An appraisal of common reweighting methods for nonresponse in household surveys based on Norwegian Labour Force Survey and Statistics on Income and Living Conditions Survey

N. D. Nguyen[1] and L.-C. Zhang[2]

Summary Despite increasing efforts during data collection, nonresponse remains sizeable in many household surveys. Statistical adjustment is hence unavoidable. By reweighting the design weights of the respondents are adjusted to compensate for nonresponse. However, there is no consensus on how it should be carried out in general. Theoretical comparisons are inconclusive in the literature, and the associated simulation studies involve hypothetical situations not all equally relevant to reality. In this paper we evaluate the three most common reweighting approaches in practice, based on real data in Norway from the two largest household surveys in the European Statistical System. We demonstrate how cross-examination of various reweighting estimators can help inform the effectiveness of the available auxiliary variables and the choice of the weight adjustment method.

Keywords: unit nonresponse, auxiliary variable selection, inverse propensity weighting, generalised regression estimation, doubly robust estimation

1 Introduction

Response rates in household surveys have declined steadily in many western countries (de Leeuw and de Heer, 2002; Stoop et al., 2010; Meyer et al., 2015). Post data collection, statistical adjustment is needed due to a sizeable amount of nonresponse. A standard process to compensate for unit nonresponse is reweightings (Little, 1986; Kalton and Flores-Cervantes, 2003; Särndal and Lundström, 2005; Brick, 2013). Generally speaking, this requires making two interrelated decisions on *auxiliary variable selection* and *weight adjustment method*. However, there is no consensus on a general approach.

We distinguish between the three most common reweighting approaches in practice. Firstly, the *two-step* approach combines response propensity weighting (from respondents to sample) and calibration (from sample to population); see e.g. Kalton and Kasprzyk (1986). In general two *different* sets of auxiliary variables are used at the two steps. The first step

[1] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland. Email: duong.nguyen@ucdconnect.com

[2] S3RI/Dept. Social Statistics and Demography, Univ. of Southampton, Southampton, UK & Statistisk sentralbyrå, Oslo, Norway & University of Oslo, Norway. Email: L.Zhang@soton.ac.uk

weight may either be directly given by the inverse of the estimated response propensities (Cassel et al., 1983; Little and Rubin, 1987), or indirectly based on adjustment cells formed using these propensities (Little, 1986; Eltinge and Yansaneh, 1997). Secondly, applying calibration of the sampling weights (from respondents to population) directly yields the *one-step* approach (Lundström and Särndal, 1999), for which a set of auxiliary variables should ideally have high association with both the response indicator and the target outcome variable. Adopting the linear calibration function yields the modified generalized regression (MGR) estimator (Bethlehem, 1988). Thirdly, using the *same* covariates for both response propensity modelling and calibration, the two-step approach could yield the so-called *doubly robust (DR)* estimators; see e.g. Robins et al. (1994); Robins and Wang (2000); Bang and Robins (2005); Carpenter et al. (2006); Kang and Schafer (2007).

Despite their long tradition, the choice between the two- and one-step approaches is still not conclusive in the literature. For instance, one may easily find motivations for the one-step approach (Little and Vartivarian, 2005; Särndal and Lundström, 2008, 2010), but there exist also several warnings against its potential pitfalls (Brick, 2013; Kott and Liao, 2015; Haziza and Lesage, 2016). Although the DR estimators have caught much attention outside the field of survey sampling, we did not come across any reports on their performance in real household (or business) surveys.

We believe theoretical comparisons are unable to reach a clear-cut choice because the 'true' nonresponse mechanism cannot be identified based on the observed data alone. Moreover, while simulation studies are useful for illustrating certain properties of one approach or another, not all the hypothetical set-ups are relevant to the reality. It is therefore essential to examine situations in actual household surveys, which are limited in number. For instance, in the context of European Statistical System (ESS), there are currently only about 10 major household surveys. Moreover, relevant auxiliary variables consist mostly (or entirely) of categorical variables, unlike what is common in simulation studies.

In this paper we assess *empirically* the three reweighting approaches outlined above, based on the Norwegian Labour Force Survey (LFS) and Survey of Income and Living Conditions (SILC), which are the two largest household surveys in the ESS. The protocol of the appraisal is generally applicable to other surveys or countries.

We begin with a description of the sampling designs of the Norwegian LFS and the SIC in Section 2. In Section 3, we describe a set of reweighting estimators to be investigated and some common variations. Then we introduce simple ANOVA-type measures to understand the potential effects of an auxiliary variable based on its association with the outcome variable and the response indicator in Section 4, and use real data to illustrate how these may be related to the resulting change in the point estimate and the associated variance. Our discussion brings forward greater nuances of the reweighting effects than those that have been delineated previously by Thomsen (1973, 1978), Oh and Scheuren (1983) and Little and Vartivarian (2005). In Section 5 we present an empirical study of the Norwegian LFS and SILC data. As will be demonstrated, cross-examination of the different point estimates and their variances can inform the effectiveness of the available auxiliary variables and the choice of the weight adjustment method. Some general conclusions that emerge from the empirical appraisal will be summarised in Section 6.

In summary, regarding auxiliary variable selection, we find that it is always useful to increase the association with the outcome variable, but seeking the highest possible association with nonresponse is not necessarily helpful. Moreover, we find that the choice of weight adjustment method matters, especially when there exist strong auxiliary variables for the outcome available; whereas provided only weak auxiliary variables for the outcome variable, limiting the loss of efficiency and avoiding spurious adjustment may be a relevant priority. Overall, we found no evidence in the situations examined to support an uncritical adoption of the two-step approach. Since the 'true' nonresponse model envisaged for a two-step approach cannot be identified based on the observed data, regardless of whether the available auxiliary variables have low or high association with nonresponse, it makes sense to choose based on cross-examination of the alternatives in a given situation.

2 Sampling designs

In this paper we use or relate the discussions to the LFS data in Section 4.2, 4.3 and 5.1, and the SILC data in Section 5.2. We now briefly describe the sampling designs of these two surveys and a relevant variable called *Panel Response Status*.

2.1 The LFS

The Norwegian LFS has a stratified cluster sampling design, where the 19 counties make up the strata and family units form the clusters. The population register provides the sampling frame. The target population consists of residents aged 15-74 years old in Norway. Every in-scope person stays in the LFS for eight quarters, and there is approximately an 7/8 overlap between two consecutive quarters. The quarterly sample contains approximately 24,000 individuals, and the current response rate is around 80 percent. All interviews are conducted by telephone.

The overlap between two consecutive quarters means that approximately one in eight persons is new in each quarterly sample. It is possible to create a variable called Panel Response Status that identifies every person as new in sample, or previous quarter respondent, or previous quarter nonrespondent. This variable has very high association with the current quarter response indicator, in that previous quarter respondents (or nonrespondents) are more likely to respond (or not respond) again. Later on we will use this Panel Response Status to demonstrate the effects of a variable that has high association with the response indicator on the point estimate and the variance of an estimator.

2.2 The SILC

The annual SILC collects data on housing, finance, health, and work, etc. The target population is residents who are aged 16 years and over and not living in institutions. It has a four-year rotating panel design. Individuals are selected from the population register by the SRS design. The interviews are largely conducted over telephone, although face-to-face interviews can take place as exceptions. Just like with the LFS, the panel design of the SILC allows one to create the Panel Response Status variable, distinguishing new persons in the sample, previous year respondents and previous year nonrespondents.

3 Reweighting estimators to be investigated

Consider a finite population U of size N. Let Y be an outcome variable of interest which takes the value y_i for unit $i \in U$. Assume that a sample s of size n is selected from U by probability sampling, where π_i is the inclusion probability and $d_i = 1/\pi_i$ is the design weight of unit $i \in s$. Let R be the response indicator defined as $r_i = 1$ if unit i responds and $r_i = 0$ otherwise, for $i \in s$. Let r denote the respondent sample of n_r units such that $r \subset s$ and $n_r < n$. We describe the various methods to be included in a schematic investigation, in terms of the estimator of the population total $t = \sum_{i \in U} y_i$.

