Quaderni di Statistica
Vol. 13, 2011

Deriving variance estimators in a complex design for a mixture distribution

Romina Gambacorta
Bank of Italy, Economic and Financial Statistics Department
E-mail: romina.gambacorta@bancaditalia.it

Maria Iannario
Department of Theory and Methods for Human and Social Sciences
University of Naples Federico II
E-mail: maria.iannario@unina.it

Richard Valliant
Joint Program in Survey Methodology, University of Maryland
E-mail: rvalliant@survey.umd.edu

Summary: In this paper, we analyse how complex sampling designs affect estimates of a statistical model used for the analysis of ordinal data. If it is not possible to assume that sample information comes from a superpopulation model, where the sampling scheme is ignorable, we introduce a specific method to assess model parameter estimates and their sampling variance. Specifically, we derive for the class of CUB models the variance estimates for complex surveys by means of the Repeated Replication methods. We present examples using data from the 2008 Survey on Households Income and Wealth performed by the Bank of Italy and we apply the Jackknife Repeated Replication method to show differences between design-based and unweighted CUB model parameter estimates and statistical inference.

Keywords: Complex sampling design, CUB models, Repeated Replication methods

1. Introduction

CUB models[1] have been introduced to analyse how subjects' and objects' covariates affect ordinal responses in rating contexts (Piccolo 2003; D'Elia and Piccolo 2005;

[1] The acronym CUB stands for **C**ombination of **U**niform and shifted **B**inomial distributions.

Iannario and Piccolo 2011). Generally, CUB model parameter estimators assume that data are independent and identically distributed and that the probability assigned to each element of the sample is the same, i.e. that data are collected using a simple random sampling (with replacement) structure. Nevertheless, this is rarely the case; in practice, data are collected as part of a complex sample survey that uses stratification, clustering or disproportionate sampling to increase statistical efficiency and reduce sampling costs. In this case the sampling scheme can affect both the accuracy and the precision of sampling estimates.

In fact, many investigations have shown that ignoring the sample design, leads to biased estimates of descriptive statistics and model parameters, generally understating the true value of their sampling variance (Kish 1992; Korn and Graubard 1995; Pfeffermann 1996; Brogan 1998). As a consequence, confidence intervals will be biased and test statistics will tend to overstate the significance of tests. Therefore, design features should be accounted for to produce approximately unbiased and design-consistent estimates of the sampling variance.

It should be noted that if the use of weighting adjustment to correct for unequal selection probabilities, non-response or post-stratification reduces the bias of the estimates, on the other hand it usually inflates sampling variance. This argument is often taken as a justification for a model based approach where attributes of the sample design are not accounted for. In fact, advocates of this method argue that these elements are not relevant when the model is correctly specified and applies universally (superpopulation model). This approach requires strong assumptions, namely the correct specification of the model and the ignorability of the sample design with respect to that model - i.e. the selection probabilities do not depend on the dependent variable, conditionally on the covariates of that model (Sudgen and Smith 1984; Skinner *et al.* 1989). When these assumptions are not satisfied, the model approach may suffer both the consequences of ignoring the design and of model misspecification: these shortcomings may be avoided by the use of survey weights (Pfeffermann and Holmes 1985).

In the context of the design-model approach debate (see Binder and Roberts 2003), this paper aims at discussing the procedures to account for sampling design in estimating and testing CUB model parameters, leaving to the analyst the choice on which approach better suits his/her needs.

The paper is organized as follows: the next section recalls the main features of the CUB model. Section 3 introduces the rationale of the design based approach, and reviews the different strategies that can be used to estimate the sampling variance. In section 4 a fitting model is presented for the analysis of ordinal data with the implementation of weights for inferential issues according to the pseudo-maximum likelihood approach. In section 5 we derive the estimate of the variance-covariance matrix using the Repeated Replication method. In section 6 we describe the data used to check the validity of the implemented model, the Survey of Households Income and Wealth, generated by a complex survey design whereas in section 7 we summarize the main findings. Some concluding remarks end the paper.

2. Statistical models for ordinal data

In this section we synthesize CUB models, generally used to analyse rating scores (ordinal responses) in evaluation contexts (Iannario 2007, 2008).

From a formal point of view, we define a CUB random variable R if and only if its probability mass function, for a given $m > 3$, is defined by:

$$P_r\left(R=r\right)=\pi\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}+(1-\pi)\left(\frac{1}{m}\right),\quad r=1,2,\ldots,m. \quad (1)$$

This model is identifiable if $m > 3$ (Iannario 2010), and the parametric space is the (left open) unit square: $\Omega(\pi,\xi)=\{(\pi,\xi):\ 0<\pi\leq 1;\ 0\leq\xi\leq 1\}$.

In case of ordinal variables generated by the response of an evaluation about a fixed item, m is the number of possible alternatives (that may represent different degrees of agreement to a specific item) while r is the provided score.

According to a CUB model, the uncertainty of respondents -modelled by a proportion of a discrete Uniform random variable- increases with $(1-\pi)$, while agreement with respect to the ‘object’ (*feeling*) -parameterized by a shifted Binomial random variable- increases with $(1-\xi)$.

