\Delta_r$*.*

		Income				Employed			
Resp.set	*IMB*	Δ_r	$\bar{e}_r$	u_r	$\bar{e}_r/\Delta_r$	Δ_r	$\bar{e}_r$	u_r	$\bar{e}_r/\Delta_r$
r_{Act}	0.788	3.73	3.02	0.71	0.810	1.30	1.14	0.16	0.878
r_{A68}	0.251	2.82	2.34	0.48	0.829	1.19	1.07	0.12	0.899
r_{A65}	0.158	2.37	1.95	0.43	0.821	1.09	0.98	0.11	0.901
r_{A63}	0.128	1.98	1.62	0.36	0.818	1.01	0.93	0.09	0.915
r_{A60}	0.061	1.58	1.35	0.22	0.858	0.89	0.83	0.06	0.935

As *IMB* drops over the series of five response sets (the rows), there is a drop (with one minor dent for *Employed*) in the deviation Δ_r and in the regression inconsistency $\bar{e}_r$. While the reduction in *IMB* is important, the reduction in Δ_r and in $\bar{e}_r$ is modest by comparison.

To illustrate, *IMB* is reduced by a factor (highest to lowest) of 1.878/0.426 = 4.40. The corresponding reduction of the deviation Δ_r is 7.46/5.20 = 1.43 (for *Income*) and 2.08/1.75 = 1.19 (for *Employed*). *Income* gains relatively more from balancing than *Employed*, but for both, the large reduction in *IMB* is accompanied by rather modest improvement in accuracy (modestly lower Δ_r). This is not entirely surprising, considering the conclusion from theory in Subsection 4.3 that Δ_r will not approach zero even if *IMB* comes close to zero. Modest expectations for better accuracy are indicated, rather than hopes for greatly improved accuracy.

The reduction is similar for the regression inconsistency $\bar{e}_r$: The reduction factor is 5.00/3.75 = 1.33 (for *Income*) and 1.47/1.33 = 1.11 (for *Employed*). The residual term u_r has a higher reduction factor.

The regression inconsistency portion of the deviation, the ratio $\bar{e}_r/\Delta_r$, shows a mild increase over the five decreasing levels of *IMB*. As pointed out in Subsection 4.3, in the limit $\bar{e}_r/\Delta_r$ would be 1, namely, if *IMB* were reduced to zero.

As anticipated in Subsection 4.4 and in hypothesis (4) of Subsection 5.1, the ratio $\bar{e}_r/\Delta_r$ should be little sensitive to the y-variable. Looking row by row at Table 2, we find this essentially confirmed. The near equality of the ratio for two variables of such different characteristics as *Income* (continuous) and *Employed* (dichotomous 0/1) is remarkable.

Table 3 presents a corresponding analysis of the LFS2012 data set, treated as a sample s of size 32,265, and with constant d_k, for the two pseudo y-variables, *Income* (continuous) and *Employed* (categorical, 1 or 0). The rows refer to the five response sets, r_{Act} (actual), and r_{A68}, r_{A65}, r_{A63} or r_{A60} (generated).

Some patterns seen in Table 2 for LCS2009 are confirmed in Table 3 for LFS2012. We note the following:

For the actual response r_{Act}, *IMB*, Δ_r and $\bar{e}_r$ are much lower in LFS2012 (Table 3) than in LCS2009 (Table 2), possibly a sign of a data collection in some sense "better" in the LFS (although the $\mathbf{x}$-vectors are not quite identical). The drop in *IMB* over the five response sets is accompanied by a consistent drop in both Δ_r and $\bar{e}_r$.

The reduction factor (highest divided by lowest) for the deviation Δ_r is 3.73/1.58 = 2.36 (for *Income*) and 1.30/0.89 = 1.46 (for *Employed*). The ratio $\bar{e}_r/\Delta_r$ shows a mild increase toward the limit value of 1, although with some irregularity for both *Income* and *Employed*.

Row by row inspection of Table 3 shows that the supposition (hypothesis (4)) that the ratio $\bar{e}_r/\Delta_r$ would be nearly the same for the two y-variables is less well confirmed than for LCS2009; here, $\bar{e}_r/\Delta_r$ is about 10% higher for *Employed* than for *Income*.

7. Estimating the Variance of the Deviation

The topic in this section is the sampling variability of the CAL estimator deviation Δ_r. We first derive a variance estimator for Δ_r, under probability sampling from a finite population.