As a baseline for comparison, consider the design weighted estimator

$$\hat{t}_d = \frac{n}{n_r} \sum_{i \in r} d_i y_i \ . \tag{1}$$

This estimator takes the sampling design into account, and is approximately unbiased for t provided nonresponse *missing completely at random* (MCAR, Little and Rubin, 1987). An alternative baseline estimator is the sample respondent expansion estimator

$$\hat{t} = \frac{N}{n_r} \sum_{i \in r} y_i \ . \tag{2}$$

It is unbiased for t provided MCAR and equal probability selection method (epsem), and allows one to gauge both the effects of sampling design and nonresponse on reweighting. In many household surveys, epsem holds either exactly or approximately, such that the difference between $\hat{t}_d$ and $\hat{t}$ may be small, when compared to the various reweighting estimators described below, which aim to adjust for the potential bias caused by nonresponse.

To begin with, when it comes to auxiliary variable selection, it is often recommended to select variables that have high association with both the survey variable (Y) and the response indicator (R); see e.g. Little and Vartivarian (2005); Schouten (2007); Särndal and Lundström (2008); Bethlehem et al. (2011). In practice, instead of building a bivariate model of (Y, R), it is common to model R and Y *separately.* Denote by Z the selected predictors of the R-model and by X those of the Y-model. The two generally do not coincide. Not all the variables in Z (or X) are equally important to R (or Y). In a sense one may consider the variables in the joint subset, denoted by $A = Z \wedge X$, to be explanatory of both R and Y, but we are unaware of any recommended reweighting approach that *only* makes use of A. There exist also other variable selection approaches which are not based on explicit R- and Y-modelling; see e.g. Schouten (2007); Särndal and Lundström (2010). However, we shall focus on the modelling approach to auxiliary variable selection in this paper, because it is more generally applicable and has a more direct connection to the weight adjustment methods, as will be explained shortly. Notice that in this paper we consider the y-values in the population to be fixed, when calculating the expectation and variance of an estimator, even when Y-modelling is used to 'assist' its construction.

Denote the response propensity of unit i, for $i \in s$, by

$$p_i = p(z_i; \alpha) = \Pr(r_i = 1 | z_i)$$

e.g. defined via a logistic regression model. Let

$$\mu_i = E(Y_i | x_i) = m(x_i; \beta)$$

be the conditional expectation of Y_i given x_i. For illustration, we shall assume the most common linear regression, i.e. $\mu_i = x_i^T \beta$; but other types of regression models of μ_i are equally feasible. The two-step weight adjustment that uses Z and X separately can now be given as

$$\hat{t}_{2sts} = \sum_{i \in U} m(x_i; \hat{\beta}) + \sum_{i \in r} \frac{d_i}{p(z_i; \hat{\alpha})} \{y_i - m(x_i; \hat{\beta})\} \ , \tag{3}$$

where $\hat{\beta} = [\sum_{i \in r} d_i x_i x_i^T / p(z_i; \hat{\alpha})]^{-1} \sum_{i \in r} d_i x_i y_i / p(z_i; \hat{\alpha})$, and $\hat{\alpha}$ is the estimator of α, which is typically obtained from fitting an appropriate logistic regression model to the sample by solving for $\sum_{i \in s} z_i [r_i - p(z_i; \alpha)] = 0$. The estimator (3) is approximately unbiased for t provided nonresponse is *missing-at-random* (MAR, Little and Rubin, 1987) given Z, and the model of p_i is correctly specified.

By itself the first step of (3) yields the Inverse Propensity Weighting (IPW) estimator

$$\hat{t}_{IPW} = \sum_{i \in r} \frac{d_i}{p(z_i; \hat{\alpha})} y_i \ . \tag{4}$$

It is approximately unbiased under the same condition as (3), but may be less efficient if X can help reduce the variance. Extreme weights can arise by IPW, when large weights are assigned to relatively few respondents with similar characteristics to nonrespondents. Some authors propose to stratify the sample into several groups (or adjustment cells) based on similar $p(z_i; \hat{\alpha})$, i.e. Response Propensity Stratification (RPS), and use the inverse within-group response rate as the 1st-step weight. RPS is reported to be more efficient than IPW in some studies (Little, 1986; Kang and Schafer, 2007), although Lunceford and Davidian (2004) warn against their routine use based on their theoretical and empirical results. In general, while potential modification of the IPW-weight $p(z_i; \hat{\alpha})^{-1}$ is always a relevant practical issue, the IPW weight is more easily interpretable when comparisons are made to other weight adjustment methods. We recommend $\hat{t}_{IPW}$ to be computed and included in a schematic investigation of reweighting methods.

Next, applying the second weight adjustment of (3) directly to the respondents yields the one-step MGR estimator

$$\hat{t}_{MGR} = \sum_{i \in U} m(x_i; \hat{B}) + \sum_{i \in r} d_i \{y_i - m(x_i; \hat{B})\} \ , \tag{5}$$

where $\hat{B} = [\sum_{i \in r} d_i x_i x_i^T]^{-1} \sum_{i \in r} d_i x_i y_i$. As mentioned before, other one-step calibration estimators are possible by other calibration functions. But the linear calibration (5) is the most routine choice, and we shall focus on it to compare the one-step approach to other adjustment methods. The MGR estimator is approximately unbiased, if nonresponse is MAR given X, and if the linear model of μ_i is correctly specified or if the response propensity p_i is the inverse of a linear combination of x_i (Lundström and Särndal, 1999). An extra feature sometimes included in the discussion of the one-step approach is when some variables in X are observed in the whole sample but have unknown population totals (Särndal and Lundström, 2005; Andersson and Särndal, 2016). However, this is not an essential difference to the two-step approach, because the same possibility can as well be accommodated by the two-step approach.

Now, the variables Z selected by R-modelling generally differ from X by Y-modelling. Moreover, none of the associated MAR assumptions can be entirely true. Under the DR approach, one uses the *same* variables to build an R-model and a Y-model; see e.g. Kim and Haziza (2014). The resulting estimator is approximately unbiased if either one of the two models is correctly specified. In practice, without actually building a bivariate (R, Y)-model, taking the auxiliary variables $V = Z \vee X$, as the union of Z and X following separate R- and Y-modelling, appears a likely course of variable selection. The DR estimator for t that we adopt for this study is thus given by applying the two-step approach (3) to (V, V) instead of (Z, X), i.e.

$$\hat{t}_{DR} = \sum_{i \in U} m(v_i; \hat{\xi}) + \sum_{i \in r} \frac{d_i}{p(v_i; \hat{\eta})} \{y_i - m(v_i, \hat{\xi})\} \ , \tag{6}$$

where $\hat{\xi} = [\sum_{i \in r} d_i v_i v_i^T / p(v_i; \hat{\eta})]^{-1} \sum_{i \in r} d_i v_i y_i / p(v_i; \hat{\eta})$ under the linear Y-model $\mu_i = v_i^T \xi$, and $\hat{\eta}$ is the estimator of the R-model parameter η in $p_i = p(v_i; \eta)$. Notice that this requires known population total of z_i, unlike the IPW estimator for which one only needs the z_i's in the sample. Provided nonresponse is MAR given V, the estimator (6) is approximately unbiased when either the R- or Y-model is correctly specified. Notice that unless separate modelling happens to result in $Z = X$, adopting $V = Z \vee X$ would imply over-fitting for p_i or μ_i. However, in the situation of $v_i = x_i$, Lunceford and Davidian (2004) uses to demonstrate the potential gains by the DR approach, i.e. to "over-model" $p(z_i; \alpha)$ by $p(v_i; \eta)$. So it is of interest to investigate the performance of $\hat{t}_{DR}$, despite the heuristic construction of V.

Table 1: A minimal set of reweighting estimators

	Weight adjustment method		
Selection and use of auxiliary variable	One-step IPW	One-step MGR	Two-step
Separate R- and Y-modelling	$\hat{t}_{IPW}(Z, -)$	$\hat{t}_{MGR}(-, X)$	$\hat{t}_{2sts}(Z, X)$
Refitting *after* R- and Y-modelling	$\hat{t}_{IPW}(V, -)$	$\hat{t}_{MGR}(-, V)$	$\hat{t}_{DR}(V, V)$

We arrive thus at a *minimal* set of estimators for a schematic investigation in any given situation (Table 1). Also specified are the respective auxiliary variables to be used for each reweighting estimator. For the estimators using $V = Z \vee X$, refitting of $p_i(v_i; \eta)$ and $\mu_i(v_i; \xi)$ is needed in practice. Cross-examination of the different point estimates and their associated variances in a given survey will be illustrated in Section 5.