If we consider a matrix $\boldsymbol{T}(n,v)$ of v covariates and we assume that p and q indicate the number of covariates introduced to explain *uncertainty* and *feeling*, respectively, then a CUB (p,q) model may be introduced. More specifically, we assume that uncertainty and perception/evaluation parameters may be related to p and q covariates, respectively, which are included in $\boldsymbol{T}$, by means of two *systematic components*:

$$\pi_i=\frac{1}{1+e^{-\boldsymbol{y}_i\,\boldsymbol{\beta}}}\,;\qquad \xi_i=\frac{1}{1+e^{-\boldsymbol{x}_i\,\boldsymbol{\gamma}}}\,;\qquad i=1,2,\ldots,n\,, \quad (2)$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ are the observed subjects’ covariates for explaining π_i and ξ_i, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the related parameters.

The matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ are subsets of $\boldsymbol{T}$. Parameters for this class of models can be estimated by using maximum likelihood; specifically, estimators of parameters have been implemented (Piccolo 2006) *via* EM algorithm (McLachlan and Peel 2000).

3. Inference with complex sample survey data

One concern in sample survey data analysis is the derivation of point estimates of population quantities by using a sample estimate. When the sample is selected following a complex design with unequal selection probabilities or adjustments to account for non-response and post-stratification are employed, it is necessary to inflate observations using the appropriate weights in order to obtain unbiased estimation of the finite population parameter (Kish 1992). The same concept can be applied to simple estimators of the totals using the Horvitz-Thompson estimator (Horvitz and Thompson 1952), to linear regression coefficients (Kish and Frankel 1974) and in general to more complex non-linear parameters using the pseudo-maximum likelihood estimation approach, with the implementation of iterative algorithms. In this context, the pseudo-likelihood is an approximation of the likelihood function in the finite population based on the likelihood function of the observed sample units adjusted for their sampling weights (Binder 1981, 1983).

In order to evaluate the statistical reliability and the accuracy of those estimates, to compute confidence intervals and to quantify the sampling error, it is necessary to correctly estimate sampling variance.

Standard survey sampling textbooks provide explicit formulae for unbiased estimators of sampling variance of simple linear estimators (totals, means) in case of a population of known size and for simple designs. In fact, the formulae for the variance computation in complex surveys are complicated by different factors.

First of all, we often deal with non linear estimators of parameters as in case of CUB parameters. Furthermore, usually in household surveys, multistage sampling designs are adopted to account for the fact that units in the population are arranged hierarchically (municipality, households) to reduce data collection costs. When multistage sampling is adopted, stratification, clustering and weighting all affect standard errors of estimates. In particular, stratification is usually adopted to ensure sample adherence to the population distribution with some a priori known characteristics. This technique can reduce the sampling variance when the variables used to stratify the population are such that strata are correlated with survey measures. Clustering reduces survey costs and time needed to conduct the survey; on the other hand, units within a cluster are likely to have similar characteristics (as they share the same socio-economic environment or background) so the inclusion of additional units in the sample coming from the same cluster does not increase proportionally the effective sample size.

The effect of sample design on sampling variance is called *Design EFFect* and is given by the ratio between the actual sampling variance ($var(\hat{\theta})_{complex}$) and the simple random sampling variance ($var(\hat{\theta})_{srs}$) for samples of the same size (Kish 1965):

$$DEFF(\hat{\theta}) = \frac{var(\hat{\theta})_{complex}}{var(\hat{\theta})_{srs}}. \tag{3}$$

Other factors affecting the sampling estimates are respondent behavior (non-response) and unequal probability of selection. They are taken into account by the use of weights, but the inclusion of weights into the estimator adds a further effect on sampling variance. In particular, the effect on sampling variance due to weighting can be approximated by (Kish 1965, 1992):

$$var(\hat{\theta}_w) = var(\hat{\theta}_{uw}) * (1 + CV(w)^2)\,, \tag{4}$$

where $var(\hat{\theta}_w)$ and $var(\hat{\theta}_{uw})$ are, respectively, the sampling variance of a weighted and of an unweighted estimator of the parameter θ and $CV(w)$ is the coefficient of variation of weights. The approximation in (4) is based on the assumption that selection probabilities are unrelated with the variable of interest. Spencer (2000) provides an adjusted version for the case where weights are correlated to the survey variable. In general, the effect of weighting is to reduce sampling bias, while the impact on sampling variance depends on weights variability. Theoretically, weighting can led also to a reduction in sampling variance, as in the case of post-stratification, if the post-strata are homogeneous with respect to the analysed item. Expression (4) is derived in cases where equal weighting would be optimal.

Finally, the effect of complex design and weighting on sampling variance estimators can be calculated by means of the misspecification effect (Skinner, 1989):

$$MEFF(\hat{\theta}) = \frac{var_{complex}(\hat{\theta}_w)}{E_{complex}(var_{srs}(\hat{\theta}_{uw}))} \tag{5}$$

given by the ratio between the sampling variance of the weighted estimator that accounts for the sampling design ($var_{complex}(\hat{\theta}_w)$) and the expected value, computed under the complex sample design, of the estimator derived under the assumption that the sample is selected by simple random sampling (*SRS*) with replacement (and thus unweighted) ($var_{srs}(\hat{\theta}_{uw})$). As this factor is usually different from one, we can conclude that ignoring sampling design and weighting leads to biased estimates of the standard errors. Therefore, confidence intervals and statistical tests will not be correct if the complexity of sample design is not taken into account.

3.1. Methods used to estimate sampling variance: linearization vs replication methods

In the case of simple designs and simple statistics we can apply analytic formulae to calculate an unbiased estimate for the sampling variance (Cochran 1977). But, as mentioned, this is a rare case because in practice complications may arise from both the complexity of the design and the correction added with weighting. In those cases there are no direct analytic methods that can be used to produce unbiased estimators of the standard errors of the estimates and we need to rely on variance estimation methods that use some approximation (see Wolter 2007).

