Handling nonignorable nonresponse using generalized calibration with latent variables

M. Giovanna Ranalli[1], Alina Matei[2], Andrea Neri[3],
[1] Dept. of Economics, Finance and Statistics, University of Perugia, Italy
e-mail: giovanna.ranalli@stat.unipg.it
[2] Institute of Statistics, University of Neuchâtel and IRDP, Neuchâtel, Switzerland
[3] Bank of Italy, Rome, Italy

Abstract

Calibration weighting has been usefully employed to adjust for unit nonresponse. Generalized calibration allows to distinguish among auxiliary variables between those that are useful to model unit nonresponse (instrumental or model variables) and those that are used in the calibration constraints (calibration variables). Since model variables need only be known on the respondents, generalized calibration offers a particularly useful tool to deal with nonignorable nonresponse. Response to a survey is the outcome of a complex process that involves several aspects: we assume that a part (or all) of such a process may be measured by unobservable variables. Latent variable models can be employed to extract either continuous constructs (latent trait models) or categorical ones (latent class models) from a set of dichotomous/ordered manifest variables. We propose to use such constructs as instrumental variables in the generalized calibration procedure. This allows to include variables of interest among the set of manifest variables. The properties of the proposed methodology are illustrated, then it is tested on a series of simulation studies and finally applied to adjust estimates from the Italian Survey of Households Income and Wealth.

Keywords: auxiliary information, finite population, instrumental variables, reverse approach.

1. Introduction

In this work we focus on the treatment of unit nonresponse in surveys when estimation of finite population totals of a set of variables of interest is of concern. Let $U = \{1, \ldots, k, \ldots, N\}$ be the set of labels identifying a finite population of interest and let $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{mk}, \ldots, y_{Mk})$ be the value taken on unit k by the M-vector of variables of interest $\boldsymbol{y}$. We are interested in estimating its population total $\boldsymbol{t}_y = \sum_U \boldsymbol{y}_k$. We use the shorthand $\sum_A$ for $\sum_{k \in A}$, with $A \subseteq U$ an arbitrary set. To this end a sample s of dimension n is selected from U using a sampling design $p(s)$ with first (second) order inclusion probabilities π_k (π_{kl}). Let $I_k = 1$ be the indicator variable for unit k selected in the sample, so that $P(I_k = 1) = \pi_k = E\{I_k\}$. Let $d_k = \pi_k^{-1}$ and $d_{kl} = \pi_{kl}^{-1}$.

Unit nonresponse occurs and we denote by $r \subseteq s$ the set of respondents. Let the response indicator $R_k = 1$ if unit $k \in r$ and 0 otherwise. Thus $r = \{k \in s | R_k = 1\}$ and the response mechanism is given by the distribution $q(r|s)$. Usually these random indicator variables are assumed independent of one another and of the sample selection mechanism, and we do the same here. Then, the probability of responding to the survey is given by $p_k = P(R_k = 1 | k \in s)$. This is the two-phase approach where the sampling design is the first phase and the response mechanism is the second phase and is based on the quasi-randomization model of Oh and Scheuren (1983).

Unit nonresponse is well known to harm the quality of estimates from a sample survey any time those who respond are with some respect different from those who do not respond. Successful reduction of nonresponse bias may be achieved using powerful auxiliary information coupled with a well-specified model either for unit response probabilities (response model) or for the variables of interest (superpopulation models). In the survey setting, auxiliary information takes the form of a set of auxiliary variables $\boldsymbol{x}$ whose value $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{qk})$ is known for $k \in r$ and whose population total $\boldsymbol{t}_x$ is known or can be unbiasedly estimated with the Horvitz-Thompson estimator $\hat{\boldsymbol{t}}_x^{\mathrm{HT}} = \sum_s d_k \boldsymbol{x}_k$. Let $\boldsymbol{t}_x^\star$ denote the population total of $\boldsymbol{x}$ or its Horvitz-Thompson estimator accordingly.

