

The Effects of Non-Response Reweighting on Estimators' Precision in a Web Survey

Silvia Biffignandi, Enrico Fabrizi and Daniele Toninelli¹

Abstract This paper analyzes the reweighting adjustments for non-response in web surveys carrying out a bootstrap evaluation of non-response adjusted estimators. We consider a target population of graduated students of Bergamo University. Students were surveyed on many topics, including student satisfaction for the university experience, before (BGS) and after (PGS) the graduation. Administrative data are also available for all the students. We firstly evaluate if the non-response process is Missing At Random (MAR) and then we focus on the estimators of the total for some target variables (factors computed on the satisfaction variables). We compare the ability of different correction methods (quasi-randomization, calibration, combined method) in improving the estimators of non-response.

Key words: web survey, non-response, missing data, calibration, bootstrap, sample design and estimation

1 Introduction

The aim of this paper is to study the effects of reweighting corrections for non-response on the precision of estimators of population totals with reference to web surveys. We compare various reweighting options in a controlled environment as close as possible

¹ Silvia Biffignandi, DMSIA, Bergamo University, via dei Caniana 2, 24127 Bergamo, Italy, email: silvia.biffignandi@unibg.it

Enrico Fabrizi, DISES, Catholic University, via Emilia Parmense 84, 29122 Piacenza, Italy, email: enrico.fabrizi@unicatt.it

Daniele Toninelli, DMSIA, Bergamo University, via dei Caniana 2, 24127, Bergamo, Italy, email: daniele.toninelli@unibg.it

The present paper is financially supported by: Bergamo University (60%/2008 Silvia Biffignandi;60%/2009, Daniele Toninelli).

to a real situation. Our analysis relies on a simulation based on data collected from two twin web surveys on a population of students from the Bergamo University. The two surveys were conducted on the same cohort of students just before graduation (*Before Graduation Survey, BGS*) and some months later (*Post Graduation Survey, PGS*). In particular, each survey is a census of the target population. The participation in the first survey was compulsory: every student that was going to be graduated was asked to fill the on-line questionnaire as part of the administrative procedures to be accomplished before graduation. Some months later, the whole original group of students was contacted by email and invited to complete an on-line questionnaire (PGS). The participation to the PGS was not compulsory; this explains the massive non-response (we obtained a response rate of about 44%).

In the economy of the simulation we are imposing the non-response pattern of the second survey to the analysis of data from the first one. This allows us to have data on the target variables for both respondents and non-respondents. Moreover, as the two surveys were conducted with the same mode on the same population and just few months apart, the assumption on the non-response process is adequately realistic. Register auxiliary variables, known for both respondent and non-respondents, are available; they are used in the non-response correction process. As target variables we consider two factors obtained applying the factor analysis to a battery of questions about the student's satisfaction of the University experience. Non-response may bias the inferences on population descriptive quantities whenever data are not Missing Completely At Random (MCAR, see Little and Rubin, 1987). Under the assumption of data "missing at random" (that is when the non-response indicator variable is independent of the target variable conditional on a set of auxiliary variables) a series of reweighting methods for the reduction of the bias have been proposed in the literature. Reweighting methods rely on some assumption about the non-response process (Zhang, 2002) and effectively reduce bias when the auxiliary variables are powerful in discriminating respondents and non-respondents. Even when the "missing at random" assumption does not hold, useful adjustments can be obtained due to the correlation between the auxiliary and the survey variables (Zhang, 1999). In view of the particular structure of our experiment we are in the position of testing the critical assumption of missing at randomness for the non-response process; moreover we may also test whether questionnaire variables other than the target ones have an impact on non-response. The reweighting methods that we compare in the simulation are: 1) correction based on the estimation of the response propensities (Oh and Scheuren, 1983); 2) calibration (Deville and Sarndal, 1992; Kott, 2006); 3) calibration of propensity-adjusted weights (Lunstrom and Sarndal, 2005; Lee and Valliant, 2009). A concise description of these methods may be found in section 3.