We test the performance of this variance estimator – its approximate unbiasedness – by empirical work in which we treat the data sets LCS2009 and LFS2012 as populations of respective sizes $N = 8{,}220$ and $N = 32{,}265$. The notation U will refer to either of these two data sets. We draw repeated samples s from each of them and observe the properties of the variance estimator for Δ_r.

We choose to treat "response or not" on the part of a unit $k \in U$ as a fixed property of the unit: A value $i_k = 1$ or 0 is specified in the data file U for every k, to specify whether or not k responds and delivers the study variable value y_k. Hence, i_k is a fixed, non-random value, and a set of responding units, $r(U)$, is "pre-specified" in U. A probability sample s from U will have an implied response set $r = r(s) = s \cap r(U) = \{k : k \in s \text{ and } i_k = 1\}$. This formulation of "intrinsic response" in the population is advantageous for our empirical work.

Comment: One should be aware that this setup is different from probabilistic response, as when, given s, k is modeled to respond with (unknown) probability $\theta_k = \Pr(i_k = 1|s) = \Pr(k \in r|k \in s)$. In that formulation, i_k has a random structure specified by model assumptions.

To develop a variance estimator we write the expression for Δ_r in (15) slightly differently as

$$\Delta_r = \frac{1}{\sum_s d_k} \sum_s d_k G_k e_k, \tag{21}$$

where G_k is given in (14), e_k is the sample regression residual in (6), and the property $\sum_s d_k G_k \mathbf{x}'_k = \mathbf{0}'$ has been used. Formula (21) presents Δ_r as the sample mean of the quantities $G_k e_k$. In particular, when s is a simple random sample of size n drawn from N, then d_k is constant, and

$$\Delta_r = \frac{1}{n} \sum_s G_k e_k = \overline{Ge}_s \tag{22}$$

Basic knowledge about simple random sampling theory suggests the variance estimator

$$\widehat{V}_{srs}(\Delta_r) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{\sum_s (G_k e_k - \overline{Ge}_s)^2}{n-1} \tag{23}$$

It would be an unbiased estimator of variance if $G_k e_k$ were fixed non-random quantities. This does not hold here, because G_k and e_k are functions of s, which is a random set. Still, we can expect (23) to give satisfactory performance as a variance estimator for Δ_r in not-so-small simple random samples s. This is confirmed by our simulation results that follow.

One should note that the variance estimator (23) is limited to pseudo y-variables, with values y_k known for all sample units $k \in s$. The reason is that the residuals e_k require y_k for all sample units, which does not hold under nonresponse for ordinary survey variables.

The simulation was carried out by drawing repeated simple random samples from LCS2009 and from LFS2012, both viewed as populations. We drew 10,000 repeated samples from each of these. The samples from LCS2009 are of size $n = 5{,}000$, those from LFS2012 also of size $n = 5{,}000$.

Table 4. Simulation results, 10,000 simple random samples of size 5,000 from the LCS2009 data base, vector $\mathbf{x}_{\text{dm14}}$*: Simulation variance (var), simulation mean of variance estimate (estvar), both multiplied by* 10^{-6} *for Income, by* 10^{6} *for Employed.*

	Income		Employed	
Resp.set	*var*	*estvar*	*var*	*estvar*
r_{Act}	1.06	1.02	7.40	7.30
r_{A65}	1.28	1.23	8.24	8.14
r_{A60}	1.43	1.38	8.89	8.76
r_{A55}	1.67	1.65	9.67	9.61
r_{A50}	1.93	1.90	10.80	10.65

The **x**-vector is given as before by (19) for LCS2009 and by (20) for LFS2012. The two pseudo y-variables are *Income* and *Employed*, as in Section 6.

For each population, we specify five intrinsic response sets, namely, those specified in Section 6. This implies for LCS2009 that for a sample s drawn from U, we get five alternative response sets, $r_A(s) = s \cap r_A(U)$, with $r_A(U) = r_{Act}$ (actual), r_{A65}, r_{A60}, r_{A55} or r_{A50} (generated). Similarly for LFS2012, the alternative response sets obtained for a sample s are $r_A(s) = s \cap r_A(U)$, with $r_A(U) = r_{Act}$ (actual), r_{A68}, r_{A65}, r_{A63} or r_{A60} (generated).