Calibration is a general tool to include auxiliary information at the estimation stage (Deville and Särndal, 1992) and has been shown to be very useful in treating unit nonresponse as well (see e.g. Särndal and Lundström, 2005). It pursues the construction of a single set of weights w_k for all variables of interest modifying specified initial weights (usually the d_k's), while satisfying benchmark constraints on known auxiliary information, i.e.

$$\sum_r w_k \boldsymbol{x}_k = \boldsymbol{t}_x^\star. \tag{1}$$

No explicit model is specified for the treatment of the nonresponse mechanism; it is implicitly given by the calibration procedure. In fact, in general the probability of response p_k is assumed to be the inverse of a given *link* function of an unknown (but estimable) linear combination of auxiliary variables, i.e. $w_k = d_k F(\boldsymbol{x}_k\boldsymbol{\gamma})$. Given $F(\cdot)$ and estimated $\boldsymbol{\gamma}$ via the constraints given in (1), the calibration estimator is $\hat{\boldsymbol{t}}_y^{\text{CAL}} = \sum_r d_k F(\boldsymbol{x}_k\hat{\boldsymbol{\gamma}})\boldsymbol{y}_k$.

As discussed so far, no discrimination is made within the set of auxiliary variables available: a single set of variables is employed at the same time for nonresponse treatment, sampling error reduction and coherence among estimates. There are cases, however, in which one has a good reason to believe that nonresponse depends on a set of p *model* variables $\boldsymbol{z}$, whose components need not coincide with the components of the *calibration* vector $\boldsymbol{x}$. This can be accommodated in the calibration framework using "generalized" calibration introduced by Deville (2000) (see also Kott, 2006) in which weights are given by $w_k = d_k F(\boldsymbol{z}_k\boldsymbol{\gamma})$ and satisfy (1). Borrowing from the econometric literature, the components of $\boldsymbol{z}$ that are not linear combinations of $\boldsymbol{x}$ are called *instrumental variables*. Then, given $F(\cdot)$ and estimated $\boldsymbol{\gamma}$, the generalized calibration estimator is $\hat{\boldsymbol{t}}_y^{\text{GCAL}} = \sum_r d_k F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}})\boldsymbol{y}_k$. Note that the value of the model variables needs to be known only for $k \in r$ to compute $\hat{p}_k = F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}})^{-1}$. For this reason generalized calibration is particularly useful for nonresponse treatment, because unlike other reweighting methods, it allows to deal with it even when the variables that cause nonresponse are known only for the respondents. This is particularly relevant when nonresponse is nonignorable, i.e. when the topic of the survey and, therefore, the variables of interest influence the response probability of a unit. In fact, it is possible to introduce the variables of interest (known only on the respondents) as instrumental variables for correcting this type of nonresponse (Deville, 2000; Kott and Chang, 2010).

In this paper we work within the framework of generalized calibration and try to select a plausible set of instrumental variables to deal with nonignorable nonresponse. In particular, we move from noticing that response to a survey is the outcome of a complex process that involves several aspects, from the topic of the survey and fieldwork organization to some of the variables of interest or of the auxiliary variables. We assume that a part of such a process depends by unobservable variables. Latent variable models can be employed to extract either continuous constructs (latent trait models) or categorical ones (latent class models) from a set of dichotomous/ordered manifest variables (an introduction to these methods is provided in Section 2). This type of latent variables is particularly relevant, although not limited, to attitude and behavioral surveys and provide a measure of unobservable variables (like the "attitude towards politics" or other sensible topics) that likely influence the "willingness to respond" of a unit. We, therefore, propose to use such constructs as instrumental variables in the generalized calibration procedure. This allows to include variables of interest among the set of manifest variables and also for the construction of a single set of weights. The proposed methodology is introduced in Section 3, where its properties are illustrated and variance estimators proposed, then tested on a series of simulation studies (Section 4) and finally applied to adjust estimates from the Italian Survey of Households Income and Wealth (Section 5).

2. Latent variable models

Latent variable models are multivariate regression models that link continuous or categorical manifest, response variables to unobserved, latent variables. We focus here on categorical responses. According to the nature of the latent phenomenon, we can distinguish between *latent trait* and *latent class* models. A latent trait refers to a latent continuum which all individuals, based on their

pattern of responses on a set of observed variables, are mapped on. Latent classes, on the other hand, refer to the categories of a latent variable that is discrete; such categories may but need not be ordered along a continuum.