4 Effects of auxiliary variable

4.1 Subclass reweighting and association measures

Not all the selected variables in Z or X are equally effective. To gauge the potential effects of a categorical auxiliary variable, $c = 1, 2, ..., C$, let the population be partitioned accordingly into C subclasses with known population sizes $N_1, \cdots, N_C$, and $N = \sum_{c=1}^{C} N_c$. Let each subclass consist of a respondent stratum and a nonrespondent stratum (Cochran, 1953), respectively, of the population sizes $N_c^{'}$ and $N_c^{''}$ and means $\bar{Y}_c^{'}$ and $\bar{Y}_c^{''}$. Let $\bar{Y}' = \sum_c N_c'\bar{Y}_c'/N'$ be the population respondent mean, where $N' = \sum_c N_c'$, and $\bar{Y}'' = \sum_c N_c''\bar{Y}_c''/N''$ the population nonrespondent mean, where $N'' = \sum_c N_c''$. Let $\bar{Y} = \bar{Y}'N'/N + \bar{Y}''N''/N$ be the population mean. Consider the unweighted sample respondent mean

$$\bar{y} = \sum_{i \in r} y_i/n_r = \hat{t}/N \ ,$$

as an estimator of $\bar{Y} = t/N$, against the reweighted respondent mean

$$\bar{y}_W = \sum_{c=1}^{C} W_c \, \bar{y}_c \ ,$$

where $W_c = N_c/N$ and $\bar{y}_c$ is the respondent mean in sample subclass c.

The set-up is convenient for several reasons. Previously, Thomsen (1973, 1978), Oh and Scheuren (1983) and Little and Vartivarian (2005) all use it to study the effects of reweighting, which is natural for household surveys where the auxiliary variables are either categorical or can be categorised, and the subclasses may arise from cross-classifying several variables. Based on subclasses $1, ..., C$, all the reweighting estimators described in Section 3 reduce to $\bar{y}_W$, provided simple random sampling (SRS), which allows us to isolate away the choice of adjustment method. Moreover, one can estimate the randomisation variances of $\bar{y}$ and $\bar{y}_W$ based on the observed sample (Thomsen, 1978), where the population y- and r-values are treated as fixed. As pointed out by Little and Vartivarian (2005), the SRS-assumption allows one to gain an appreciation of the relative efficiency, i.e. $\text{RE} = \text{Var}(\bar{y}_W)/\text{Var}(\bar{y})$, without complicating the technical details due to complex designs. Notice that, even when the sampling design is complex, or if one prefers the model-based or quasi-randomisation-based inference in the end, it is still possible to make use of the randomisation-based results below, obtained under the SRS assumption, in order to easily gauge the potential effects of an auxiliary variable.

Now, to examine the change of the point estimate due to subclass reweighting, let

$$B = E(\bar{y} - \bar{y}_W) = \frac{1}{\bar{h}} \sum_{c=1}^{C} W_c \bar{Y}_c^{'} (h_c - \bar{h}) = \frac{1}{\bar{h}} \sum_{c=1}^{C} W_c (\bar{Y}_c^{'} - \bar{Y}^{'})(h_c - \bar{h}) \ , \tag{7}$$

where $h_c = N_c^{'}/N_c$, for $h_c > 0$, is the population subclass respondent proportion, and $\bar{h} = \sum_c W_c h_c$ is the population respondent proportion. The second last expression in (7) is given by Thomsen (1973), and the last one follows since $\sum_c W_c(h_c - \bar{h}) = 0$. Considering $\{W_1, ..., W_C\}$ as a probability mass function, one may interpret B as the covariance between $\bar{Y}_c^{'}$ and h_c as c varies, denoted by $\mathrm{Cov}_W(\bar{Y}_c^{'}, h_c)$. Since $\bar{h}$ is fixed at the estimation stage, different subclass formations can only affect $\mathrm{Cov}_W(\bar{Y}_c^{'}, h_c)$. Thus, B would be large if either $\bar{Y}_c^{'}$ or h_c varies much across the subclasses, i.e. if the subclasses are heterogeneous either with respect to the outcome variable or the response indicator, or both.

Next, regarding the RE of subclass reweighting, Thomsen (1978) shows that

$$\begin{aligned}
\mathrm{Var}(\bar{y}) &= \frac{1}{n\bar{h}^2}\Big\{\sum_{c=1}^{C} W_c h_c S_c^2 + \sum_{c=1}^{C} W_c h_c (\bar{Y}_c^{'} - \bar{Y}^{'})^2\Big\} = \frac{1}{n\bar{h}^2}(\tau_1 + \tau_2)\ , \\
\mathrm{Var}(\bar{y}_W) &\approx \frac{1}{n}\sum_{c=1}^{C} W_c S_c^2 / h_c\ ,
\end{aligned}$$

where $S_c^2 = \sum_{i=1}^{N_c^{'}} (Y_{ci} - \bar{Y}_c^{'})^2/(N_c^{'} - 1)$ is the population subclass respondent variance. Notice that $\mathrm{Var}(\bar{y})$ can be decomposed into two terms of within- and between-subclass respondent variances, denoted by τ_1 and τ_2, respectively, with fixed sum $\tau_1 + \tau_2$. A corresponding ANOVA-type measure of the association between c and Y can be given by

$$\lambda_{cY} = \tau_2/(\tau_1 + \tau_2)\ .$$

The association measure λ_{cY} provides an easy appreciation of the potential effects of the auxiliary variable (or variables) underlying the subclasses $c = 1, ..., C$. In the extreme case of $\lambda_{cY} = 1$ and $\tau_1 = 0$, we would have $B = \mathrm{Bias}(\bar{y}) = E(\bar{y}) - \bar{Y}$ and $\mathrm{Var}(\bar{y}_W) = 0 < \mathrm{Var}(\bar{y})$. At the other end, where $\lambda_{cY} = 0$, $\tau_2 = 0$ and $S_c^2 \equiv S^2$, we would have $B = 0$ and

$$\mathrm{Var}(\bar{y}) = \frac{S^2}{n}\cdot\frac{1}{\bar{h}} \leq \frac{S^2}{n}\cdot\prod_{c=1}^{C}\big(\frac{1}{h_c}\big)^{W_c} \leq \frac{S^2}{n}\cdot\sum_{c=1}^{C}\frac{W_c}{h_c} \approx \mathrm{Var}(\bar{y}_W)\ ,$$

by applying twice the inequality of weighted arithmetic and geometric means, or directly the Titu's lemma as a special case of Cauchy-Schwarz inequality. Between the two extreme cases, increasing λ_{cY} makes the subclasses more heterogeneous with respect to Y, which tends to decrease the within-subclass variances S_c^2 and $\mathrm{Var}(\bar{y}_W)$, as well as increasing the change of point estimate, i.e. provided fixed $h_1, ..., h_C$.

Similarly, an ANOVA-type measure of the association between c and R is given as

$$\lambda_{cR} = \sum_{c=1}^{C} W_c (h_c - \bar{h})^2 / \Big\{\sum_{c=1}^{C} W_c h_c (1 - h_c) + \sum_{c=1}^{C} W_c (h_c - \bar{h})^2\Big\} = \nu_2/(\nu_1 + \nu_2)$$

where ν_1 and ν_2 are the within- and between-subclass variances of R, respectively, with fixed sum $\nu_1 + \nu_2$. In the extreme case of $\lambda_{cR} = 1$ and $\nu_1 = 0$, h_c would be either 0 or 1, such that the subclasses are nested in the respondent and nonrespondent strata. We would have $B = 0$, despite perfect association between c and R, so that subclass reweighting affects only the variance depending on λ_{cY}. At the other end, where $\lambda_{cR} = 0$, $\nu_2 = 0$ and $h_c \equiv \bar{h}$, we would again have $B = 0$, where subclass reweighting affects only the variance. Between the two extreme cases, both B and $\mathrm{Cov}_W(\bar{Y}_c', h_c)$ are likely to increase with $\nu_2 = \mathrm{Var}_W(h_c)$ and λ_{cR}. To appreciate what might happen to the variance at the same time, rewrite

$$\mathrm{Var}(\bar{y}_W) \approx \frac{1}{n}\Big\{ \sum_{c=1}^{C} W_c S_c^2/\bar{h} - \sum_{c=1}^{C} W_c S_c^2 (h_c - \bar{h})/\bar{h}^2 + \sum_{c=1}^{C} W_c S_c^2 (h_c - \bar{h})^2/\bar{h}^3 \Big\} \,,$$

based on Taylor expansion of h_c around $\bar{h}$. As ν_2 increases, the term involving $(h_c - \bar{h})^2$ may increase accordingly, while that involving $(h_c - \bar{h})$ remains small since $\sum_c W_c(h_c - \bar{h}) = 0$. In particular, even if λ_{cY} is high and S_c^2's are relatively small, it is possible for the term involving $(h_c - \bar{h})^2$ to increase to such an extent that we would have $\mathrm{Var}(\bar{y}_W) > \mathrm{Var}(\bar{y})$. Thus, as λ_{cR} increases, subclass reweighting is likely to achieve greater change of the point estimate while increasing the variance at the same time.