It follows that the response $r_A(s)$ in a realized random sample s is also a random set. For example, for LCS2009 with $r_A(U) = r_{Act}$, the responding proportion of a sample s (the response rate) varies with a mean over the 10,000 repetitions very nearly equal to 0.679, which is the relative size of the actual response set r_{Act} (see Table 1).

Similarly, quantities such as *IMB*, Q_s, Q_r, Δ_r, $\bar{e}_r$, u_r and $\bar{e}_r/\Delta_r$ vary over the 10,000 repeated samples. For some of these, we computed mean, variance and standard deviation over the 10,000 samples. We expect that those means agree closely with the corresponding "population parameters" in Tables 2 and 3. This was confirmed, so we do not show simulation means and focus instead on the variance.

Simulation results are presented for LCS2009 in Table 4 and for LFS2012 in Table 5, where *var* is the variance of Δ_r given by (22) over the 10,000 repetitions (the simulation variance) and *estvar* is the simulation mean of the variance estimator, that is, the mean over the 10,000 repetitions of $\widehat{V}_{srs}(\Delta_r)$ given by (23).

The results in Tables 4 and 5 prompt a couple of comments. We note that *var* and *estvar* agree closely in the two tables, a sign that (23) is a satisfactory variance estimator for Δ_r.

Table 5. Simulation results, 10,000 simple random samples of size 5,000 from the LFS2012 data base; vector $\mathbf{x}_{\text{dm16}}$*: Simulation variance (var), simulation mean of variance estimate (estvar), both multiplied by* 10^{-6} *for Income, by* 10^{6} *for Employed.*

	Income		Employed	
Resp.set	*var*	*estvar*	*var*	*estvar*
r_{Act}	2.83	2.76	14.07	14.43
r_{A68}	3.26	3.17	14.95	15.35
r_{A65}	3.43	3.34	15.47	15.95
r_{A63}	3.56	3.46	16.14	16.65
r_{A60}	3.83	3.73	17.55	18.02

Table 6. Significance analysis; LCS2009 data, vector $\mathbf{x}_{\text{dm14}}$, five response sets; score $= \Delta_r / S_{Ge}$; IMB and deviation Δ_r obtained from Table 2; standard deviation S_{Ge} is multiplied by 10^{-3} for Income, by 10^2 for Employed.

		Income		Employed	
Resp.set	*IMB*	S_{Ge}	*score*	S_{Ge}	*score*
r_{Act}	1.878	1.257	5.94	0.336	6.20
r_{A65}	1.120	1.381	5.44	0.355	5.66
r_{A60}	0.890	1.461	4.78	0.368	5.21
r_{A55}	0.648	1.598	3.85	0.386	5.10
r_{A50}	0.426	1.713	3.03	0.406	4.30

The simulation variance *var* increases when the imbalance *IMB* is reduced. For Income, it is essentially doubled in going from r_{Act} to r_{A50} in Table 4. The tendency may not be surprising, but we did not anticipate it. A trade-off occurs, in that the mean (Tables 2 and 3) of the deviation Δ_r (that is, the bias) is reduced, which is a gain, while the variance (Tables 4 and 5) is increased, which is a disadvantage. But it can be claimed that the gain more than outweighs the disadvantage.

8. Is the Nonresponse Bias Important?

The question "is the nonresponse bias important" is relevant because it could be maintained that it may be trivially small, not significantly different from zero, and that in any case it is a result of random influences. Its significance or not depends on the degree of response imbalance. In this section, we subject the CAL estimator deviation Δ_r to testing, to see if it is significantly large, at different levels of *IMB*.

For this purpose we confront the deviation Δ_r with an appropriate standard deviation. Written in the form (22), $\Delta_r = \overline{Ge_s}$ is the sample mean of $G_k e_k$ over s. For this exercise, let us consider s as a simple random sample of size n from the Swedish population in scope for the survey, with a size N of the order of 7×10^6. Table 6 (for LCS2009) and Table 7 (for LFS2012) show results from this analysis. In those tables, data on *IMB* and Δ_r are taken from Tables 2 to 5, as the case may be, and $score = \overline{Ge_s}/S_{Ge} = \Delta_r / S_{Ge}$, where $S_{Ge}^2 = \widehat{V}_{srs}(\Delta_r)$ is given by formula (23) with $n = 8{,}220$ for LCS2009, $n = 32{,}265$ for LFS2012, and $1/N$ negligible.