A latent trait model is essentially a factor analysis model for categorical data (see Bartholomew et al., 2002). For simplicity, we consider binary responses and let $\boldsymbol{\omega}_k = (\omega_{1k}, \ldots, \omega_{\ell k}, \ldots, \omega_{Lk})$ be the vector of these manifest variables observed for $k \in r$. Components of the manifest vector may include components of the response variable vector $\boldsymbol{y}$ and/or of the model vector $\boldsymbol{z}$. Denote by $q_{\ell k} = Pr(\omega_{\ell k} = 1|\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{jk}, \ldots, \theta_{Jk})$ is the value taken on unit k by the vector of $J < L$ latent variables computed from $\boldsymbol{\omega}_k$. The latent trait model is defined as $\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \sum_{j=1}^{J} \beta_{\ell j}\theta_{jk}$, for $\ell = 1, \ldots, L$, where $\beta_{\ell 0}, \ldots, \beta_{\ell J}$ are model parameters. An important special case is obtained by taking $J = 1$, i.e. a unidimensional latent trait model

$$\text{logit}(q_{\ell k}) = \beta_{\ell 0} + \beta_{\ell 1}\theta_k, \tag{2}$$

where it is usually assumed that $\theta_k \sim N(0, 1)$. Model (2) is also referred to as a two parameter logistic Rasch model, and is essentially a logistic regression except that the θ_k's are not observed.

In Latent Class Models, on the other hand, the latent variable is supposed to be discrete (Lazarsfeld and Henry, 1968; Goodman, 1974). Let the latent class variable of unit k be denoted by ϑ_k, a particular latent class by c and the number of latent classes by C. The full vector of responses of unit k is again $\boldsymbol{\omega}_k$, whilst $\boldsymbol{h} = (h_1, \ldots, h_L)$ refers to a possible response pattern. Then the latent class model can be expressed as $P(\boldsymbol{\omega}_k = \boldsymbol{h}) = \sum_{c=1}^{C} P(\vartheta_k = c)P(\boldsymbol{\omega}_k = \boldsymbol{h}|\vartheta_k = c)$, in which the probability of observing a response pattern $\boldsymbol{h}$ is a weighted average of class-specific probabilities. The number of latent classes may be chosen using model selection criteria like AIC, BIC or cAIC. Classification of units in latent classes provides an alternative way of building response homogeneity groups to deal with nonresponse. In this case our latent variable $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{Ck})$ is the indicator variable vector such that $\theta_{ck} = 1$ if $\vartheta_k = c$.

3. Generalized calibration with latent variables

Once an estimate $\hat{\boldsymbol{\theta}}_k$ of the latent variable $\boldsymbol{\theta}_k$ is obtained, either using latent trait or latent class models, it can be used in a generalized calibration framework as the instrumental variable $\boldsymbol{z}$ or as a component of the vector of instrumental variables. Therefore, let $\boldsymbol{z}_k = \hat{\boldsymbol{\theta}}_k$ or $\boldsymbol{z}_k = (\hat{\boldsymbol{\theta}}_k, \boldsymbol{z}_k^0)$ (where $\boldsymbol{z}_k^0$ represent instrumental variables different from $\hat{\boldsymbol{\theta}}_k$) and $w_k = d_k F(\boldsymbol{z}_k\boldsymbol{\gamma})$. If we ignore estimation of $\boldsymbol{\theta}_k$, the properties of $\hat{\boldsymbol{t}}_y^{\text{GCAL}} = \sum_r w_k \boldsymbol{y}_k$ are those illustrated, e.g., in Särndal and Lundström (2005) or Kott (2006). It is a consistent estimator for $\boldsymbol{t}_y$ if $p_k = F(\boldsymbol{z}_k\boldsymbol{\gamma})^{-1}$ or, alternatively when $\boldsymbol{t}_x^\star = \boldsymbol{t}_x$, if $E\{y_{mk}|\boldsymbol{x}_k\} = \boldsymbol{x}_k\boldsymbol{\beta}_m$, for $m = 1, \ldots, M$. In addition, all choices of $F(\cdot)$ are asymptotically equivalent to the linear one, and this is helpful for variance estimation.