The approaches that may be used to obtain approximate estimates of sampling variance in such complex cases can be divided into:

1. Taylor linearization method;
2. Repeated Replication method.

The Taylor linearization method is used when complex (non-linear) estimators are concerned and is based on the idea of adopting a simplifying assumption with respect to the statistic for which the variance should be estimated. In particular, the non linear estimator is approximated by a linear one, and the standard survey variance estimation methods are used to estimate the precision of the linearized statistic. In practice, the linear approximation is derived by means of a first order Taylor series expansion for the estimator $\hat{\theta}$ of the population parameter θ. Considering a simple case of a (twice continuously differentiable) function f of one real variable x, Taylor's theorem states that:

$$f(x) = f(a) + f'(a)(x-a) + (remainder) \tag{6}$$

where a is the point about which the expansion is done. The linear approximation is obtained by dropping that remainder term.

For the Replication methods, the original sample is used to select G sub-samples, called replicate samples. Then, the estimator $\hat{\theta}_g$ of the population parameter θ is calculated for each sub-sample. Finally, the variance of the estimator is computed as a measure of the variability of these estimates among the sub-samples with respect to the estimate $\hat{\theta}$ computed on the full sample :

$$v\hat{a}r(\hat{\theta}) = c\sum_{g=1}^{G} h_g(\hat{\theta}_{(g)} - \hat{\theta})^2 \tag{7}$$

where c is a constant that depends on the replication method and h_g is a replicate specific constant and depends on the sampling scheme. Replication methods differ in the way replicates are selected from the total sample and in the specification of the constants c and h_g.

Replicate estimates can be obtained from the full sample also by applying replicate weights ($w_i^{(g)}$) that are rescaled versions of the full sample weights (w_i). For some methods of replication, the replicate weights are equal to zero for the units that do not belong to the replicate sample and nonzero for the units in the sub-sample. In other methods, all units in the full sample receive a nonzero weight in each replicate.

There are different methods that can be used to select replicates. The methods that are mainly used to derive variance estimates for complex surveys are the *jackknife repeated replication* (JRR), the *balanced repeated replication* (BRR) and the (rescaled) *bootstrap*.

Both linearization and replication methods provide biased estimates of the variance, even if the bias is usually negligible for large samples (for the replication methods the bias depends also on the number of replications). All the methods are asymptotically equivalent and generally lead to consistent variance estimators. In particular, the latter statement is true for linear or differentiable nonlinear estimators. For the variance of quantiles, the jackknife is not consistent whereas the bootstrap and the BRR are consistent. In terms of statistical performance, the Taylor expansion and JRR show a lower MSE while BRR and Bootstrap perform better in terms of confidence intervals and coverage probability (Wolter 2007).

However, other features rather than precision make it often preferable to choose replication methods (Brick *et al.* 2000). In particular, the main advantages of using the replication approach are:

1. generic applicability and adaptability to all kind of estimators, including nonlinear ones;
2. possibility to account for all kinds of sample designs, estimators and weight adjustments used;
3. sound theoretical basis: justified both in the design and model based approach;
4. applicability to domain estimates and in case of missing data;
5. simplicity: the idea of replicate samples and estimating variability from subsamples is easy to understand;
6. confidentiality safeguard: as all the information on the sample design is already included in replication weights, design variables such as stratum or cluster of the respondents do not need to be released.

On the other hand, linearization methods show many practical drawbacks:

1. the procedure assumes that in the Taylor series expansion of the estimator, terms beyond the linear one make a negligible contribution to the variance of the estimator, which may be not the case in small samples;
2. it may led to complicated formulae (often computationally intractable) in the case of complex statistics and it cannot be applied for nonsmooth functions (like quantiles) for which it is not possible to compute derivatives;
3. it requires the knowledge of all the design variables, which are often not released in publicly disseminated datasets;
4. it is not always possible with linearization methods to account for weighting adjustments due to the fact that weights increase the complexity of the estimator, making linear approximations difficult to compute.

In what follows we will therefore refer to Repeated Replicaton methods to estimate the variance-covariance matrix of CUB parameters.

4. *The use of weights in* CUB *models*

In this section, to analyse the main inferential results which take into account the sampling design, we present an updated version of the algorithm in order to take the unequal selection probabilities into account.

The pseudo-log-likelihood for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})'$ with survey weights $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)'$ can be written as follows:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \log \left\{ \frac{1}{1+e^{-\boldsymbol{y}_i\boldsymbol{\beta}}} \left[\binom{m-1}{r_i-1} \frac{e^{-\boldsymbol{x}_i\boldsymbol{\gamma}(r_i-1)}}{\left(1+e^{-\boldsymbol{x}_i\boldsymbol{\gamma}}\right)^{m-1}} - \frac{1}{m} \right] + \frac{1}{m} \right\}, \quad (8)$$

where r_i is the observed value of the variable R for the i-th subject. For given m, weights $\boldsymbol{w}$, and a fixed small tolerance ϵ (= 10^{-6}, for instance), the EM algorithm may be effectively implemented by integration of standard steps (in this context we summarize all the steps by underlying the innovations of procedures with weight related to 2 and 3 steps of the classical procedure).