The hypotheses we are testing are: *i*) missing data are completely random; *ii*) non-response propensity and/or calibration adjustment are beneficial with respect to no correction at all if the auxiliary variables have power in explaining non-response. The paper is organized as follows: in section 2 we describe the data set, the target variables and the non-response process of the PGS. Section 3 offers a concise review of the reweighting based methods for correcting non-response that are relevant in this paper. Section 4 introduces the simulations and presents the results. Section 5 offers some concluding remark.

2 Data and target variables

Our analysis uses a simulation based on a real dataset consisting in the results of the two mentioned web surveys. More specifically, our population refers to 459 students from the Bergamo University graduated between September 2003 and April 2004.

The BGS survey covers several topics like the evaluation of various aspects of the university experience, the job experiences, the family background, the participation in student international exchanges programs and other subjects.

The PGS survey collects data about the position of the graduate on the job market, its satisfaction, training activities, specialized advanced studies, and so on. Satisfaction about the compliance of University training and requirements of the job market and help received from the job placement bureau of the University were also investigated.

The target variables are two factors drawn from the BGS survey data. More specifically, we focus on a battery of nine questions concerning students' satisfaction about the University experience. They are continuous; they have mean equal to 0 and their distribution are not far from normality (skewnesses are, respectively, 0.55 and 0.34; kurtosis are 1.12 and 0.14 respectively). The two factors are quite powerful in explaining the total variation and they can be rather easily interpreted. The first factor (Factor 1) may be broadly interpreted as an overall satisfaction factor, while the second one (Factor 2) may be viewed as a contrast between the evaluation of the teaching quality and of the physical facilities of the University. The communalities show that not all the variables are well accounted for by two components. The explained variance ranges from 50% (for the evaluation of equipments and study load) to a little less than 30% (for the "spaces for studying"). Our objective is the estimation of the total of the two factors (i.e. the target variables) under different non-response adjusting methods.

2.1 *The non-response process*

To analyze the non-response process, we apply a logistic regression analysis. The dependent variable is the response indicator while as regressors we consider the following variables: a) register variables (age, gender, time between the enrolment and the graduation, foreign country study experience, faculty, presence of a personal email address in the archive, high school's type and final mark, exams average mark); b) a selection of variables included in the BGS questionnaire (software and foreign languages self-perceived knowledge, work experiences before and during the university, percentage of class attendance, parents' job position and education level); c) Factor 1 and Factor 2. The factors are included in the regressors' set because we are interested in testing whether data lost due to non-response are missing at random. If this factors turn out to be significant, this means that the response indicator is not independent of the target variables after controlling for the other auxiliary variables.

Our model shows that the response process remains largely unexplained. The pseudo- R^2 is around 0.15. More specifically, some of the register variables (faculty, availability of an email address other than the University one, number of years between

enrolment and graduation) are significant; no variables coming from the BGS questionnaire data are significant. This implies that we can base our correction using register variables that are actually available for all the students. Moreover, the target variables, are not significant at the $\alpha = 0.05$ significance level, so the MAR assumption cannot be rejected. In fact, the situation is close to the borderline. We do not have a strong evidence against the hypothesis that the response indicator is independent of the target variables, conditionally on the auxiliary variables. However there is some weak evidence that target variables are to some extent distributed differently for respondents and non-respondents (see table 1).

Table 1: Lower bound (L), Mean (M) and Upper bound (U) for Factor 1 and Factor 2 (95% confidence level)

Group	$L_{Factor1}$	$M_{Factor1}$	$U_{Factor1}$	$L_{Factor2}$	$M_{Factor2}$	$U_{Factor2}$
Respondents	-0.056	0.074	0.204	-0.068	0.058	0.183
Non-respond.	-0.186	-0.059	0.069	-0.176	-0.046	0.084

2.2 Prediction of the target variables

The effectiveness of weighting correction methods depends also on the availability of powerful predictors of the target variables. To this end, we analyze the relationships between Factor 1 and Factor 2 and the set of register variables potentially useful for weights' correction, along with the variables stemming from the BGS questionnaire.