Remark Särndal and Lundström (2010) consider three indicators, H_1 - H_3, for the usefulness of auxiliary information. They consider H_2 to be ad hoc, which is only included for exploration. According to their conclusion, they prefer H_1 for a given y-variable, and they argue for H_3 as a tentative choice for the "many y-variables situation", but call for more research to develop other indicators (than H_3).

The indicator H_1 is given by $H_1 = |H_0|$ and $H_0 = \Delta_A/S_y$. Combining eqs. (2.1), (5.2), (5.7), (5.8) and (5.11) in Sarndal and Lundstrom (2010), we have

$$H_0 = \frac{\Delta_A}{S_y} = -R_{y,m} \times cv_m = -\frac{Cov(y,m)}{S_y S_m} \times \frac{S_m}{\bar{m}_{r;d}} = -\frac{P}{S_y} Cov(y,m)$$

where P is the weighted response rate, i.e. an estimate of $\bar{h}$ in our set-up, and $\Delta_A = (\tilde{Y}_{EXP} - \tilde{Y}_{CAL})/\hat{N}$, with the "expansion" estimator $\tilde{Y}_{EXP}$ and the "calibration" estimator $\tilde{Y}_{CAL}$. Thus, Δ_A is similar to the B-term by eq. (7) in this paper, defined as the expectation of $\bar{y} - \bar{y}_W$ under SRS, where $\bar{y} = \hat{t}_d/N = \tilde{Y}_{EXP}/\hat{N}$ and $\bar{y}_W = \tilde{Y}_{CAL}/\hat{N}$ by subclass reweighting. Notice that by eq. (7) in this paper, B is a function of $\bar{h}$ and $Cov_W(\bar{Y}_c', h_c)$. The key difference between Δ_A and B is that the latter is based on the response propensity p_i's, whereas the former is based on m_i's which are on the scale of $1/p_i$.

Next, $H_3 = cv_m$, which is based on the auxiliary variables and the response indicator but not the y-values. In this sense it is similar to λ_{cR} in our paper, which measures the association between the auxiliary variables and the response indicator. While H_3 is related to the variance of m_i, λ_{cR} is related to the variance of p_i; while H_3 depends in addition on $\bar{h}$, λ_{cR} depends in addition on the decomposition of the variance of p_i.

Thus, by introducing λ_{cY} and λ_{cR}, we move into areas not covered by Särndal and Lundström (2010). In particular, we find that as λ_{cY} increases, reweighting tends to increase both bias adjustment (B or Δ_A) and efficiency gains; whereas as λ_{cR} increases, reweighting is likely to increase bias adjustment but inflate the variance at the same time.

4.2 A simulation study

In Section 4.1, we presented the formula for B, the change in point estimate due to subclass reweighting, as well as the fomulas for the variance of the unweighted respondent mean $\bar{y}$ and weighted mean $\bar{y}_W$. These formulas hold exactly under SRS. In practice, strict SRS is not the most common design, despite the household survey inclusion probabilities tend not to vary greatly across the population. They can still provide useful indications for the relative importance and potential effects of the different auxiliary variables in reweighting, as we will discuss in more details in Section 5, even though they do not suffice as the final uncertainty measures to be reported together with the survey estimates. We feel that such uses are warranted based on our past experience of in-house empirical evaluations. Below we carry out a simple simulation study to illustrate this point.

First we generate a Norwegian Labour Force population that resembles the LFS in the first quarter of 2015, including the response indicator. This proceeds as follows.

- The population of approximately 3.8 million Norwegians aged 15-74 are distributed in the 19 counties according to the situation in the first quarter of 2015. The county population size varies from approximately $58,000$ to $506,000$. We refer to more details of the population at `https://www.ssb.no/en/befolkning/statistikker/folkemengde`.

- Within each county, assign each person a binary register employment status, such that the total number of register employed people is as given in the first quarter of 2015.

- Within each county, simulate independently the LFS classification (employed, unemployed, inactive) for each person, by the multinomial distribution with the corresponding proportions observed among the LFS respondents in that county.

- Within each county h, simulate independently the response indicator (yes, no) for each person, using the Bernoulli distribution with a probability $0.81 + d_{1h}$ if the person is register employed and $0.76 + d_{0h}$ if the person is not register employed. The figures 0.81

and 0.76 are respectively the average response rates for the registered employed and not registered employed in the first quarter of 2015. Within each stratum, the response rates for these two groups vary slightly, about 2% above or below the averages. Hence, d_{0h} and d_{1h} are simulated to have a normal distribution with mean 0 and standard deviation 0.01 to the reflect the range of the corresponding stratum response rates observed in the LFS sample.

We then draw repeatedly samples (of the same size as the LFS) from this population using SRS or Stratified SRS (StrSRS), where the strata are the 19 counties and the stratum sample sizes are as in the Norwegian LFS. The county sample size varies from 610 to $2,745$. Based on m simulated samples, with sufficiently large m, we may compare the true values of B, $Var(\bar{y})$ and $Var(\bar{y}_W)$ under each sampling design, with the expected sample estimates of them using the formulas in Section 4.1 under the assumption of SRS. The results for the proportions of unemployed and employed are given in Table 2. It can be seen that the formulas under the SRS assumption ("Estimated") hold as well approximately under the Stratified SRS sampling design.

Table 2: Simulation results ($\times 10^{-3}$), $m = 1000$.

	Unemployment				Employment		
	Estimated	SRS	StrSRS		Estimated	SRS	StrSRS
B	-1.00	-1.00	-1.00	B	12.48	12.44	12.62
$s.e(\bar{y})$	1.14	1.14	1.20	$s.e(\bar{y})$	3.34	3.33	3.45
$s.e(\bar{y}_W)$	1.16	1.16	1.22	$s.e(\bar{y}_W)$	1.90	2.01	1.93

4.3 Examples from the Norwegian LFS data

In practice, λ_{cY} and λ_{cR} are neither 0 nor 1, and they vary simultaneously with the auxiliary variables. In the literature such as those cited in Section 3, it is often suggested that one should select variables that have high associations with *both* Y and R. Little and Vartivarian (2005) summarise in their "Table 1" the effects of reweighting, depending on the association of the auxiliary variables to Y and R, which is reproduced here as Table 3. However, our own experiences (Zhang et al., 2013) suggest that there exist greater nuances in reality, which we demonstrate below using four examples based on the Norwegian LFS data. The examples illustrate also how $(\lambda_{cY}, \lambda_{cR})$ may be related to the changes of the point estimate and the associated variance.

We use the Norwegian LFS in the first quarter of 2015. The sample size is $n = 24,353$ and the response rate is $\bar{h} = 0.79$. We consider two binary Y-variables: employment and unemployment status. All the terms B, $\text{Var}(\bar{y})$, $\text{Var}(\bar{y}_W)$, etc. are estimated based on the

Table 3: Effects of nonresponse reweighting, from Little and Vartivarian (2005).

Association with Nonresponse	Association with Outcome Variable	
	Low	High
Low	Effect on Bias: — Effect on Variance: —	Effect on Bias: — Effect on Variance: ↓
High	Effect on Bias: — Effect on Variance: ↑	Effect on Bias: ↓ Effect on Variance: ↓

observed sample. However, for simplicity we do not introduce extra notations to emphasise that the values presented are estimates instead of population quantities.