If we take "2 or greater" as an indication of a significant difference from zero, we note in Tables 6 and 7 that all *score* values but one are well above the critical value of 2. For all

Table 7. Significance analysis; LFS2012 data, vector $\mathbf{x}_{\text{dm16}}$, five response sets; score $= \Delta_r / S_{Ge}$; IMB and deviation Δ_r obtained from Table 3; standard deviation S_{Ge} is multiplied by 10^{-3} for Income, by 10^2 for Employed.

		Income		Employed	
Resp.set	*IMB*	S_{Ge}	*score*	S_{Ge}	*score*
r_{Act}	0.788	0.711	5.25	0.162	7.99
r_{A68}	0.251	0.761	3.70	0.167	7.12
r_{A65}	0.158	0.780	3.04	0.171	6.37
r_{A63}	0.128	0.795	2.49	0.174	5.80
r_{A60}	0.061	0.826	1.91	0.181	4.89

rows and in both surveys, *score* is greater for *Employed* than for *Income*; the difference is more pronounced for LFS2012.

The most prominent feature of the two tables, holding an attractive promise for adaptive data collection, is the fact that *score* has a consistently decreasing pattern over the five response sets. Although the standard deviation S_{Ge} increases with lower imbalance – see Tables 4 and 5 – this is more than compensated for by the decrease in the deviation $\Delta_r = \overline{Ge}_s$. This holds for both surveys, LCS2009 and LFS2012, and for both y-variables, *Income* and *Employed*. It suggests that efforts to decrease *IMB* in data collection can reduce the CAL estimator's deviation – the bias – towards insignificant levels.

9. Concluding Discussion

In this article we have emphasized the link between the accuracy of a nonresponse adjusted estimator and the validity – or lack of it – of the regression fit behind the estimator. Both issues depend on the degree to which an adaptive data collection can succeed in reducing the response imbalance. A key question is: How important for accuracy is an adaptive data collection?

In practice, an auxiliary vector $\mathbf{x}$ is chosen, by selection from a supply of available auxiliary variables. The regression between survey variable y and $\mathbf{x}$ would, if "true", lend unquestionable support to the nonresponse adjusted estimator. But invariably that model does not hold for the responding set. The main problem is the selection effect, a non-random response, rather than one of not enough, or omitted, predictor variables. The estimation becomes biased due to *regression inconsistency*.

The inaccuracy (or bias) is measured here by the deviation, denoted Δ_r, of the nonresponse adjusted estimator (the calibration estimator) from the unbiased estimate possible under full response. The article focuses on the deviation Δ_r. We are able to express it as a sum of two terms, the regression inconsistency denoted $\bar{e}_r$, and a residual. The former is the dominating term: The regression inconsistency is a major factor in inaccuracy or bias in estimates.

The deviation Δ_r and the regression inconsistency $\bar{e}_r$ (both given in Section 3) depend on the degree of response imbalance realized in data collection. Imbalance is measured here by the statistic *IMB*, defined in Subsection 4.1 with respect to an auxiliary vector $\mathbf{x}$ known for the sample. The *IMB* statistic is a tool for the data collection: Methods exist to reduce it, with promise of more accurate estimation.

The empirical part of this article used data sets from two important Statistics Sweden surveys, the Living Conditions Survey and the Labour Force survey. These are examples only to illustrate a technique which can be reproduced for other surveys in other national statistical institutes.

We find that both Δ_r and $\bar{e}_r$ decrease when *IMB* is made to decrease. This is encouraging. But they do not decrease to zero with *IMB*. Both remain at distinctly non-zero levels even if *IMB* is brought to near-zero. The message is that modest expectations for better accuracy are in order, rather than hopes for a great payoff, when imbalance is reduced through an adaptive data collection.

This is understandable; the data collection is confined to operate on a designated auxiliary vector $\mathbf{x}$, known for response and for nonresponse, rather than on the survey

variable y itself. That would naturally be more effective, but is untenable because y-values are missing for the nonresponse. Whatever the choice of $\mathbf{x}$, it can never be a "full substitute" for the y-variable.

We studied the ratio $\bar{e}_r/\Delta_r$, the proportion of the estimator's deviation attributable to regression inconsistency, as a function of *IMB*. This ratio tends to one as *IMB* tends to zero. The ratio is quite insensitive to the survey variable y. In studies not reported here, we have noted that the ratio is, however, quite sensitive to the choice of $\mathbf{x}$-vector: It tends to be lower if a more extensive $\mathbf{x}$-vector is postulated.