To simplify the notation, we implement a CUB models without covariates; thus, $\boldsymbol{\theta} = (\pi, \xi)'$:

0. $$\boldsymbol{\theta}^{(0)} = (\pi^{(0)}, \xi^{(0)})'; \;\; \ell^{(0)} = \ell(\boldsymbol{\theta}^{(0)}).$$

1. $$bb^{(k)} = \binom{m-1}{r_i-1} \left(1-\xi^{(k)}\right)^{r_i-1} \left(\xi^{(k)}\right)^{m-r_i}; \tau^{(k)} = \frac{\pi^{(k)}\, bb^{(k)}}{\pi^{(k)}\, bb^{(k)} \,+\, (1-\pi^{(k)})/m}.$$

2. $$\overline{R}_n^{(k)}(\theta) = \frac{\displaystyle\sum_{i=1}^{n} w_i\, r_i\, \tau(r; \boldsymbol{\theta}^{(k)})}{\displaystyle\sum_{i=1}^{n} w_i\, \tau(r; \boldsymbol{\theta}^{(k)})}.$$

3. $$\pi^{(k+1)} = \frac{\sum_{i=1}^{n} w_i\, \tau(r; \boldsymbol{\theta}^{(k)})}{\sum_{i=1}^{n} w_i}; \;\; \xi^{(k+1)} = \frac{m - \overline{R}_n^{(k)}(\theta)}{m-1}.$$

4. $$\boldsymbol{\theta}^{(k+1)} = (\pi^{(k+1)}, \xi^{(k+1)})'; \qquad \ell^{(k+1)} = \ell(\boldsymbol{\theta}^{(k+1)}).$$

5. $$\begin{cases} \text{if } \mid \ell^{(k+1)} - \ell^{(k)} \mid \geq \epsilon, \; k \to k+1; \;\; \text{go to 1;} \\ \text{if } \mid \ell^{(k+1)} - \ell^{(k)} \mid < \epsilon, \;\; \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k+1)}; \;\; \text{stop.} \end{cases}$$

One of the main problems of the EM algorithm is the choice of a convenient set of starting values $\boldsymbol{\theta}^{(0)}$ for the estimates, since this procedure is generally slower than the second order convergence rates of the Newton Raphson routines. Iannario (2012) suggests alternative starting values based on the qualitative nature of responses which have been proved useful in terms of efficiency for reducing time to convergence.

Previous implementation provides consistent parameter estimates with the support of sampling weights.

5. Design based inference in CUB *models*

The next step in analysing CUB models for complex survey data is to estimate the sampling variance of the parameters in order to assess their significance using the Repeated replication method.

Following the replication method and using the weighted version of the EM algorithm, the estimation of the parameters should be computed both for the entire sample $\hat{\boldsymbol{\theta}}$, using sampling weights (w_i), and for each of the G sub-samples $\hat{\boldsymbol{\theta}}_{(g)}$, using replicate weights $(w_i^{(g)})$ in the EM algorithm. Then, an estimate of the sampling variance for $\hat{\boldsymbol{\theta}}$ is obtained by applying the formula (7).

In the same way, we obtain an estimate of the sampling variance and covariance matrix of the parameters $\hat{V}(\hat{\boldsymbol{\theta}})$, by generalizing the formula in equation (7) to account for covariance elements:

$$\hat{V}(\hat{\boldsymbol{\theta}}) = c\sum_{g=1}^{G} h_g(\hat{\boldsymbol{\theta}}_{(g)} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{(g)} - \hat{\boldsymbol{\theta}})'. \tag{9}$$

where the constants c and h_g are, as mentioned, defined from the chosen replication method and the survey design.

In addition to the computation of standard errors that accounts for the complex design, degrees of freedom in the Student t distribution must be adjusted when testing for parameters' statistical significance. The exact calculation of the degrees of freedom can be in some cases difficult and is a function of the variability of the estimator of the parameters' variance. One rule-of-thumb that is often used is the following (Valliant and Rust 2010): the number of degrees of freedom is calculated as a difference between the number of clusters $(n.clusters)$ and number of strata $(n.strata)$ of the survey design:

$$df_{adj} = n.clusters - n.strata\,.$$

This approximation, derived using the methods of Satterthwaite (1946) is based on the assumption that the variance estimator has approximately a Chi square distribution.

6. The Survey on Household Income and Wealth

The Survey on Household Income and Wealth (SHIW, hereafter) has been conducted by the Bank of Italy since 1965 to collect information on the economic behaviour of Italian households and specifically to measure income and wealth components. The main objective is to estimate how these variables are distributed across Italian households.

The questionnaire has been extended during the years and contains topics about households' demographic structure, occupation and income for each member, households' wealth and debts, consumption, insurance and pension plans. In each wave additional questions on relevant items are added in two round sections, each administered to half of the sample. Over the years some of the topics collected in these additional sections have been about family choices, capital gains, inheritance, financial information, happiness and job satisfaction.

Until 1987 the survey was conducted with time-independent samples of households. Since 1989 part of the sample has comprised households interviewed in previous surveys (panel households) to allow the analysis of the dynamics of the studied phenomena. In particular, in each wave, a random subset of households interviewed for the first time in the previous wave is included in the sample in order to compensate for the loss of panel units due to attrition. The design adopted is known as a split panel survey (Kish 1987).