Factor 1 turns out to be almost unpredictable, unless variables related to satisfaction itself are included into the model. This results hold if we include satisfaction variables both used or not used in building the factor itself. Using satisfaction variables considered to compute the factor is not advisable, as it would imply a certain degree of circularity. Even in this last case the correlation is very low. Excluding these satisfaction variables, the only significant variable is the dummy that indicates the low level of education of the student's mother. The R^2 obtained from the linear regression based on this variable (that may be seen as the outcome of formal a stepwise method of selection) is only 0.0162. For Factor2 several variables are significant at the $\alpha = 0.05$ level ($R^2 = 0.285$): faculty dummies, graduation mark, high school type and, among the BGS variables, the work of father and mother and the self-perceived software knowledge level. Anyway, the exclusion of the variables stemming from the BGS questionnaire reduces R^2 to 0.245, so we will omit these variables in re-weighting correction methods.

3 Reweighting methods

Oh and Scheuren (1983) provide a justification for many reweighting methods by treating non-response as a second phase of sampling. If we denote r the subset of respondents in the sample s taken from population U , we may write:

$$\Pr(k \in r) = \Pr(k \in s) \Pr(k \in r | k \in s) = \pi_k \theta_k \quad \forall k \in U,$$

where the π_k s are the randomization inclusion probabilities and θ_k s are the response propensities. A first reweighting approach is called quasi-randomization; it is not model-free but relies on a model for the response propensities. The unknown θ_k s should be estimated, typically assuming that $\theta_k = p(\mathbf{x}_k^t \beta)$, where p is an assumed known function and $\mathbf{x}_k^t$ is a vector of auxiliary variables (Chang and Kott, 2008).

In many applications, θ_k s are implicitly estimated using calibration algorithms. The original purpose of calibration (Deville and Sarndal, 1992) is to obtain a set of weights w_k calibrated on known population totals of the auxiliary variables ($\sum_k w_k \mathbf{x}_k = \mathbf{X}$) with a minimum modification of basic survey weights ($d_k = \pi_k^{-1}$) with respect to some distance criterion. The calibration may be used to adjust for non-response (Lundstrom and Sarndal, 2005; Kott, 2005). It is expected to be effective in reducing the non-response bias if the calibration set includes good predictors of the response indicator variable. On the other hand, improvements in precision of estimators of population descriptive quantities for the target variable are obtained when variables correlated with the target one are included in the calibration set. Vartivarian and Little (2003), although working on non-response reweighting based on adjustment cells, a special case of calibration, propose to form the cells cross-classifying the sample with respect to both the predictor of the target variable and the response propensity. Using a similar argument, you can think of forming your calibration set with the union of the subset of variables including relevant predictors for the response predictors and the subset of variables that are good for linearly predicting the target variable. This approach may be criticized according to the following arguments: *i*) including variables with low correlation with the target variable in the calibration set (as may be the case if they are relevant only for predicting the response propensity) may lead to an increase in the weights' variability and thus to reduced efficiency that over-compensate the gain due to bias reduction; *ii*) as noted in Chang and Kott (2008), when calibration is used, it is implicitly assumed that $p(\mathbf{x}_k^t \beta) = (1 + \mathbf{x}_k^t \beta)^{-1}$ (see also Fuller et al., 1994); moreover the estimation of the unknown may be sub-optimal with respect to maximum likelihood or other estimation methods. The criticisms may be circumvented using a two step approach, recently advocated by Lee and Valliant (2009) in the context of web surveys. It consists in estimating the propensity θ_k using only the variables that are good for predicting it (estimation methods depending on the available information and thus including propensity score matching); then $d_k^* = (\pi_k \hat{\theta}_k)^{-1}$ are used as initial weights for a calibration step where variables with (possibly) strong correlation with the target one are included. This approach leads to a combination of the two methods (calibration of quasi-randomization adjusted weights).

4 Simulation and results

We evaluate the performance of estimators based on various re-weighting methods using a bootstrap simulation. The synthetic population is provided by the BGS dataset.