We carried out a simulation study, involving repeated simple random sampling from the two data bases. The mean of the deviation Δ_r decreases as a result of reduced *IMB*, a positive message; the variance increases to some degree.

An estimator of variance for Δ_r was developed. We used it to test the deviation Δ_r, to see if it is small enough to be insignificant when *IMB* is made to be low in data collection. We found encouraging signs that this may be the case.

One may argue that the deviation Δ_r and the regression inconsistency $\bar{e}_r$ are studied here under rather idealized conditions: To compute them requires the survey variable values y_k for the full sample s. We can do this in an experimental setting, as here, where y is a register variable, called "pseudo y-variable", but not in a real survey context where y-values are missing for the nonresponding part. Nevertheless, important insights for practice can be gained by our kind of analysis.

Appendices

Appendix 1: Proofs

The proofs (a), (b), and (c) that follow make use of the $\mathbf{x}$-vector condition (3). Quantities defined in Sections 3 and 4 are used, including I_k and f_k given in (12).

(a) Proof that f_k is a predicted value from the linear regression fitted over s of I_k on $\mathbf{x}_k$. The linear fit gives the slope vector $\mathbf{b}_{Is} = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} / \sum_s d_k \mathbf{x}_k I_k = \mathbf{\Sigma}_s^{-1} \bar{\mathbf{x}}_r$. The predicted value $\hat{I}_k$ of I_k from this regression fit is thus $\hat{I}_k = \mathbf{b}_{Is}' \mathbf{x}_k = \bar{\mathbf{x}}_r' \mathbf{\Sigma}_s^{-1} \mathbf{x}_k = f_k$.

(b) Proof of the correlation coefficient (13). We need expressions for the variances S_{Is}^2 and S_{fs}^2, and the covariance S_{Ifs}. First, $S_{Is}^2 = \sum_s d_k (I_k - \bar{I}_s)^2 / \sum_s d_k = (1/P) - 1$, using that $\bar{I}_s = \sum_s d_k I_k / \sum_s d_k = 1$. Next, to find $S_{fs}^2 = \sum_s d_k (f_k - \bar{f}_s)^2 / \sum_s d_k$, use (3) to find $\bar{f}_s = \sum_s d_k f_k / \sum_s d_k = 1$ and $\sum_s d_k f_k^2 / \sum_s d_k = 1 + Q_s$ with Q_s given in (11). We get $S_{fs}^2 = 1 + Q_s - 1^2 = Q_s$.

For the covariance, we note that $\bar{f}_r = \sum_r d_k f_k / \sum_r d_k = \bar{\mathbf{x}}_r' \mathbf{\Sigma}_s^{-1} \bar{\mathbf{x}}_r = 1 + Q_s$, which leads to

$$S_{Ifs} = \sum_s d_k (I_k - \bar{I}_s)(f_k - \bar{f}_s) \Big/ \sum_s d_k = \bar{f}_r - \bar{I}_s \bar{f}_s = (1 + Q_s) - 1 \times 1 = Q_s.$$

Now $IMB = P^2 Q_s$ by (11), so $corr(I,f) = S_{Ifs}/(S_{Is} S_{fs}) = (IMB/P(1-P))^{1/2}$.

(c) Derivation of the correlation coefficient (16): We need first and second moments of the weight factors F_k and G_k given in (14). Both means are zero over s: $\sum_s d_k F_k / \sum_s d_k = 0$, $\sum_s d_k G_k / \sum_s d_k = 0$. To illustrate, the second of these is

verified with a use of (3):

$$\sum_s d_k G_k \Big/ \sum_s d_k = (1/P) \sum_r d_k g_k \Big/ \sum_s d_k - 1$$

$$= \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \sum_r d_k \mathbf{x}_k (\mathbf{x}_k' \boldsymbol{\mu}) \Big/ \sum_r d_k - 1 = \bar{\mathbf{x}}_s' \boldsymbol{\mu} - 1 = 0.$$

The two variances and the covariance are obtained by similar algebraic manipulations as

$$S_{Fs}^2 = \sum_s d_k F_k^2 \Big/ \sum_s d_k = (1 - P - PQ_s)/P;$$

$$S_{Gs}^2 = \sum_s d_k G_k^2 \Big/ \sum_s d_k = (1 - P + Q_r)/P;$$

$$S_{FGs} = \sum_s d_k F_k G_k \Big/ \sum_s d_k = (1 - P)/P.$$

To exemplify with the first of these, use that $Q_s = \bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_r - 1$ by (3), to obtain

$$S_{Fs}^2 = \sum_s d_k F_k^2 \Big/ \sum_s d_k = (1/P) - 2\bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_r + \bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1} \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum_s d_k \right) \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_r$$

$$= (1/P) - (1 + Q_s)$$

The expression (16) for the correlation coefficient $S_{FGs}/(S_{Fs}S_{Gs})$ then follows.

