Calibration and Regression Estimation in Dual Frame Surveys

M. Giovanna Ranalli [1] & Antonio Arcos [2] & Maria del Mar Rueda [3] & David Molina[4]

[1] *Department of Political Sciences, University of Perugia,*
giovanna.ranalli@stat.unipg.it
[2] *Department of Statistics and Operational Research, Universidad de Granada,*
arcos@ugr.es
[3] *Department of Statistics and Operational Research, Universidad de Granada,*
mrueda@ugr.es
[4] *Department of Statistics and Operational Research, Universidad de Granada,*
dmolinam@ugr.es

Abstract. Recently, multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. We will discuss recent developments in the application of the calibration paradigm to the estimation of the total of a variable of interest in dual frame surveys as a general tool to include auxiliary information, also available at different levels. When the variable of interest is binary or, more generally, categorical, extension of model calibration and generalized regression estimation in this context is also shown by means of multinomial assisting models. A dedicated R-package – Frames2 – has also been developed to obtain estimates from dual frame surveys and incorporating auxiliary information.

Key-words. Auxiliary information, Model calibration, Multinomial model, Raking ratio, R-package.

1 Introduction

A main aim of survey statisticians is to obtain more accurate estimates, without increasing survey costs. Two popular tools to achieve this goal are (*i*) the use of more than one population frame to select independent samples and (*ii*) the use of auxiliary information either at the design or at the estimation stage. The use of more than one list of population units is important because a common practical problem in conducting sample surveys is that frames may be incomplete or out of date, so that resulting estimates may be seriously biased. Multiple frame surveys are useful when no single frame covers the whole target population but the union of several available frames does, or when information about a subgroup of particular interest comes only from an incomplete frame. They also have other advantages. In fact, Hartley (1962) introduces dual frame surveys as a cost-saving device, showing that they can often achieve the same precision as a single-frame survey at a much reduced cost. Kalton and Anderson (1986) suggest using two frames for sampling rare populations where even greater efficiencies can be obtained. Several estimators of the population total and mean have been proposed in the literature in dual frame surveys, usually classified, according to the level of frame information needed, as *dual-frame* and *single-frame* estimators.

On the other hand, the growing availability of information coming from census data, administrative registers and previous surveys provide a wide range of variables, concerning the population of interest, that are eligible to be employed as auxiliary information to increase efficiency in the estimation procedure. In this scenario, a very relevant example, especially for official statistics, is given by *calibration estimation* that adjusts basic design weights to account for auxiliary information and meet benchmark constraints on auxiliary variables population statistics (Deville and Särndal, 1992). Särndal (2007) provides an overview on developments in calibration estimation. In this work, we will discuss how calibration estimation can be used to handle estimation from two frame surveys and how different types of auxiliary information can be easily integrated in the calibration process as benchmark constraints (Section 2). Moreover, when the variable of interest is binary or, more generally, categorical, extension of model calibration and generalized regression estimation in the context of dual frame surveys is also illustrated by means of multinomial assisting models (Section 3). Functionalities of a dedicated `R`-package – `Frames2` – which has been developed to obtain estimates from dual frame surveys will be briefly illustrated together with some concluding remarks (Section 4).

2 Calibration for dual frame surveys

Consider a finite set of N population units identified by the integers, $\mathcal{U} = \{1, \dots, k, \dots, N\}$, and let A and B be two sampling-frames, both can be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$. Because the population units in the overlap domain ab can be sampled in either survey or both surveys, it is convenient to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, to denote the domain in the overlapping area coming from frame B. Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in $\mathcal{U}$, $\mathcal{A}$, $\mathcal{B}$, a, b, ab, ba, respectively. It follows that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ba}$ and $N = N_a + N_b + N_{ab} = N_a + N_b + N_{ba}$.