The sample is drawn in two stages. In the first stage municipalities (Primary Sampling Units, PSUs) are stratified by region and demographic size. Municipalities are then selected with probability proportional to the size of the resident population with the exception of the bigger municipalities (more than 40,000 residents) that are always included in the sample (Self Representing Units, SRUs). In the second stage, households are selected randomly from the municipalities' chosen in the first stage. Data are collected by means of personal interviews conducted by professionally trained interviewers and using computer-assisted devices (Computer-Assisted Personal Interviewing, CAPI). The final sample comprises about 8,000 households. For further details about the sampling design of the SHIW and the questionnaire content see Bank of Italy (2010). Microdata, documentation and publications can be downloaded free of charge from the Bank of Italy's website together with weights adjusting for unequal probability of selection, non-response and post-stratification. Description of the impact of the latter on estimates is provided by Faiella and Gambacorta (2007). In this paper we used data from the 2008 wave of the SHIW. In particular, we refer to the individual global satisfaction which has been collected only for half of the sample. The final sample size was 3,887 households.

6.1. SHIW sampling error calculation model

To allow the analysts to compute sampling variance, since 2008 the SHIW disseminates replication weights. This choice protects confidentiality as it avoids the dissemination of design variables related to geographical information (municipality of residence).

The SHIW sampling error computation model - an approximation of the complex sample design that preserves the features of the actual design while avoiding any source of analytical problems in the estimation of sampling variance (Heeringa *et al.* 2010) - is constructed by assuming a sampling with replacement of ultimate clusters and a paired selection of clusters design (Kish 1965). This assumption avoids sampling error calculation complications associated with a multistage design with a selection without replacement of PSUs. This is a conservative approach as the estimate of the variance will slightly overestimate the true variance. A Sampling Error Stratum (SES) is formed for all municipalities that are selected in the sample with certainty: i.e. all those with more than 40,000 inhabitants (SRUs) or that are home of panel households. Then, SES elements are randomly assigned to form two Sampling Error Computation Units (SECUs). For the remaining municipalities (Non Self-Representing units, NSRUs), sampling error strata are obtained collapsing two similar NSRUs. Each municipality constitutes a SECU.

In 2008 this model leds to 319 sampling error strata each containing two SECUs each. The Jackknife Repeated Replication (JRR) method is then employed to construct replicates. Another option could have been to use the Balanced Repeated Replication method since pairs of SECUs were formed.

In the case of the JRR method each replicate is formed by deleting one PSU from the sample in turn. In particular, as in the sampling error design each SES contains only two SECUs one replicate is created for each stratum by deleting its first computation unit. The paired replicates that could be obtained by deleting the second SECU are not considered as they add no precision to the final variance estimate (Wolter, 2007). The version of the Jackknife used in SHIW is also known as JK2. In terms of replicate weights this means assigning to all the elements of the deleted SECU a weight equal to zero and adjusting the weights of the elements in the remaining SECU in the strata to compensate for the loss in weights of the cancelled units. The weights of all the other elements remain unchanged (Faiella 2008).

Using replicate weights it is possible to calculate for each replicate the relevant estimator $\hat{\theta}_g$ of the population parameter θ and to calculate its variance by applying the formula in equation (5) where $c = 1$ and $h_g = 1$.

7. Design based estimation in the SHIW by means of CUB Models

Results based on an unweighted CUB model without covariates fitted for one of the ordinal variable present in the questionnaire (*happiness* expressed on a Likert scale with $m = 10$ ordered categories: from 1 *low happiness* to 10 *high happiness*), and the implementation the same model with weights are presented in Table 1. The first two rows report results for the model estimated without accounting for weights and design features while in the following rows weights are included in the estimates by applying the modified version of the EM algorithm reported in section 4. Standard errors are calculated using the JRR method as described in section 5.

Considering the first model, the estimated value of $\hat{\pi}$ denotes a *low* level of uncertainty $(1 - \hat{\pi} = 0.126)$ in the answers. Respondents give accurate rating to the investigated question. The feeling toward *happiness* is on average moderately high $(1 - \hat{\xi} = 0.670)$. The results obtained using a design based approach show slightly higher expected values for *happiness* and lower respondents indecision. Standard errors, and therefore confidence intervals, are substantially larger in the latter case.

Table 2 summarizes the main findings in terms of estimated bias and effect of the design that can be deduced comparing the models in Table 1. We assume that the weighted estimates are approximately unbiased

In this model, neglecting the use of weights leads little differences in term of parameters estimates; in particular, we observe a small overestimate of a person uncertainty about *happiness* (as the estimated bias on $\hat{\pi}$ is positive) and a minor underestimate of respondent's feeling towards *happiness* (the estimated bias on $\hat{\xi}$ is negative). On the

Table 1. Estimation of CUB *model for* Happiness

Model Parameter	Parameter Estimate	Standard Error	t-Statistics (*df*)	95% Confidence interval
CUB $_{unweighted,srs}$				
$\hat{\pi}$	0.8735	0.0109	80.128	$(0.8520, 0.8948)$
$\hat{\xi}$	0.3296	0.0030	109.867	$(0.3237, 0.3355)$
CUB $_{weighted,JRR}$				
$\hat{\pi}$	0.8847	0.0159	55.642 (*319*)	$(0.8534, 0.9160)$
$\hat{\xi}$	0.3240	0.0043	75.349 (*319*)	$(0.3155, 0.3325)$

other hand, accounting for design features does affect standard errors as the misspecification effect indicates a substantial increase in the estimates of the standard errors of the parameters, although their significance is not altered.