Example 1 Let Y be the LFS Unemployment Status. Let two subclasses be formed based on the Registered Employment Status, where $c = 1$ for not registered employed and $c = 2$ for registered employed. We have $W_c = (0.35,\, 0.65)$ and $h_c = (0.74,\, 0.81)$, for $c = (1,\, 2)$, with the corresponding subclass respondent means $\bar{y}_c = (0.07,\, 0.00)$ and respondent variances $S_c^2 = (0.06,\, 0.00)$. We obtain

$$\lambda_{cY} = 0.04,\ \lambda_{cR} = 0.01,\ B = -1.41 \times 10^{-3},\ \text{s.e}(\bar{y}) = 1.13 \times 10^{-3},\ \text{RE} = 1.07\ .$$

Both λ_{cY} and λ_{cR} are close to zero. This provides an example of the top-left scenario in Table 3, according to which reweighting has little effect. However, the point estimate is actually changed by about 120% of the standard error (s.e) of $\bar{y}$, while it increases the variance only slightly. Previous studies of the Norwegian data (Zhang, 1999; Thomsen and Zhang, 2001; Zhang, 2005) all conclude that employment is overestimated and unemployment underestimated, based on the unadjusted respondent sample. The adjustment B is therefore in the direction one would expect, and it is by no means ‘negligible’ in size, despite the low association of the auxiliary variable with both Y and R.

Example 2 Let Y be the LFS Employment Status, and keep the same subclasses as in Example 1. We have $\bar{y}_c = (0.14,\, 0.96)$ and $S_c^2 = (0.12,\, 0.04)$, and

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.01,\ B = 1.68 \times 10^{-2},\ s.e(\bar{y}) = 3.31 \times 10^{-3},\ \text{RE} = 0.34\ .$$

It can be seen that λ_{cR} stays the same but λ_{cY} is greatly increased, compared to when the outcome variable is Unemployment Status. This provides an example of the top-right scenario in Table 3, according to which reweighting leads to little bias adjustment, although it may reduce the variance. However, it can be seen that in addition to the huge variance reduction, the change in the point estimate is also several times the standard error.

Example 3 Let Y be the LFS Employment Status. Let the subclasses be formed using the Panel Response Status, where $c = 1$ if previous nonrespondent, $c = 2$ if previous respondent, and $c = 3$ if new sample unit. For $c = (1, 2, 3)$, we obtain $W_c = (0.20, 0.67, 0.13)$, $h_c = (0.29, 0.94, 0.77)$, $\bar{y}_c = (0.66, 0.71, 0.66)$, and $S_c^2 = (0.22, 0.21, 0.22)$, so that

$$\lambda_{cY} = 0.00,\ \lambda_{cR} = 0.39,\ B = 5.50 \times 10^{-3},\ s.e(\bar{y}) = 3.31 \times 10^{-3},\ \mathrm{RE} = 1.28\ .$$

Compared to Example 1, λ_{cR} is considerably increased but λ_{cY} remains almost zero. This provides an example of the low-left scenario in Table 3, according to which reweighting leads to little bias adjustment, although it may increase the variance. Actually, however, in addition to the increasing variance, the change in the point estimate is again by no means 'negligible' in size, despite the low association between the auxiliary variable and Y.

Example 4 Let Y be the LFS Employment Status. Crossing the Panel Response Status and the Registered Employment Status yields the subclasses, where $c = 1$ if previous nonrespondent and not registered employed, $c = 2$ if previous nonrespondent and registered employed, $c = 3$ if previous respondent and not registered employed, $c = 4$ if previous respondent and registered employed, $c = 5$ if new sample unit and not registered employed, and $c = 6$ if new sample unit and registered employed. Then, for $c = (1, 2, 3, 4, 5, 6)$, we obtain $W_c = (0.08, 0.12, 0.21, 0.47, 0.05, 0.08)$, $h_c = (0.25, 0.31, 0.93, 0.94, 0.72, 0.79)$, $\bar{y}_c = (0.14, 0.95, 0.14, 0.96, 0.10, 0.95)$, $S_c^2 = (0.12, 0.05, 0.12, 0.04, 0.09, 0.05)$, and

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.39,\ B = 1.78 \times 10^{-2},\ s.e(\bar{y}) = 3.31 \times 10^{-3},\ \mathrm{RE} = 0.43\ .$$

Compared to Example 2, λ_{cR} is considerably increased in addition to high λ_{cY}. This provides an example of the low-right scenario in Table 3, which is 'ideal' according to the prevailing recommendation in the literature. However, while the adjustment B is increased by about 6% compared to the reweighting in Example 2, there is also a loss of efficiency by about 26%. In other words, it is *not* unreservedly beneficial to increase the association with R, while the association with Y remains the same. In fact, we now demonstrate the caveat of doing so with the following thought experiment.

Example 4* The first two h_c's in Example 4 are the response rates of the previous nonrespondents, the next two of the previous respondents, and the last two of the new sample members. To vary the response rates more extremely, suppose we have full response among the previous respondents, so that $h_3 = h_4 = 1$; suppose the response rates among the new sample units stay the same, so that $h_5 = 0.72$ and $h_6 = 0.79$; suppose the response rates among the previous nonrespondents are reduced to $h_1 = 0.05$ and $h_2 = 0.10$. This yields $h_c = (0.05, 0.10, 1.00, 1.00, 0.72, 0.80)$, with the same overall response rate $\bar{h} = 0.79$. Keeping everything else the same as in Example 4, we obtain

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.78,\ B = 1.99 \times 10^{-2},\ s.e(\bar{y}) = 3.31 \times 10^{-3},\ \mathrm{RE} = 1.13\ .$$

As remarked earlier in Section 4.1, *without* increasing λ_{cY} at the same time, increasing λ_{cR} on its own can result in $\mathrm{Var}(\bar{y}_W) > \mathrm{Var}(\bar{y})$, despite high association λ_{cY}.

5 Empirical study of reweighting

For this study of reweighting, a number of auxiliary variables are extracted from the statistical register system at Statistics Norway and linked to the samples at the individual level. For the LFS, these include age (11), sex (2), county (19), education level (4), marital status (3), family type (3), immigration (3), birth country (2), income (5), household income (5), registered employment (2), where the numbers in parentheses indicate the number of categories each variable has. The same variables are used for the SILC, except for registered employment due to data protection regulations. In addition, some of the variables are adjusted to have fewer categories due to the smaller SILC sample size, e.g. 4 age groups instead of 11, 7 regions instead 19 counties, etc.

For both R- and Y-modelling, variable selection is carried out stepwise according to the Akaike Information Criterion. While this is somewhat simplistic, it suffices for the purpose of this study and reflects well the existing process at national statistical offices. All the 6 estimators listed in Table 1 are applied to each of the outcome variable to be presented, in terms of the corresponding population mean estimators, denoted by $\bar{y}_{method} = \hat{t}_{method}/N$ where the subscript *method* identifies the weight adjustment method. The baseline estimator to be presented is $\bar{y} = \hat{t}/N$ for $\hat{t}$ given by (2). The difference to $\hat{t}_d/N$ is negligible compared to their differences to the various reweighting estimates. To save space, other estimators that have been calculated may be mentioned in comments but not presented in details. This include e.g. using RPS instead IPW under the two-step approach. All the estimated variances are calculated in R using 500 bootstrap samples with the same sampling design as the LFS/SILC, except for one case to be specified later. The bootstrap follows the procedure of Canty and Davison (1999), where to mimic the effect of sampling without replacement, the bootstrap population is made by concatenating copies of the observed sample, from which the bootstrap replicate samples are taken without replacement according to the given sampling design. For each sample, we calculate the estimates for each of the estimators discussed in Section 3, and the standard deviation of these estimates is used to estimate the standard error of each estimator.

5.1 The LFS

We have carried out the same analysis for five quarterly samples. The results are very similar, so only those based on the first quarter in 2015 are presented here, where we focus on two binary outcome Y-variables, employment and unemloyment, denoted by Y_{em} and Y_{un}, respectively.