Under the two phase approach, we can estimate the variance of $\hat{t}_{ym}^{\text{GCAL}}$ using

$$\hat{V}_{2\text{p}}(\hat{t}_{ym}^{\text{GCAL}}) = \hat{V}_{\text{sam}} + \hat{V}_{\text{nr}} = \sum_{k\neq l}\sum_{k,l\in r} d_{kl}\Delta_{kl}\frac{\hat{e}_{mk}^\star}{\hat{p}_k}\frac{\hat{e}_{lk}^\star}{\hat{p}_l} + \sum_r d_k\Delta_{kk}\frac{(\hat{e}_{mk}^\star)^2}{\hat{p}_k} + \sum_r \frac{(1-\hat{p}_k)}{\hat{p}_k^2}(d_k\hat{e}_{mk})^2,$$

where $\hat{e}_{mk}^\star = y_{mk}$ if $\boldsymbol{t}_x^\star = \hat{\boldsymbol{t}}_x^{\text{HT}}$ or $\hat{e}_{mk}^\star = \hat{e}_{mk} = y_{mk} - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}_m$ if $\boldsymbol{t}_x^\star = \boldsymbol{t}_x$, $\hat{\boldsymbol{\beta}}_m = (\sum_r w_k\boldsymbol{z}_k^T\boldsymbol{x}_k)^{-1} \times \sum_r w_k\boldsymbol{z}_k^T y_{mk}$ is an estimate of the instrumental variable coefficient vector of the regression of y_m on $\boldsymbol{x}$, and $\hat{p}_k = F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}})^{-1}$ (see e.g. Särndal and Lundström, 2005, Section 11.1).

Obtaining $\hat{V}_{\text{sam}}$ for a particular design may be cumbersome because one has to start from the respondents set. The "reverse" approach proposed by Shao and Steel (1999) simplifies considerably the task. In general, since the response mechanism is independent of the sampling process, the two can be exchanged in the decomposition of the variance. Using the reverse approach, we can estimate the variance of $\hat{t}_{ym}^{\text{GCAL}}$ using $\hat{V}_{\text{rev}}(\hat{t}_{ym}^{\text{GCAL}}) = \hat{V}_1 + \hat{V}_2$, where $\hat{V}_1 = \sum\sum_r d_{kl}\Delta_{kl}\hat{\eta}_{mk}^\star\hat{\eta}_{ml}^\star$, with $\hat{\eta}_{mk}^\star = \boldsymbol{x}_k\hat{\boldsymbol{\beta}}_m + R_k(y_{mk} - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}_m)$ if $\boldsymbol{t}_x^\star = \hat{\boldsymbol{t}}_x^{\text{HT}}$ or $\hat{\eta}_{mk}^\star = R_k\hat{e}_{mk} = R_k(y_{mk} - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}_m)$ if $\boldsymbol{t}_x^\star = \boldsymbol{t}_x$, and $\hat{V}_2 = \sum_r d_k(1-\hat{p}_k)\hat{e}_{mk}^2$.

Note that the two variance estimators considered above require computation of coefficients and residuals for each variable of interest. Since we are usually interested in estimating the total of

a whole set of variables of interest, a replication based variance estimator is also considered. In particular, let a jackknife variance estimator be $\hat{V}_{\text{jack}}(\hat{t}_{ym}^{\text{GCAL}}) = \frac{n_r-1}{n_r}\sum_{l\in r}\left({}^{(l)}\hat{t}_{ym}^{\text{GCAL}} - \hat{t}_{ym}^{\text{GCAL}}\right)^2$, where ${}^{(l)}\hat{t}_{ym}^{\text{GCAL}} = \sum_{k\in r}{}^{(l)}w_k y_{mk}$, n_r is the size of r and the jackknife replicate weights are given by