Let y be a variable of interest in the population and y_k its value on unit k, for $k = 1, \dots, N$. The entire set of population y values is our finite population $\mathcal{F}$. The objective is to estimate the finite population total $Y = \sum_{k=1}^{N} y_k$ of y, that can be written as

$$Y = Y_a + \eta Y_{ab} + (1 - \eta) Y_{ba} + Y_b, \tag{1}$$

where $0 \leq \eta \leq 1$, and $Y_a = \sum_{k \in a} y_k$, $Y_{ab} = \sum_{k \in ab} y_k$, $Y_{ba} = \sum_{k \in ba} y_k$ and $Y_b = \sum_{k \in b} y_k$. Two probability samples s_A and s_B are drawn independently from frame A and frame B of sizes n_A and n_B, respectively. Each design induces first-order inclusion probabilities π_{Ak} and π_{Bk}, respectively, and sampling weights $d_{Ak} = 1/\pi_{Ak}$ and $d_{Bk} = 1/\pi_{Bk}$. Units in s_A can be divided as $s_A = s_a \cup s_{ab}$, where $s_a = s_A \cap a$ and $s_{ab} = s_A \cap (ab)$. Similarly, $s_B = s_b \cup s_{ba}$, where $s_b = s_B \cap b$ and $s_{ba} = s_B \cap (ba)$. Note that s_{ab} and s_{ba} are both from the same domain ab, but s_{ab} is part of the frame A sample and s_{ba} is part of the frame B sample. In this way, we have a sort of "poststratified" sample $s = s_a \cup s_{ab} \cup s_{ba} \cup s_b$ with "poststratum" sample sizes n_a, n_{ab}, n_{ba} and n_b. Note that $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$ (see Rao and Wu, 2010).

The Hartley (1962) estimator of Y is given by

$$\hat{Y}_H(\eta) = \hat{Y}_a + \eta \hat{Y}_{ab} + (1 - \eta)\hat{Y}_{ba} + \hat{Y}_b, \tag{2}$$

where $\hat{Y}_a = \sum_{k \in s_a} d_{Ak} y_k$ is the expansion estimator for the total of domain a and similarly for the other domains. If we let

$$d_k^\circ = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ \eta d_{Ak} & \text{if } k \in s_{ab} \\ (1-\eta) d_{Bk} & \text{if } k \in s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases},$$

then $\hat{Y}_H(\eta) = \sum_{k \in s} d_k^\circ y_k$. In the following, we will drop η for ease of notation. Choice of a value for η has attracted much attention in literature but will not be discussed here (see Lohr, 2009, for a review).

Now, calibration estimation, as discussed in one frame surveys by Deville and Särndal (1992), can be used to handle estimation from two frame surveys and different types of auxiliary information can be easily integrated in the calibration process as benchmark constraints. Let $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})$ be the value taken on unit k by a vector of auxiliary variables $\boldsymbol{x}$ of which we assume to know the population total $\boldsymbol{t}_x = \sum_{k=1}^N \boldsymbol{x}_k$. This vector of totals may pertain only $\mathcal{A}$, only $\mathcal{B}$, the entire population $\mathcal{U}$, or a combination of the three. We will look at a general formulation of the problem. Relevant examples of auxiliary vectors $\boldsymbol{x}$ can be found in Ranalli et al. (2014), together with the asymptotic properties of the resulting estimator.

Using the calibration paradigm, we wish to modify, as little as possible, basic Hartley weights d_k° to obtain new weights w_k°, for $k \in s$ to account for auxiliary information and derive a more accurate estimation of the total Y. A general dual-frame calibration estimator can be defined as $\hat{Y}_{\text{CAL}} = \sum_{k \in s} w_k^\circ y_k$ where w_k° is such that

$$\min \sum_{k \in s} G(w_k^\circ, d_k^\circ) \qquad \text{s.t.} \qquad \sum_{k \in s} w_k^\circ \boldsymbol{x}_k = \boldsymbol{t}_x, \tag{3}$$

where $G(w, d)$ is a distance measure satisfying the usual conditions required in the calibration paradigm (see e.g. Deville and Särndal, 1992, Section 2). Given the set of constraints, different calibration estimators are obtained by using different distance measures. In many instances, numerical methods are required to solve the the minimization problem in (3). Carefully defining the elements of the auxiliary variable vector $\boldsymbol{x}$ allows for the inclusion of, for example, information on the frame sizes or of the overlap domain size (see Ranalli et al., 2014, for a whole set of examples).