Table 4: Association with (R, Y_{em}, Y_{un}), selected†, B in 10^{-2}

Auxiliary variable	λ_{cR}	$\lambda_{cY_{em}}$	$\lambda_{cY_{un}}$	B_{em} (RE)	B_{un} (RE)
Registered employment	$0.01^\dagger$	$0.69^\dagger$	$0.04^\dagger$	1.68 (0.33)	-0.14 (1.07)
Age	$0.02^\dagger$	$0.28^\dagger$	$0.01^\dagger$	-1.04 (0.71)	-0.10 (1.08)
Sex	$0.00^\dagger$	$0.00^\dagger$	$0.00^\dagger$	0.01 (1.00)	0.00 (1.00)
County	$0.00^\dagger$	$0.01^\dagger$	0.00	0.07 (0.99)	0.00 (1.00)
Family type	$0.02^\dagger$	0.02	0.00	0.19 (0.99)	-0.02 (1.02)
Birth country	0.02	0.00	0.01	-0.25 (1.01)	-0.15 (1.11)
Immigration status	$0.02^\dagger$	0.00	$0.01^\dagger$	-0.20 (1.02)	-0.16 (1.12)
Education	$0.01^\dagger$	$0.11^\dagger$	$0.01^\dagger$	0.59 (0.91)	-0.06 (1.04)
Marital status	$0.02^\dagger$	0.01	0.00	0.30 (1.00)	-0.06 (1.05)
Income	$0.02^\dagger$	$0.26^\dagger$	$0.02^\dagger$	1.47 (0.80)	-0.13 (1.08)
Household income	$0.04^\dagger$	$0.09^\dagger$	0.01	1.39 (0.96)	-0.14 (1.13)

The association measures of each covariate with R, Y_{em} and Y_{un} are given in Table 4, together with B and RE by the respective subclass reweighting, as described in Section 4.1. It can be seen that the available covariates have very different associations with the two outcome variables. Whilst registered employment, age, income and education all have a high association with Y_{em}, the association with Y_{un} is much lower across the board, although registered employment and income remain the two with the highest associations there. The covariates selected for the R-model and the two Y-models are marked (by †) for the corresponding λ_{cR}, $\lambda_{cY_{em}}$ and $\lambda_{cY_{un}}$ (Table 4). No interaction terms are selected for any of the models based on these data. Largely the same variables are selected for both Y-models, denoted by X_{em} and X_{un}, respectively. Each model includes the covariates that have the highest association with either Y_{em} or Y_{un}. The R-model includes all the available covariates (Z), except for birth country that is similar to immigration status. In particular, both X_{em} and X_{un} are nested in Z, such that $V = Z$ for both Y_{em} and Y_{un}.

The different estimates and their associated s.e's (in parentheses) are given in Table 5. Compared to the baseline estimate, all the reweighting estimates adjust the employment rate downwards and the unemployment rate upwards, i.e. in the direction expected. In the case of employment, all the one-step MGR and two-step estimators reduce the variance, while the one-step IPW estimator increase the variance. In the case of unemployment, all the reweighting estimators increase the variance, but have similar RE to each other. For

Table 5: LFS estimates (s.e) in 10^{-2}, the first quarter 2015

Auxiliary for (IPW, MGR)	Mean employment, $\bar{y} = 69.84$ (0.35)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step estimator
(Z, X_{em})	67.47 (0.44)	67.10 (0.19)	$\bar{y}_{2sts} = 67.08$ (0.19)
(Z, Z)	"	67.10 (0.19)	$\bar{y}_{DR} = 67.09$ (0.19)
Auxiliary for (IPW, MGR)	Mean unemployment, $\bar{y} = 2.45$ (0.12)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step estimator
(Z, X_{un})	2.99 (0.14)	3.06 (0.14)	$\bar{y}_{2sts} = 3.18$ (0.15)
(Z, Z)	"	3.05 (0.14)	$\bar{y}_{DR} = 3.19$ (0.15)

both Y_{em} and Y_{un}, the point-estimate changes are very large compared to the s.e's. Bias exploration by the method described in Zhang (1999) suggests that, provided informative nonresponse, the reweighted employment estimators may still have a positive bias, so that the risk is low that the reweighted estimators are more biased than the baseline estimator. Likewise for the reweighted unemployment estimators, since the upward adjustments of unemployment resulted from reweighting appear plausible in magnitude compared to the downward adjustments of employment.

To a large extent these results have confirmed the potential adjustment effects, which are suggested by simple subclass reweighting and association measures in Section 4.3. As indicated in Example 2 there, it is possible to achieve large adjustment of the point estimate *and* variance reduction for Y_{em}, without high association with R but provided high association with the outcome variable. Moreover, as indicated in Example 1, the reweighting estimators can yield appreciable adjustment of the point estimate of Y_{un} but also slightly increase the variance, despite the association is low with both Y_{un} and R.

Cross-examination of the estimators gives rise to additional noteworthy observations. Firstly, a striking result in Table 5 is the large variances of the IPW estimators, e.g. $\bar{y}_{IPW}$ is even less efficient than the baseline estimator $\bar{y}$ for Y_{em}. We notice that using RPS with 5 groups is unable to reduce the variance compared to the IPW estimator for these LFS data. Recall that in the case of $V = Z \vee X = X$, Lunceford and Davidian (2004) show that "over-modelling" $p(z_i; \alpha)$ by $p(v_i; \eta)$ can reduce the variance of the IPW estimator. However, since X_{em} is a subset of Z here, the predictive covariates are already included in Z and the strategy of "over-modelling" does not work. This shows that having predictive variables for Y in the R-model does not guarantee efficiency by itself, without an appropriate weight adjustment method. For instance, the MGR estimator based on "over-modelling" $\mu(v; \xi)$ with $v = z$ is basically as efficient as $\bar{y}_{MGR}$ that only uses X_{em}. Moreover, the two-step estimator $\bar{y}_{2sts}$ is able to recover almost all the lost efficiency of $\bar{y}_{IPW}$ by calibration of the IPW-adjusted weights $d_i/p(z_i; \hat{\alpha})$ with respect to X_{em}.

Secondly, the two-step approach $\bar{y}_{2sts}$ does not offer any noticeable advantage over the one-step MGR for the Norwegian LFS. In theory, correct modelling of the unit nonresponse could yield approximately unbiased estimation for any outcome variable. In reality, however, the true nonresponse model is unobtainable. This is certainly the case with the LFS data here, given the low association between the available covariates and R. Empirically, $\bar{y}_{2sts}$ does not yield any notable improvement over $\bar{y}_{MGR}$ here, but is more complicated due to an extra step of model-fitting and reweighting.

Thirdly, the DR approach does not offer any noticeable advantage compared to the traditional one and two-step approaches for the Norwegian LFS. In the case of Y_{em}, where there is a good Y-model, the results here agree with the literature (Bang and Robins, 2005; Kang and Schafer, 2007) that the DR estimator does not perform better than the regression estimator, but could improve the performance of $\bar{y}_{IPW}$ obtained from the R-model alone. Compared to the two-step estimator $\bar{y}_{2sts}(Z, X)$, the DR estimator $\bar{y}_{2sts}(Z, Z)$ has the same IPW-weights, but differ with respect to the extra calibration variables in $Z \setminus X_{em}$ for Y_{em} and $Z \setminus X_{un}$ for Y_{un}. However, this makes little difference since the extra variables do not have any appreciable association with the respective outcome variable.

The one-step MGR estimator $\bar{y}_{MGR}$ seems therefore reasonable for the Norwegian LFS, among the options considered here. The auxiliary variables may be selected with respect to several key Y-variables. It is the simplest in production, and it has the lowest variance, although the difference to the two-step alternatives are small in this case. It may be noticed that the existing production method in the LFS is essentially the same as subclass reweighting based on post-stratification by sex, age, and registered employment. It performs similarly to $\bar{y}_{MGR}$ for both Y_{em} and Y_{un}, with somewhat smaller adjustment of the point estimates but also smaller variance for Y_{un}. Therefore, the key to improve the existing method must be to find other auxiliary variables in the statistical register system, as more administrative data are being made available, which are more predictive of the unemployment status Y_{un}. The MGR can be used instead of the post-stratification if the number of auxiliary variables increases for this reason.

5.2 The SILC

For the SILC, we use data from the 2015 sample, where the response rate is 57 percent and the net sample size is about 9,200. We focus on two binary Y-variables: whether people find it difficult to make ends meet and whether they have poor health conditions, denoted by Y_{en} and Y_{he}, respectively.