$${}^{(l)}w_k = w_k\frac{{}^{(l)}d_k}{d_k} + \left(\boldsymbol{t}_x^{\star} - \sum_{k'\in r} w_{k'}\frac{{}^{(l)}d_{k'}}{d_{k'}}\boldsymbol{x}_{k'}\right)\left(\sum_{k'\in r}{}^{(l)}d_{k'}F(\boldsymbol{z}_{k'}\hat{\boldsymbol{\gamma}})\boldsymbol{z}_{k'}^T\boldsymbol{x}_{k'}\right)^{-1}{}^{(l)}d_k F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}})\boldsymbol{z}_k^T$$

and ${}^{(l)}d_k = 0$ when $k = l$ and ${}^{(l)}d_k = n_r d_k/(n_r - 1)$ otherwise (see Kott, 2006, for a similar proposal).

4. Simulation studies

We consider the Abortion data set formed by four binary variables extracted from the 1986 British Social Attitudes Survey and concerning attitude to abortion. $N = 379$ individuals answered to the following questions after being asked if the law should allow abortion under the circumstances presented under each item: 1) the woman decides on her own that she does not want to keep the baby; 2) the couple agrees that they do not wish to have a child; 3) the woman is not married and does not wish to marry the man; 4) the couple cannot afford any more children. The data is analyzed in Bartholomew et al. (2002) and found to hide a latent continuous variable, interpretable as the attitude to abortion.

We focus on estimation of the total of the first two variables of interest, $t_{y1} = 166$ and $t_{y2} = 225$. At the population level, the latent trait θ_k is estimated for each unit using the two parameter logistic Rasch model in (2). An auxiliary variable x_k is generated as $x_k = 1 + \sum_{m=1}^{4} y_{mk} + \epsilon_k$, with $\epsilon_k \sim N(0, 1)$. Its population total is assumed known. We draw 10,000 simple random samples of dimension $n = 100$. For each sample nonresponse is simulated using a Poisson design with probabilities defined for each unit according to various settings. In this work, for reasons of space, we report only on the following three nonresponse models, that are obtained by varying the model variable and the link function: **Lin**$-y_2$: $p_k = 1/(1.1 + 0.9y_{2k})$; **Rak**$-y_2$: $p_k = 1/\exp(0.2 + 0.5y_{2k})$; **Rak**$-\theta$: $p_k = 1/\exp(0.2 + 0.5\theta_k)$. In all settings, the population average for p_k is around 0.6. Three choices for the model variable are considered to compute a calibration estimator: the variable of interest y_2 (GCAL$-y_2$), the auxiliary variable x (CAL$-x$) and the latent trait θ (GCAL$-\theta$). With these three choices weights are computed using the appropriate link function in the R package `sampling` (Tillé and Matei, 2011) and then applied to estimate t_{y2} and t_{y1}. For each estimator we compute the Monte Carlo Bias, Variance and Mean Squared Error. Variance estimation is conducted using estimators $\hat{V}_{\text{2p}}$, $\hat{V}_{\text{rev}}$ and $\hat{V}_{\text{jack}}$ illustrated in Section 3. For each variance estimator its Monte Carlo expectation is computed, together with the empirical coverage of a 95% confidence interval.

Table 1 reports the results for all simulation settings. We can note that when GCAL$-y_2$ is based on the true response model (**Lin**$-y_2$ and **Rak**$-y_2$) its bias is negligible, but its variance is relatively larger than that of the other estimators. Although correlation between x and y_2 is relatively large (0.74) CAL$-x$, has the opposite behavior, with a large bias and a small variance that places concern on coverage. GCAL$-\theta$ has, as expected, a very good performance in the **Rak**$-\theta$ setting, but is also a very good compromise in the other two cases, when the interest is at the estimation of both y_1 and y_2. As of variance estimators, $\hat{V}_{\text{rev}}$ has a better behavior that $\hat{V}_{\text{2p}}$, but $\hat{V}_{\text{jack}}$ provides the best coverage.