Table 2. Estimated BIAS and MEFF of CUB *model parameters for* Happiness

	$\hat{\pi}$	$\hat{\xi}$
$BIAS$	0.0113	−0.0056
$MEFF_{JRR}$	2.1279	2.0544

Finally, Figure 1 reports smoothed density estimates of the Jackknife (JRR) distribution for the analysis of standard errors concerning the CUB model with weights. The distributions are based on 319 replicate estimates.

7.1. Design based estimation by means of CUB *models with subjective covariates*

For the implementation of a more complete model, several covariates related to *happiness* are significant. Among them, we can observe the significance of *gender*, *health*, *education* and marital status (*married*) for explaining the feeling parameter, and the perceived family economic condition (*familycond*) for explaining the uncertainty. In particular, *gender* is a dummy equal to one for females, *health* is the individuals' perceived status of health from 1 (excellent) to 5 (very poor), *education* represents the education degree of the individual on a scale from 1 to 5 , *married* is a dummy for married people while *familycond* represent the perception of the individual about household's income as sufficient to allow the family to make ends meet, from 1 (with great difficulty) to 5 (very easily).

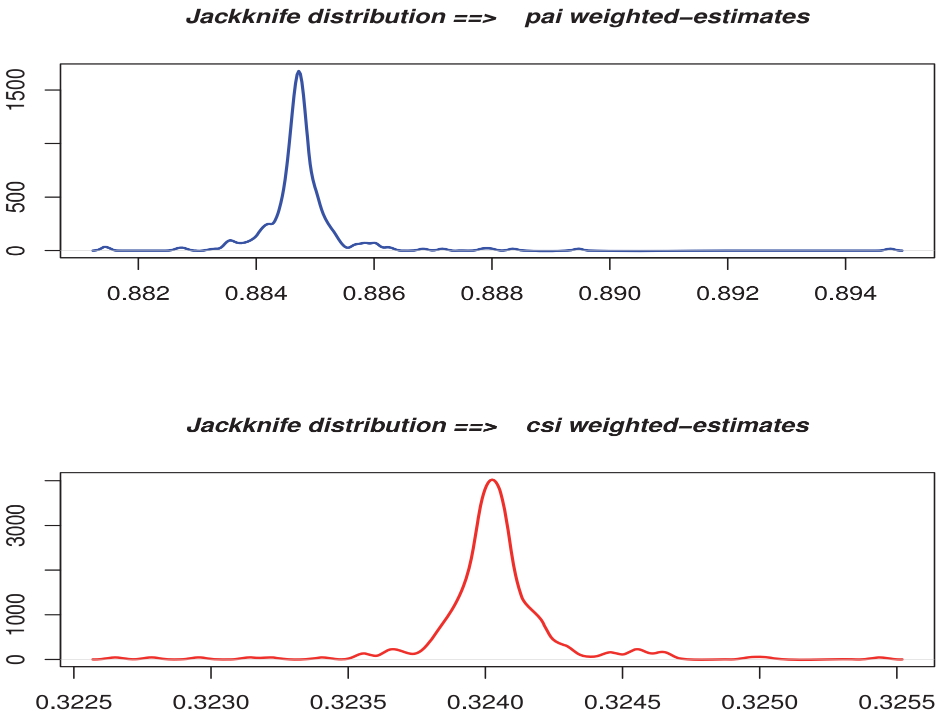

Figure 1. Jackknife Repeated Replication (JRR) for the analysis of standard error. Top panel is density for $\hat{\pi}$; bottom panel is density for $\hat{\xi}$.

In the following we compare the results obtained with CUB unweighted and CUB weighted models, respectively.

Table 3. Estimation of CUB *model for* Happiness

Model Parameter		Parameter Estimate	Standard Error	t-Statistics (df)	95% Confidence interval
CUB $_{unweighted,srs}$					
	$\hat{\beta}_0$	-0.0220	0.3076	$-\ 0.0715$	$(-0.6249,\ 0.5809)$
familycond	$\hat{\beta}_1$	1.1909	0.1566	7.6022	$(\ 0.8838,\ 1.4979)$
	$\hat{\gamma}_0$	-0.7258	0.0674	-10.7720	$(-0.8579, -0.5938)$
gender	$\hat{\gamma}_1$	0.1327	0.0290	4.5772	$(\ 0.0759,\ 0.1896)$
health	$\hat{\gamma}_2$	0.2621	0.0160	16.6520	$(\ 0.2312,\ 0.2930)$
education	$\hat{\gamma}_3$	-0.1145	0.0129	$-\ 8.8760$	$(-0.1398, -0.0892)$
married	$\hat{\gamma}_4$	-0.3976	0.0292	-13.6060	$(-0.4548, -0.3403)$
CUB $_{weighted,JRR}$				($df = 319$)	
	$\hat{\beta}_0$	-0.5599	0.4485	-1.2482	$(-1.4421,\ 0.3223)$
familycond	$\hat{\beta}_1$	1.5168	0.2145	7.0719	$(\ 1.0949,\ 1.9387)$
	$\hat{\gamma}_0$	-0.7659	0.0977	-7.8413	$(-0.9581, -0.5737)$
gender	$\hat{\gamma}_1$	0.1478	0.0342	4.3174	$(\ 0.0805,\ 0.2151)$
health	$\hat{\gamma}_2$	0.2664	0.0208	12.804	$(\ 0.2255,\ 0.3073)$
education	$\hat{\gamma}_3$	-0.1166	0.0206	-5.6534	$(-0.1571, -0.0761)$
married	$\hat{\gamma}_4$	-0.3655	0.0385	-9.4844	$(-0.4412, -0.2898)$

The sign of the parameters are the same in both models: people that have problems to make ends meet answer with higher uncertainty; men, healthier and more educated people report higher level of happiness. Also, marriage exerts a positive effect on reported feeling towards happiness.