Appendix 2: Description of the LCS2009 and LFS2012 Data Bases, the Auxiliary Variables and the Survey Variables

The empirical sections 5 to 8 use data from two Statistics Sweden surveys: The Living Conditions Survey (LCS) in 2009 and the Labor Force Survey (LFS) in 2012. From these we created the two data bases referred to as LCS2009 and LFS2012, each consisting of a subset of the probability sample drawn for those surveys. A difference in target populations is that LCS targets persons aged 16 and above; LFS targets ages 16 to 74. We can treat both data bases as simple random samples of individuals from the Swedish Register of Total Population. The purpose is to have data sets on which to illustrate theoretical findings. The results are not claimed to be generalizable to the surveys themselves.

Both surveys use a single data collection mode: Telephone. All attempts by interviewers to establish contact with a sampled person – the call attempts – are registered by Statistics Sweden's WinDATI-system. For every sampled person, the system stores a series of "events". These include not only successful contacts but also call without reply, busy line, contact with household member other than the sampled person, and appointment booking for a later contact. Every registered event is considered a call attempt. There are more than 30 for some units. If contact and data delivery occurs, the data collection effort is complete for that sample unit. By tradition, an objective in both surveys has been to reach a response

rate as high as possible when the field work must necessarily stop. Given that objective, the two surveys use somewhat different data collection strategies.

A number of auxiliary variables are available and recorded for all units in the two data bases. Most of those were obtained by matching Swedish registers using the unique personal identifier key. They allow different $\mathbf{x}$-vectors to be constructed and used in computing the imbalance *IMB*, the calibration estimator $\hat{Y}_{CAL}$, the deviation $\boldsymbol{\Delta}_r$ and other quantities of importance in Sections 3 to 4. One $\mathbf{x}$-vector was used for each of LCS2009 and LFS2012. These are given in (19) and (20). For both, the weighting matrices $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_r$ are invertible. For the computations, we also need survey variables y. For both LCS2009 and LFS2012, we designate two register variables, available for all sample units, to be y-variables; these "pseudo y-variables" are *Income* (continuous, obtained from the Swedish tax register), and *Employed* (binary; 1 for employed person, 0 otherwise).

Also recorded in the data bases are five alternative response sets. That is, every unit k in the data base has an attached value equal to either 1 (for response) or 0 (for nonresponse). Each response outcome can be seen as a vector, of dimension equal to the size of the data base, of values 1 or 0 designating whether or not k is a respondent.

The first response set is the response actually recorded for the units in the data base when the data collection was carried out in 2009 and 2012, respectively. The other four are experimental response sets, constructed with the threshold method as described in Subsection 5.3 and in Särndal and Lundquist (2014). As a result, each of LCS2009 and LFS2012 contains five alternative response sets representing diminishing levels of imbalance *IMB*, to meet the objective in the article to study how imbalance influences quantities of interest. The threshold methods is applied somewhat differently in the two cases. Reasons are different sample sizes and different intensity in the field work. The thresholds used are different.

LCS2009. The LCS sample consists of individuals aged 16 years and older. The ordinary field work lasted five weeks; a follow-up that concluded the data collection brought the response rate to an ultimate 67.4%. The survey is annual and is published on a yearly basis. The data set counting 8,220 individuals that we use here as the LCS2009 data base can be considered a simple random subsample from the entire LCS sample.

We used five alternative response sets, the actual one and four experimental ones. The latter were constructed with the threshold method described in Subsection 5.3, using the vector $\mathbf{x} = \mathbf{x}_{\text{dm14}}$ given in (19) to compute the response propensities. We used the thresholds 65%, 60%, 55% and 50%. The intervention points were attempts 3, 6 and 9 of the ordinary data collection, the end of the ordinary data collection, and attempt 4 of the follow-up. Together with the actual response, this gives five alternative response sets r, denoted r_{Act}, r_{A65}, r_{A60}, r_{A55} and r_{A50}.