5. Application to the Italian Survey on Household Income and Wealth

The Survey on Household Income and Wealth (SHIW) is conducted by the Italian central bank every two years in order to study the economic behaviors of Italian households by collecting detailed information on their income and wealth. The sample consists of about 8,000 households selected from population registers using a complex two stage (municipality-household) sampling design. Because of the sensitiveness of the issues surveyed, measurement error and unit nonresponse are two major issues (for more details see Neri and Ranalli, 2011). Our goal here is to make inferences

Table 1: *Monte Carlo Bias (B), Variance and MSE for each calibration estimator and simulation setting. Monte Carlo mean and 95% confidence interval coverage for each variance estimator.*

	Model variable	B	VAR	MSE	$\hat{V}_{2p}$	95% cov	$\hat{V}_{rev}$	95% cov	$\hat{V}_{jack}$	95% cov
Lin$-y_2$										
$\hat{t}_{y2}$	GCAL$-y_2$	1.1	397.4	398.5	339.9	92.8%	410.2	95.3%	500.4	97.1%
	CAL$-x$	-23.9	214.1	786.6	173.0	54.7%	213.9	62.5%	265.2	70.9%
	GCAL$-\theta$	-13.3	245.7	421.6	204.8	81.0%	252.3	85.9%	310.4	90.1%
$\hat{t}_{y1}$	GCAL$-y_2$	0.4	308.0	308.2	244.2	91.0%	299.8	94.0%	365.9	96.1%
	CAL$-x$	-8.0	257.8	322.1	198.8	87.9%	247.2	92.0%	304.6	95.0%
	GCAL$-\theta$	3.0	285.6	294.4	240.4	91.3%	296.6	94.4%	361.1	96.5%
Rak$-y_2$										
$\hat{t}_{y2}$	GCAL$-y_2$	0.8	429.4	430.0	351.3	92.2%	429.8	94.9%	521.4	96.9%
	CAL$-x$	-20.3	220.4	633.8	175.4	64.2%	218.2	71.5%	269.7	79.0%
	GCAL$-\theta$	-11.1	257.1	379.7	205.2	83.5%	254.4	88.2%	312.1	91.7%
$\hat{t}_{y1}$	GCAL$-y_2$	0.4	320.6	320.8	248.8	90.8%	307.7	93.4%	374.1	95.7%
	CAL$-x$	-6.1	267.7	304.7	204.5	89.5%	255.5	92.9%	314.0	95.3%
	GCAL$-\theta$	4.3	312.1	330.4	242.0	90.1%	300.8	93.2%	365.2	95.2%
Rak$-\theta$										
$\hat{t}_{y2}$	GCAL$-y_2$	11.9	441.4	583.4	353.9	89.3%	434.7	93.0%	526.4	95.5%
	CAL$-x$	-7.8	211.8	271.8	165.8	86.9%	205.7	91.3%	255.3	94.2%
	GCAL$-\theta$	0.5	249.6	249.9	194.8	90.7%	241.3	93.8%	297.4	96.2%
$\hat{t}_{y1}$	GCAL$-y_2$	-4.4	306.4	325.6	247.2	90.5%	307.0	94.1%	373.0	96.1%
	CAL$-x$	-9.2	268.1	352.7	209.9	87.5%	262.5	91.8%	323.0	94.8%
	GCAL$-\theta$	0.4	305.5	305.7	244.0	91.3%	303.3	94.4%	368.9	96.4%

about the average yearly individual net wealth for the Italian population in 2008. Previous research based on the SHIW data shows that nonresponse is non ignorable and depends on the true wealth. It should be noted that the true wealth is not observed either for respondents because of measurement error. Our approach uses survey responses related to wealth as proxies of the true wealth and builds latent classes to be used as response homogeneity groups. In particular, we use the following variables to build the vector $\boldsymbol{\omega}_k$: the individual observed wealth class (ordinal variable with five levels), the number of total call attempts needed to make the interview (ranging between 1 and 4) and six dummy indicators for the ownership of a secondary dwelling, of bonds, of agricultural and of non-agricultural land, of other non-residential buildings and for the household living in a deluxe dwelling. Using latent class analysis, we classify respondents into five latent classes. Then, the predicted latent classes memberships are used as the instrumental variables $\boldsymbol{z}_k$.