Table 4 summarizes the main findings in terms of estimated bias and effect of the design that can be deduced comparing the two models in Table 3. Neglecting the design produces a consistent bias in estimating the effect of perceived family economic conditions on uncertainty. The bias is moderate for other variables. MEFF are quite large for all the variables, when using the jackknife repeated replication method, indicating in general that, even when the increase in standard errors does not affect parameters statistical significance, confidence intervals are larger, and thus parameters' precision is lower.

Finally, in Table 5 we report the Variance-Covariance matrix obtained by means of *JRR* method, while Figure 2 reports the Jackknife (JRR) distribution for the analysis of standard errors of the model with weights and covariates.

Table 4. Estimated BIAS and MEFF of CUB *model parameters for* Happiness

Parameter	BIAS	$MEFF_{JRR}$
$\hat{\beta}_0$	-0.5379	2.1263
$\hat{\beta}_1$	0.3259	1.8746
$\hat{\gamma}_0$	-0.0400	2.1013
$\hat{\gamma}_1$	0.0151	1.3937
$\hat{\gamma}_2$	0.0043	1.7477
$\hat{\gamma}_3$	-0.0021	2.5545
$\hat{\gamma}_4$	0.0320	1.7398

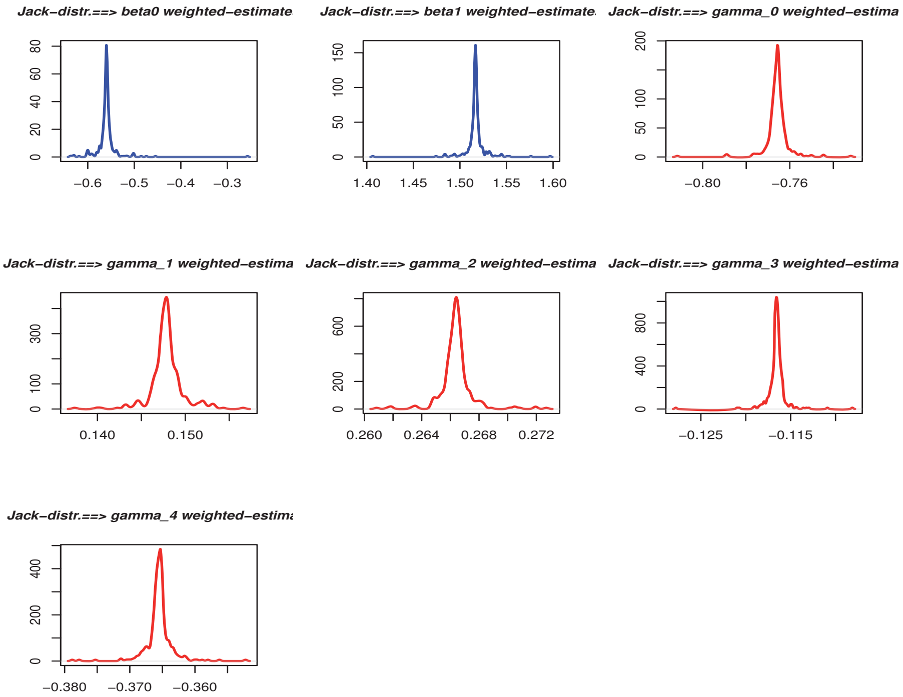

Figure 2. Jackknife Repeated Replication (JRR) for the analysis of standard errors

8. Conclusions

In this paper we propose design-based parameter estimates and inference for CUB models. Variance formulas are derived by using repeated replication methods.

Table 5. Variance-Covariance matrix *of* CUB *model with* JRR *method*

	β_0	β_1	γ_0	γ_1	γ_2	γ_3	γ_4
β_0	0.2012						
β_1	−0.0856	0.0460					
γ_0	0.0081	−0.0020	0.0095				
γ_1	−0.0022	0.0005	−0.0008	0.0012			
γ_2	−0.0012	0.0006	−0.0014	0.0000	0.0004		
γ_3	−0.0006	−0.0001	−0.0017	0.0000	0.0002	0.0004	0.0015
γ_4	−0.0027	−0.0008	0.0000	−0.0013	0.0005	0.0000	0.0000

Results, obtained on data from the Survey on Households Income and Wealth, show that neglecting design features led to a serious underestimate of the sampling variance of the parameters' estimates and to potentially biased estimates of the parameters. Nevertheless, in our examples, the statistical significance of CUB parameters is not affected by the selected approach. The latter result is probably due to the sizeable dimension of the sample and to the large number of SECUs in the SHIW sampling error design. However, in smaller surveys, or even in analyses of subgroups in SHIW (like regional estimates, analysis referred to income classes), there could be cases where unweighted estimates could produce false significance.

We therefore suggest to carefully choose the estimation approach (design *vs* model based), and seriously to take into consideration the possible consequences on results.

Acknowledgements: This work has been partly supported by the MIUR grant for PRIN2008 project of the Research Unit at University of Naples Federico II (PUC E61J10000020001) and of FARO 2011 project of Polo SUS. During the preparation of this paper, M. Iannario benefits of Fulbright scholarship at Department of Statistics and Actuarial Science, University of Iowa, USA. The views expressed are those of the authors and do not necessarily reflect those of the Bank of Italy.