As it is assumed to be a census (except for the effect of non-response) bootstrap replicated samples are obtained by drawing a simple random sample from the BGS data having exactly the same size. Subsequently the non-response pattern of the PGS is applied to the students included in the bootstrap samples. The totals of Factor 1 and Factor 2 are considered the target estimates (totals are equal to 0 by construction).

Let n_r be the number of respondents, and y_{ak} the k -th sample value of target variable y_a (a = Factor 1, Factor 2). N is the population size. The estimators that we are going to compare are: *Unadjusted (UNW)*¹, *Quasi-Randomization Adjusted (QRA)*², Calibration (*CAL_j*). Several calibration estimators are considered depending on various composition of the calibration set³. The various choices for the d_j weights⁴ and for the S_j calibration set are described in table 2.

Table 2: Calibration estimators labels, initial weights and calibration sets

<i>Estimator</i>	d_{jk}	S_j
CAL1	1	course variables (type of course, faculty dummies)
CAL2	1	course + non-response explaining variables
CAL3	1	course + non-response, factors explaining variables
CAL4	1	course + non-response, factors explaining variables (presence of private email address in the register excluded)
CAL5	1	course + factors explaining variables
CAL6	$\hat{\theta}_k^{-1}$	course + factors explaining variables

The calibration set of CAL1 is motivated by the fact that the “course variables” identify the most relevant domains of the population. CAL2 is the estimator based on the same information of QRA, but considering a different function p and estimation method for the response propensity. CAL3 includes the variables useful for explaining both the non-response and the target variables⁵. CAL4 does the same but excluding the variable “availability of a private email”; this variable is responsible for most variation in the weights without being an effective predictor of the target variables. CAL5 includes variables that are significant predictor of Factor 2 only. CAL6 represents the calibration after the propensity correction.

$$^1 \quad UNW: \hat{t}_{a,UNW} = \frac{N}{n_r} \sum_{k \in S_r} y_{ak}.$$

$$^2 \quad QRA: \hat{t}_{a,QRA} = \sum_{k \in S_r} \frac{y_{ak}}{\hat{\theta}_k}, \text{ where } \hat{\theta}_k = \hat{\theta}_k(\mathbf{x}_k^t; \beta) \text{ is the propensity estimated using the}$$

logistic regression and the variables described in section 2.1.

$$^3 \quad \text{General notation for calibration estimators: } \hat{t}_{a,CALj} = \sum_{k \in S_r} g_{jk} d_{jk} y_{ak}, \text{ where } g_{jk} \text{ is the}$$

calibration modifier of initial weights d_{jk} (Deville and Sarndal, 1992) associated to the calibration set S_j .

⁴ Information on the non-response process is incorporated in the initial weights of calibration. In all cases the calibration distance adopted is the logarithmic, thus leading to unconstrained generalized raking estimation (Deville and Sarndal, 1992).

⁵ Since Factor 1 was unpredictable we consider the variables predicting Factor 2 only (sec. 2.2).

The tables 3 shows the results referred to the estimators introduced in the previous section for Factor 2¹. They are based on $B = 2,000$ bootstrap samples.

Table 3: Factor 2: results of the bootstrap samples

<i>Estimator</i>	<i>Bias (B)</i>	<i>Variance</i>	<i>MSE</i>	<i>100(B/MSE)</i>
UNW	25.8	871.8	1535.6	43.2
QRA	18.2	922.6	1254.5	26.4
CAL1	21.1	722.6	1169.7	38.2
CAL2	25.4	826.8	1473.8	43.9
CAL3	23.2	782.0	1318.9	40.7
CAL4	17.1	721.7	1013.2	28.8
CAL5	13.8	693.8	884.9	21.6
CAL6	6.2	740.7	778.6	4.8