As of calibration variables, we use two sources of auxiliary information. The first one is the National statistical office and consists of the distribution of individuals according to some demographic variables: age (5 classes), gender, education (3 levels), nationality (italian/foreigner), job status (employed/unemployed/inactive), geographical area (north/centre/south). The second one is the Italian Department of the Treasury holding the administrative records of real estate owners and consists of the distribution of individuals according to the value of the owned dwelling (5 classes). Overall, the vector of population totals $\boldsymbol{t}_x$ is made of $q = 18$ components.

Six different approaches to estimation of the mean household wealth are compared: (1) the Hajek estimator (no nonresponse adjustment); (2) a two phase estimator in which response probabilities are estimated via a logistic model that uses covariates known also for nonrespondents (see details in Neri and Ranalli, 2011); (3) classical calibration using the aforementioned 18 calibration variables; generalized calibration using (4) latent classes $\hat{\boldsymbol{\theta}}_k$ as model variables, (5) manifest variables $\boldsymbol{\omega}_k$ as model variables, and (6) the individual observed wealth class (classes of y_k) as model variables. For cases (4), (5) and (6), since $p < q$, we use two step calibration $w_k = d_k F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}}_1)F(\boldsymbol{x}_k\hat{\boldsymbol{\gamma}}_2)$ to obtain final weights in which the first adjustment step $F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}}_1)$ is obtained using the routine proposed in Chang and Kott (2008) and then, since it does not provide calibrated weights, $F(\boldsymbol{x}_k\hat{\boldsymbol{\gamma}}_2)$ is obtained calibrating on $\boldsymbol{t}_x$ using $d_k F(\boldsymbol{z}_k\hat{\boldsymbol{\gamma}}_1)$ as starting weights.

It is worth noting that the true mean of the households' wealth is unknown for this application.

Table 2: *Estimated mean of the households' wealth using different estimators, estimated jackknife standard error and % coefficient of variation, standard deviation of the set of calibrated weights.*

Method	Estimated Mean	Jackknife St Err	%CV	Weights St Dev
(1) Hajek estimator	271692.60	10551.63	3.88	2491.41
(2) Two phase estimator	267147.98	9523.05	3.56	2852.49
(3) Classical calibration	298468.40	9748.53	3.27	2656.21
Two step generalized calibration:				
(4) Model var. – latent cl.	326824.48	11314.23	3.46	2853.14
(5) Model var. – manifest var.	354229.19	41704.42	11.77	8827.81
(6) Model var. – classes of y	337372.15	12118.12	3.59	3276.91

Table 2 reports the results for all the approaches with an estimate of the standard error using jackknife. Nonresponse adjustments via calibration all provide an increase in the estimate. This is reasonable since we expect wealthier household to be less collaborative. The estimated variance of the generalized calibration estimators $(4) - (6)$ is larger than that of classical calibration (3), because of the increased complexity in the first step nonresponse adjustment. As already noted in the simulations, generalized calibration that uses the variables of interest as model variables (cases (5) and (6)) provide more variable estimators than the one that uses latent variables. By comparing (5) with (4), it seems that the reduction in dimensionality performed by latent class analysis allows for a more stable estimator, without loosing too much in terms of information. This is true also when comparing (4) and (6), that have the same number of model variables. Finally, note that (4) shows a way less variable set of weights as opposed to (5) and (6).

References

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman and Hall/CRC.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95:555–571.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*, pages 65–76, New York. Springer.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:133–142.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105:1265–1275.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Boston, Houghton Mifflin.

Neri, A. and Ranalli, M. G. (2011). To misreport or not to report? The measurement of household financial wealth. *Statistics in Transition*, 12-2:281–300.

Oh, H. L. and Scheuren, F. (1983). Weighted adjustment for nonresponse. In Madow, W. G., Nisselson, H., and Olkin, I., editors, *Incomplete Data in Sample Survey*, volume 2, pages 143–184, New York. Academic Press.

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

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonneglible sampling fractions. *Journal of the American Statistical Association*, 94:254–265.

Tillé, Y. and Matei, A. (2011). *The R Package 'sampling'*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.