References

Bank of Italy (2010), Italian Household Budgets in 2008, *Supplements to the Statistical Bulletin - Sample Surveys*, Rome, 8.

Binder, D. A. (1981), On the Variances of Asymptotically Normal Estimators from Complex Surveys, *Survey Methodology*, 7, 157–170.

Binder, D. A. (1983), On the Variances of Asymptotically Normal Estimators from

Complex Surveys, *International Statistical Review*, 51, 279–292.

Binder, D., Roberts, G. (2003), Design-based and Model-based Methods for Estimating Model Parameters. In R.L. Chambers and C.J. Skinner (Eds.), *Analysis of survey data*, John Wiley & Sons, 29–33.

Brick, J.M., Morganstein, D., Valliant, R. (2000), *Analysis of Complex Sample Data Using Replication*, Technical Report, Westat, Rockville, MD. http://www.westat. com/ Westat/pdf/wesvar/ACS-Replication.pdf

Brogan, D.J. (1998), Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data, in: P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*, John Wiley and Sons, New York.

Cochran, W. G. (1977), *Sampling techniques* (3rd ed.), John Wiley & Sons, New York.

D'Elia, A., Piccolo, D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.

Faiella, I. (2008), Accounting for sampling design in the SHIW, *Temi di discussione*, 662, Bank of Italy.

Faiella, I., Gambacorta R. (2007), The weighting process in the SHIW, *Temi di Discussione*, n.636, The Bank of Italy, Roma.

Fellegi, I. P. (1980), Approximate tests of independence and goodness of fit based on stratified multistage samples, *Journal of the American Statistical Association*, 75, 261–268.

Heeringa, S.G., West, B.T., Berglund, P.A. (2010), *Applied Survey Data Analysis*, Chapman Hall/CRC Press, Boca Raton, FL.

Horvitz, D.G., Thompson, D.J. (1952), A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, 47, 663–685.

Iannario, M. (2007), A statistical approach for modelling Urban Audit Perception Surveys, *Quaderni di Statistica*, 9, 149–172.

Iannario, M. (2008), A class of models for ordinal variables with covariates effects, *Quaderni di Statistica*, 10, 53–72.

Iannario, M. (2009), Fitting measures for ordinal data models, *Quaderni di Statistica*, 11, 39–72.

Iannario, M. (2010), On the identifiability of a mixture model for ordinal data, *Metron*, LXVIII, 87–94.

Iannario, M. (2012), Preliminary estimators for a mixture model for ordinal data, *Advances in Data Analysis and Classification*, 6, forthcoming.

Iannario, M., Piccolo, D. (2009), A program in R for CUB models inference, available at www.dipstat.unina.it, Version 2.0.

Iannario M., Piccolo, D. (2012), CUB Models: Statistical Methods and Empirical Evidence, in: R.S. Kenett and S. Salini (Eds.) *Modern Analysis of Customer Survey*, J. Wiley & Sons, New York, 232–254.

Kish, L. (1965), *Survey Sampling*, J. Wiley & Sons, New York.

Kish, L. (1987), *Statistical design for research*, J. Wiley & Sons, New York.

Kish, L. (1992), Weighting for Unequal Pi, *Journal of Official Statistics*, 8, 183–200.

Kish, L., Frankel, M.R. (1974), Inference from Complex Samples, *Journal of the Royal Statistical Society, Series B*, 36, 1–37.

Korn, E.L., Graubard, B.I. (1990), Simultaneous testing of regression coefficients with complex sample survey data: Use of Bonferroni t statistics, *The American Statistician*, 44, 270–276.

Korn, E.L., Graubard, B.I. (1995), Examples of Differing Weighted and Unweighted Estimates from a Sample Survey, *The American Statistician*, 49, 291–295.

Leti, G. (1979), *Distanze e indici statistici*, La Goliardica Editrice, Roma.

McLachlan G., and Peel G.J. (2000), *Finite Mixture Models*, J. Wiley & Sons, New York.

Pfeffermann, D. (1996), The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239–261.

Pfeffermann, D., Holmes, D.J. (1985), Robustness considerations in the choice of method of inference for regression analysis of survey data, *Journal of the Royal Statistical Society, Series A*, 148, 268–278.

Piccolo, D. (2003), On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.

Piccolo, D. (2006), Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33–78.

Satterthwaite, F.E. (1946), An Approximate Distribution of Estimates of Variance Components, *Biometrics Bulletin*, 2, 110–114.

Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer, New York.

Skinner, C.J. (1989), Introduction to Part A. In Skinner, C., Holt, D. and Smith, T. (Eds.), *Analysis of Complex Surveys*, J. Wiley & Sons, Chichester.

Skinner, C.J., Holt, D., Smith, T.M.F. (1989), *Analysis of Complex Surveys*, J. Wiley & Sons, New York.

Spencer, B.D. (2000), An Approximate Design Effect for Unequal Weighting When Measurement May Correlate With Selection Probabilities, *Survey Methodology*, 26, 137–138.

Sugden, R.A., Smith, T.M.F. (1984), Ignorable and informative designs in survey sampling inference, *Biometrika*, 71, 495–506.

Valliant, R., Rust, K. (2010), Degrees of Freedom Approximations and Rules of Thumb, *Journal of Official Statistics*, 26, 585–602.

Wolter, K.M. (2007), *Introduction to Variance Estimation*, 2nd ed., Springer-Verlag, New York.