The impact of the proposed corrections in reducing the variance is not impressive. Two forces act in opposite directions: on one hand calibrating potentially reduces the variance, since we dispose of significant linear predictors of the factor; on the other hand, seen the weak relation between the auxiliary and the target variable, the variability of weights inflates the variance. This is apparent in the case of CAL2, where the calibration set is composed only by variables explaining the non-response process. The bad performances of CAL2 (if compared to QRA) confirm the theoretical expectation that estimating the propensity with a plausible model (logistic regression) and efficient estimation method (maximum likelihood) leads to better results with respect to the use of the same information via calibration. Including in the calibration set the variables significant for explaining non-response turns out to be not very helpful in reducing bias and increases variance. For this reason among the calibration methods with constant initial weights CAL5 turns out to be the best. The estimator CAL6 (calibration on the target variable starting from quasi-randomization adjusted weights) emerges clearly as the best option, in terms of both bias and variance.

Results for Factor 1 (not reported for brevity) depend heavily on the lack of significant predictors for it. All calibrations seem useless, if not detrimental. The inclusion of variables explaining the non-response process is effective in reducing bias; the best way of using this information is, again, via the logistic regression implemented in the QRA estimator. The QRA estimator turns out to be the best option; however CAL6 shows very close performances, although including variables that are not useful in explaining Factor 1.

5 Conclusions

In this paper we compared various reweighting methods aimed at reducing the effects of non-response in the context of a web survey.

¹ Note that an expansion estimator associated with an ideal simple random sample (without non-response) of the same size of our bootstrap samples estimators has a MSE equal to 1035. Therefore estimators with MSE above this threshold are less efficient.

With respect with the theoretical expectation of section 1 we found that:

i) The MAR assumption cannot be rejected. If we consider Factor 1, no one of the considered variables is significantly correlated with it. The situation is better for Factor 2. In this case the set of variables relevant to explain non-response and those with significant correlation with Factor 2 partially overlap.

ii) When variables with some power in explaining non-response are available, the re-weighting correction based on estimated propensity and/or calibration is better than no correction at all; this is true also if data allows only for a limited explanation of the non-response (small pseudo R^2 in the logistic regression).

In the case of Factor 1 no calibration correction is beneficial, while in the case of Factor 2 calibration correction applied to propensity-adjusted weights is more effective than response propensity correction and calibration alone. Generally speaking, we can conclude that calibration correction applied to propensity-adjusted weights is the best way of combining the information contained in auxiliary variables explaining the non-response process and the target variables.

References

1. Chang, T., Kott, P.S.: Using calibration under a plausible model. *Biometrika* **95**, 557--571 (2008)
2. Deville, J.C., Sarndal, C.E.: Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376--382 (1992)
3. Fuller, W.A., Loughin, M.M., Baker, H.D.: Regression weighting for the 1987-1988 National Food Consumption Survey. *Survey Methodology* **20**, 75--85 (1994)
4. Kott, P.S.: "No" is the easiest answer: using calibration to assess nonignorable nonresponse in the 2002 Census of Agriculture. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association, Washington D.C (2005)
5. Kott, P.S.: Using calibration to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133--142 (2006)
6. Lee, S., Valliant, R.: Estimation for volunteer panel web surveys using propensity scores and calibration adjustment. *Sociological Methods and Research* **37**, 319--343 (2009)
7. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*. John Wiley and Sons, New York (1987)
8. Lundstrom, S., Sarndal, C.E.: *Estimation in Surveys with Non Response*. John Wiley and Sons, New York (2005)
9. Oh, H.I., Scheuren, F.J.: Weighting adjustment for unit nonresponse. In: *Incomplete Data in Sample Survey*. Academic Press, New York (1983)
10. Vartivarian, S., Little, R.J.A.: On the Formation of Weighting Adjustment Cells for Unit Nonresponse. In: *The University of Michigan Department of Biostatistics Working Paper Series*. Working Paper 10 (2003). Available via web. <http://www.bepress.com/umichbiostat/paper10>
11. Zhang, L.C.: A note on post-stratification when analyzing binary survey data subject to nonresponse. In: *Journal of Official Statistics* **15**, 329--334 (1999)
12. Zhang, L.C.: A method of weighting adjustment for survey data subject to non-ignorable nonresponse. In: *DACSEIS Research Paper Series* **2** (2002)