The association measures of each available covariate with R, Y_{en} and Y_{he} are given in Table 6, together with B and RE by the respective subclass reweighting. It can be seen

Table 6: Association with (R, Y_{en}, Y_{he}), selected†, B in 10^{-2}

Auxiliary variable	λ_{cR}	$\lambda_{cY_{en}}$	$\lambda_{cY_{he}}$	B_{en} (RE)	B_{he} (RE)
Age	$0.00^\dagger$	$0.02^\dagger$	$0.01^\dagger$	0.01 (0.98)	0.12 (0.96)
Sex	$0.00^\dagger$	0.00	0.00	0.08 (0.98)	0.03 (0.99)
Region	$0.00^\dagger$	0.00	0.00	0.15 (0.98)	0.05 (0.99)
Family type	0.00	$0.02^\dagger$	0.00	-0.13 (1.00)	0.00 (1.00)
Birth country	$0.01^\dagger$	$0.02^\dagger$	0.00	-0.43 (1.05)	-0.04 (1.02)
Education	$0.04^\dagger$	0.01	$0.01^\dagger$	-0.39 (1.08)	-0.37 (1.13)
Marital status	$0.01^\dagger$	$0.03^\dagger$	$0.01^\dagger$	-0.31 (1.02)	0.07 (0.97)
Income	$0.03^\dagger$	$0.03^\dagger$	$0.02^\dagger$	-0.67 (1.08)	-0.42 (1.12)
Household income	$0.02^\dagger$	$0.06^\dagger$	$0.02^\dagger$	-0.93 (1.08)	-0.34 (1.10)

that here we are in a situation of only low association with both the outcome variables and nonresponse across the board. The covariates selected for the R-model and the two Y-models are marked (by †) in Table 6. No interaction terms are selected for any of the models based on these data. As in the case of LFS, largely the same variables are selected for both Y-models, denoted by X_{en} and X_{he}, respectively, and each of them includes the covariates that have the highest association with either Y_{en} or Y_{he}. The R-model includes all the available covariates (Z), except for family type that resembles marital status. While X_{he} is entirely nested in Z, X_{en} is almost so except for family type.

Table 7: SILC estimates (s.e) in 10^{-2}, year 2015, $V = Z \vee X_{en}$

Auxiliary for	Mean of Y_{en}, $\bar{y} = 11.20$ (0.39)		
(IPW, MGR)	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step Estimator
(Z, X_{en})	13.05 (0.44)	14.16 (0.46)	$\bar{y}_{2sts} = 14.68$ (0.48)
(V, V)	13.05 (0.44)	14.22 (0.46)	$\bar{y}_{DR} = 14.65$ (0.48)
Auxiliary for	Mean of Y_{he}, $\bar{y} = 6.07$ (0.30)		
(IPW, MGR)	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step Estimator
(Z, X_{he})	6.83 (0.35)	6.99 (0.35)	$\bar{y}_{2sts} = 7.00$ (0.36)
(Z, Z)	”	6.98 (0.37)	$\bar{y}_{DR} = 6.94$ (0.38)

The different estimators and their associated s.e's (in parentheses) are given in Table 7. Compared to the baseline estimates, reweighting leads to upwards adjustments for both Y_{en} and Y_{he}, and increases the variance in all the cases. Again, as exemplified in Section 4.3, the adjustment of the point estimate can be large, several times the s.e's here, despite the low association with both Y and R; whereas low association with Y does increase the variance. For both Y-variables, it can be seen that the one-step MGR and the two-step estimators are closer to each other than the one-step IPW estimators. In particular, the IPW estimators do not have larger variances, compared to any of the alternatives that includes calibration towards the selected population auxiliary totals. Notice that using RPS with 5 groups reduces the variance of $\bar{y}_{IPW}$ slightly, and it may somewhat change the point estimate, e.g. we would have $\bar{y}_{en} = 12.75$ (0.43) and $\bar{y}_{he} = 6.85$ (0.34) instead.

Regarding the three reweighting approaches the results suggest similar conclusions for the SILC as the LFS. The DR estimator using (V, V), for $V = Z \vee X$, does not offer any noticeable advantage compared to the traditional two-step approach using (Z, X) for the SILC. Neither does the two-step approach $\bar{y}_{2sts}$ using (Z, X) offer any trustworthy advantage over the one-step MGR using X. The variance of $\bar{y}_{2sts}$ is slightly larger than that of $\bar{y}_{MGR}$ for both Y-variables. The adjustment of the point estimate is similar in the case of Y_{he}, and about one s.e larger by $\bar{y}_{2sts}$ for Y_{en}. However, given the low association of the available covariates with nonresponse, the R-model is hardly the true nonresponse model. Indeed, given the low association with the Y-variables, it seems possible that the difference in the adjusted point estimates can be spurious.

The situation here, where one can only achieve low association with Y, can very well happen in many countries that have fewer auxiliary variables available than in Norway. It is often possible to find additional sample covariates that have higher association with nonresponse. For instance, given the rotating panel design of the SILC, one may introduce the Panel Response Status (PRS) as in Example 3 and 4 in Section 4.3, which has a higher association with R ($\lambda_{cR} = 0.20$) but almost no association with the two Y-variables ($\lambda_{cY_{en}} = 0.00$, and $\lambda_{cY_{he}} = 0.00$). The variable PRS has three categories indicating whether an individual is a previous respondent, previous nonrespondent, or is a new sample unit. Adding PRS as an extra covariate to Z given in Tabel 6 yields Z^* for the R-model.

Table 8: SILC estimates (s.e) in 10^{-2}, with Z^* for R-model

	One-step $\bar{y}_{IPW}$	Two-step $\bar{y}_{2sts}$	Two-step $\bar{y}_{DR}$
Mean of Y_{en}: $\bar{y} = 11.20$ (0.39)	14.39 (0.65)	15.57 (0.66)	15.43 (0.63)
Mean of Y_{he}: $\bar{y} = 6.07$ (0.29)	7.09 (0.43)	7.14 (0.43)	7.04 (0.44)

The new one-step IPW and two-step estimators using Z^* for the R-model are given in Table 8. The 500 bootstrap resamples are generated with the same design as the SILC but further stratified by whether an individual is a new sample unit or not. The most notable feature in Table 8 is that all the reweighting estimators produce greater point-estimate adjustments but also considerably larger variances, compared to the corresponding estimators without PRS in Table 7. A simple explanation is that PRS enhances the association with R without increasing the association with the two Y-variables. On the one hand, it is highly likely that the baseline $\bar{y}$ underestimates both proportions, since all the reweighting methods produce upwards adjustments. On the other hand, it is unclear whether the bias of any adjusted estimator may have gone from negative to positive, and the increased variances certainly suggest a heightened risk of introducing spurious adjustments.

The existing production method of the SILC is reweighting by about 200 subclasses, which are formed by cross-classifying several of the auxiliary variables considered here. Stablising the variance of estimation is therefore an important aspect for improvement. This speaks against including variables like PRS, because the affected estimators would have considerably larger variances. Recall that X_{en} and X_{he} are essentially nested in Z (Table 6). A possible resolution is to settle for a common set of variables, denoted by Q, and choose between the IPW and MGR estimators based on an overall assessment of their efficiency for different Y-variables. Two initial choices for Q are (i) the intersection $Q_0 = Z \wedge X_{en} \wedge X_{he}$, and (ii) the union $Q_1 = Z \vee X_{en} \vee X_{he}$. In addition, one can explore any of the 32 possible Q between Q_0 and Q_1, and obtain the corresponding IPW and MGR estimates that are given in Table 9.