LFS2012. The Swedish Labour Force Survey is a monthly panel survey that targets persons aged 16 to 74. It uses a rotating panel such that a selected person is in the sample on eight occasions (every third month for 2 years). Two different samples are drawn: "ordinary" and "extended". The former is selected using a simple stratification, while the latter uses a more complex stratification aimed at more satisfactory covering of people deemed "outside the labour market". The LFS-interview time is relatively short. The survey is published on a monthly basis, requiring a shorter field work period than in the LCS. For more information about the LFS see http://www.scb.se/en_/.

To compose the data base LFS2012, we used part of the data from the 12 monthly LFS surveys in 2012: The first waves of the "ordinary" parts are combined. That is, we treat the 12 wave-one samples as one survey. The monthly first-wave size is approximately 2,650 persons. For the whole year this adds up to a size of 32,265 for our LFS2012 data base. The sample in each month is actually a stratified one, but not much different from a simple random sample, so we treat LCS2012 as such.

Five alternative response sets are used, r_{Act}, r_{A68}, r_{A65}, r_{A63} or r_{A60}. The first is the actual response as recorded for the units in LFS2012, with a response rate of 70.6%. The other four are generated from the actual LFS2012 response set r_{Act} by the threshold method and with the use of the recorded call attempt data.

To compute the response propensities for the threshold method, we used the monitoring vector $\mathbf{x} = \mathbf{x}_{\text{dm16}}$ given in (20). We used the thresholds 68%, 65%, 63% and 60%, and eight intervention points: Call attempts 3, 5, 7, 9, 12, 15, 18 and 22. At each intervention point units (or more correctly groups of units) with a response propensity higher than the threshold are set aside, as described in Subsection 5.3. This gives five experimental response sets r: r_{Act}, r_{A68}, r_{A65}, r_{A63} and r_{A60}.

10. References

Brick, J.M. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29: 329–353. Doi: https://doi.org/10.2478/jos-2013-0026.

Deville, J.C. and Y. Tillé. 2004. "Efficient Balanced Sampling. The Cube Method." *Biometrika* 91: 893–912. Doi: https://doi.org/10.1093/biomet/91.4.893.

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A* 169: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Haziza, D. and É. Lesage. 2016. "A Discussion of Weighting Procedures for Unit Nonresponse." *Journal of Official Statistics* 32: 129–145. Doi: http://dx.doi.org/10.1515/jos-2016-0006.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–161. Doi: http://dx.doi.org/10.2307/1912352.

Legg, J.C. and C.L. Yu. 2010. "A Comparison of Sample Set Restriction Procedures." *Survey Methodology* 36: 69–79.

Little, R.J.A. and S. Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31: 161–168.

Lundquist, P. and C.E. Särndal. 2013. "Aspects of Responsive Design – With Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29: 557–582. Doi: https://doi.org/10.2478/jos-2013-0040.

Matei, A. and M.G. Ranalli. 2015. "Dealing with Non-Ignorable Nonresponse in Survey Sampling: A Latent Modeling Approach." *Survey Methodology* 41: 145–164.

Särndal, C.E., K. Lumiste, and I. Traat. 2016. "Reducing the Response Imbalance: Is the Accuracy of the Survey Estimates Improved?" *Survey Methodology* 42: 219–238.

Särndal, C.E. and P. Lundquist. 2014. "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation." *Journal of Survey Statistics and Methodology* 2: 361–387. Doi: https://doi.org/10.1093/jssam/smu014.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society, Series A* 179: 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Schouten, B. 2015. "Statistical Inference Based on Randomly Generated Auxiliary Variables." *Discussion Paper*. Statistics Netherlands, The Hague.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology* 39: 29–58.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Response." *Journal of Official Statistics* 27: 231–253.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

Tourangeau, R., J.M. Brick, S. Lohr, and J. Li. 2017. "Adaptive and Responsive Survey Designs: a Review and Assessment." *Journal of the Royal Statistical Society, Series A* 180: 203–223. Doi: http://dx.doi.org/10.1111/rssa.12186.

Vartivarian, S. and R.J.A. Little. 2002. "On the Formation of Weighting Adjustment Cells for Unit Nonresponse." In proceedings of *American Statistical Association, Survey Research Methods Section*, 3553–3558.

Vella, F. 1998. "Estimating Models With Sample Selection Bias." *Journal of Human Resources* 33: 127–169. Doi: http://dx.doi.org/10.2307/146317.

Received April 2016
Revised April 2017
Accepted April 2017