Table 9: SILC estimates (s.e) in 10^{-2}, with different auxiliary variables

	Mean of Y_{en}		Mean of Y_{he}	
Variables	IPW (s.e)	MGR (s.e)	IPW (s.e)	MGR (s.e)
Q_0	12.68 (0.43)	13.27 (0.43)	6.61 (0.33)	6.81 (0.34)
Q_0, region	12.65 (0.43)	13.24 (0.43)	6.62 (0.33)	6.81 (0.34)
Q_0, sex	12.68 (0.43)	13.27 (0.43)	6.61 (0.33)	6.80 (0.34)
Q_0, birth country	12.79 (0.43)	14.06 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, education	12.91 (0.44)	13.37 (0.43)	6.83 (0.34)	6.99 (0.35)
Q_0, family type	12.70 (0.43)	13.37 (0.43)	6.61 (0.33)	6.81 (0.34)
Q_0, region, sex	12.65 (0.43)	13.24 (0.43)	6.62 (0.33)	6.81 (0.34)
Q_0, region, birth country	12.76 (0.43)	14.02 (0.46)	6.59 (0.33)	6.78 (0.35)
Q_0, region, education	12.89 (0.44)	13.36 (0.43)	6.84 (0.34)	7.00 (0.35)
Q_0, region, family type	12.67 (0.43)	13.35 (0.43)	6.63 (0.33)	6.81 (0.34)
Q_0, sex, birth country	12.79 (0.43)	14.06 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, sex, education	12.90 (0.44)	13.37 (0.43)	6.84 (0.34)	6.99 (0.35)
Q_0, sex, family type	12.70 (0.43)	13.37 (0.43)	6.61 (0.33)	6.80 (0.34)
Q_0, birth country, education	13.07 (0.44)	14.18 (0.46)	6.81 (0.34)	7.00 (0.37)
Q_0, birth country, family type	12.82 (0.43)	14.16 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, education, family type	12.91 (0.44)	13.45 (0.43)	6.82 (0.34)	6.99 (0.35)
Q_0, region, sex, birth country	12.76 (0.43)	14.02 (0.46)	6.59 (0.33)	6.78 (0.35)
Q_0, region, sex, education	12.88 (0.44)	13.36 (0.43)	6.85 (0.35)	7.00 (0.35)
Q_0, region, sex, family type	12.67 (0.43)	13.35 (0.43)	6.63 (0.33)	6.81 (0.34)
Q_0, region, birth country, education	13.06 (0.44)	14.14 (0.46)	6.81 (0.34)	6.97 (0.36)
Q_0, region, birth country, family type	12.78 (0.43)	14.12 (0.47)	6.59 (0.33)	6.78 (0.35)
Q_0, region, education, family type	12.89 (0.44)	13.44 (0.43)	6.84 (0.34)	7.00 (0.35)
Q_0, sex, birth country, education	13.07 (0.44)	14.18 (0.46)	6.82 (0.35)	7.00 (0.37)
Q_0, sex, birth country, family type	12.82 (0.43)	14.16 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, sex, education, family type	12.90 (0.44)	13.45 (0.43)	6.84 (0.34)	6.99 (0.35)
Q_0, birth country, education, family type	13.07 (0.44)	14.26 (0.46)	6.81 (0.34)	6.99 (0.37)
Q_0, region, sex, birth country, education	13.05 (0.44)	14.15 (0.46)	6.83 (0.35)	6.98 (0.37)
Q_0, region, sex, birth country, family type	12.78 (0.43)	14.12 (0.47)	6.59 (0.33)	6.78 (0.35)
Q_0, region, sex, education, family type	12.88 (0.44)	13.44 (0.43)	6.85 (0.35)	7.00 (0.35)
Q_0, region, birth country, education, family type	13.06 (0.44)	14.22 (0.46)	6.81 (0.34)	6.97 (0.36)
Q_0, sex, birth country, education, family type	13.07 (0.44)	14.26 (0.46)	6.82 (0.35)	7.00 (0.37)
Q_1	13.05 (0.44)	14.22 (0.46)	6.82 (0.35)	6.98 (0.37)

We observe the same pattern in Table 9 as previously, given low association with Y: the auxiliary variables Q that yield greater adjustment of the point estimates also lead to larger variances. The simplest choice here appears to be Q_0, which achieves the minimum s.e's for both the IPW and MGR estimators for both the Y-variables. Adding extra auxiliary variables does not improve the efficiency, but it may be accepted in practice, if benchmarking towards the extra variable is considered necessary and the induced adjustment and variance are judged reasonable. For example, region may be added to Q_0 to produce consistent regional estimates without losing efficiency or affecting much the point estimates.

6 Conclusions

Two interdependent decisions are required when reweighting for unit nonresponse: auxiliary variable selection and weight adjustment method. The following conclusions emerge from the review and empirical appraisal above.

When selecting the auxiliary variables, it is always useful to increase the association with the outcome variable, but seeking higher association with nonresponse is not necessarily helpful. In particular, one can achieve large useful adjustment of the point estimate and reduce the variance at the same time, provided high association with the outcome variable but only low association with nonresponse. While it is often possible to find variables that are primarily associated with nonresponse but not the outcome variables, such as the variable PRS in the LFS and SILC, caution would be necessary regarding such variables, because they tend to inflate the variance and heighten the risk of spurious adjustment, as it has been demonstrated empirically in Section 4.3 and 5.2.

Regarding weight adjustment, the choice of method does matter, e.g. between the one-step IPW and MGR estimators, especially when there exist strong auxiliary variables for the outcome available, as for the employment variable in the LFS. In particular, it would be unwise *only* to consider the IPW (or RPS) estimator based on a nonresponse model, when high association with the outcome variable is available. Provided weak auxiliary variables for the outcome variable, bigger adjustment of the point estimate is often accompanied by an increasing variance, by either the IPW or MGR estimator. Limiting the loss of efficiency and avoiding spurious adjustment may be the priority in such situations. Thus, it is important to pay attention not only to the size of adjustment of the point estimate by the weight adjustment method, but also the effects of reweighting on the variance of estimation, whether the given auxiliary variables are strong or weak.

Finally, regarding the three main reweighting approaches identified in Section 1, we found no evidence in the situations examined, which supports an uncritical general adoption of either the two-step approach. Neither the traditional nor the DR two-step approach yields empirically any gains for the Norwegian LFS and SILC. Since the ‘true’ nonresponse model envisaged for a two-step approach cannot be identified based on the observed data, whether the available auxiliary variables have low or high association with nonresponse, it makes sense to choose based on cross-examination of the alternatives in a given situation.

References

Andersson, P. G. and Särndal, C.-E. (2016) Calibration for nonresponse treatment: in one or two steps? *Statistical Journal of the IAOS*, **32**, 1–7.

Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.

Bethlehem, J., Cobben, F. and Schouten, B. (2011) *Handbook of Nonresponse in Household Surveys.* Hoboken: Wiley.

Bethlehem, J. G. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, **4**, 251–260.

Brick, M. J. (2013) Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, **29**, 329–353.

Canty, A. and Davison, A. (1999) Resampling-based Variance Estimation for Labour Force Surveys. *The Statistician*, **48**, 379–391.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Statist. Soc. A*, **169**, 571–584.

Cassel, C. M., Särndal, C.-E. and Wretman, J. H. (1983) Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* (eds. W. G. Madow, I. Olkin and D. B. Rubin), vol. 3, pp. 143–160. New York: Academic Press.

Cochran, W. (1953) *Sampling Techniques.* New York: Wiley.

Eltinge, J. L. and Yansaneh, I. S. (1997) Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, **23**, 33–40.

Haziza, D. and Lesage, É. (2016) A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, **32**, 129–145.

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

Kalton, G. and Kasprzyk, D. (1986) The treatment of missing survey data. *Survey Methodology*, **12**, 1–16.

Kang, J. D. Y. and Schafer, J. L. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, 523–539.

Kim, J. and Haziza, D. (2014) Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, 375–394.

Kott, P. S. and Liao, D. (2015) One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, **41**, 165–181.

de Leeuw, E. and de Heer, W. (2002) Trends in household survey nonresponse - a longitudinal and international comparison. In *Survey Nonresponse* (eds. R. Groves, D. Dillman, J. Eltinge and R. J. A. Little), pp. 41–54. New York: Wiley.

Little, R. J. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R. J. and Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**, 161–168.

Little, R. J. A. (1986) Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, 139–157.

Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statist. Med.*, **23**, 2937–2960.

Lundström, S. and Särndal, C.-E. (1999) Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, **15**, 305–327.

Meyer, B. D., Mok, W. K. C. and Sullivan, J. X. (2015) Household surveys in crisis. *Journal of Economic Perspectives*, **29**, 199–226.

Oh, H. L. and Scheuren, F. S. (1983) Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys* (eds. W. G. Madow, I. Olkin and D. B. Rubin), vol. 2, pp. 143–184. New York: Academic Press.

Robins, B. J. M. and Wang, N. (2000) Inference for imputation estimators. *Biometrika*, **87**, 113–124.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.

Särndal, C.-E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse.* Chichester: Wiley.

— (2008) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, **24**, 167–191.

— (2010) Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, **36**, 131–144.

Schouten, B. (2007) A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, **23**, 51–68.

Stoop, I., Billiet, J., Koch, A. and Fitzgerald, R. (2010) *Improving Survey Response: Lessons Learned from the European Social Survey.* Chichester, UK: Wiley.

Thomsen, I. (1973) A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift*, **11**, 278–285.

— (1978) A second note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift*, **16**, 278–285.

Thomsen, I. and Zhang, L.-C. (2001) The effects of using administrative registers in economic short term statistics: The Norwegian Labour Force Survey as a case study. *Journal of Official Statistics*, **17**, 285–294.

Zhang, L.-C. (1999) A note on post-stratification when analyzing binary survey data subject to nonresponse. *Journal of Official Statistics*, **15**, 329–334.

— (2005) On the bias in gross labour flow estimates due to nonresponse and misclassification. *Journal of Official Statistics*, **21**, 591–604.

Zhang, L.-C., Thomsen, I. and Kleven, Ø. (2013) On the use of auxiliary and paradata for dealing with non-sampling errors in household surveys. *International Statistical Review*, **81**, 270–288.