The calibration process induces a different final value for the weights which depends on both the distance measure $G(\cdot, \cdot)$ used and the benchmark constraints applied. On the other hand, given a value for η, the final set of weights does not depend on the values of the variables of interest and can be, therefore, used for all variables of interest. When a value for η is to be computed from the sample data, then it is essential to consider proposals based on estimators of N_a, N_b and N_{ab} as the one in, e.g., Skinner and Rao (1996) so that it is the same for all variables of interest.

When inclusion probabilities in domain ab are known for both frames, and not just for the frame from which the unit was selected, *single-frame* methods can be used that combine the observations into a single dataset and adjust the weights in the intersection domain for multiplicity. In particular, observations from frame A and frame B are combined and the two samples drawn independently from A and B are considered as a single stratified sample over the three domains a, b and ab. To adjust for multiplicity, the weights are defined as follows for all units in frame A and in frame B,

$$d_k^\star = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ (1/d_{Ak} + 1/d_{Bk})^{-1} & \text{if } k \in s_{ab} \cup s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases} .$$

Note that units in the overlap domain, which are expected to be selected a number of times given by $1/d_{Ak}+1/d_{Bk}$ have equal weights in frame A and in frame B. The estimator proposed by Kalton and Anderson (1986) is essentially an expansion estimator for which $\hat{Y}_S = \sum_{k\in s} d_k^\star y_k$.

The calibration estimator in this single-frame approach is given by $\hat{Y}_{\text{CAL}}^{\text{S}} = \sum_{k\in s} w_k^\star y_k$ where weights $w_k^\star$ are such that

$$\min \sum_{k\in s} G(w_k^\star, d_k^\star) \quad \text{s.t.} \quad \sum_{k\in s} w_k^\star \boldsymbol{x}_k = \boldsymbol{t}_x.$$

Note that the only difference with equation (3) is the starting basic design weight. Note also that calibration can handle the case in which $(1/d_{Ak} + 1/d_{Bk}) \geq 1$ for some units k and, therefore, the basic weights are smaller than 1.

In Ranalli et al. (2014) a discussion of the asymptotic properties of this estimator is also provided. Variance estimation is also considered for both the dual and the single frame approach, by means of an analytic variance estimator based on the linearization technique and of Jackknife. In addition, some estimators proposed in the literature are shown to belong the class of calibration estimators, like the estimator in Skinner (1991) which, when N_A and N_B are known, adjusts the single-frame estimator $\hat{Y}_S$ using raking ratio estimation. In this case the single frame calibration estimator provides a simple tool to extend such Raking Ratio estimator to general sampling designs by simply plugging in different basic design weights $d_k^\star$, and to more composite auxiliary information settings. In addition, it can also provide an alternative estimate for the overlap domain size N_{ab}.

3 Model calibration and regression estimation for the case of a categorical variable of interest

Very often in surveys the interest is in the estimation of class frequencies of a categorical response variable, like when data is collected from respondents who provide a single choice from a list of alternatives. We code these alternatives $1, 2, \ldots, m$ and the objective is to estimate the frequency distribution of such y variable in the population U. To this end, we define a set of indicators z_i $(i = 1, \ldots, m)$ such that for each unit $k \in U$ $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. Our problem thus, is to estimate the proportions $P_i = N^{-1} \sum_{k \in U} z_{ki}$ $i = 1, 2, \ldots, m$. Note that, as before, $P_i = N^{-1}(Z_{ai} + \eta Z_{abi} + (1 - \eta) Z_{bai} + Z_{bi})$, where $Z_{ai} = \sum_{k \in a} z_{ki}$, $Z_{abi} = \sum_{k \in ab} z_{ki}$, $Z_{ba} = \sum_{k \in ba} z_{ki}$ and $Z_b = \sum_{k \in b} z_{ki}$.

It is well known that the efficiency of the calibration procedures relies on how well a linear model describes the relationship between the variable(s) of interest and the auxiliary ones. Therefore, it may be inefficient when the underlying relationship is indeed non linear or when the variables of interest are not continuous (Wu and Sitter, 2001). In the case of a categorical variable and assuming we also know the value of the vector of auxiliary variables $\boldsymbol{x}_k$ for $k \in \mathcal{U}$, it seems more sensible to consider that the population under study $\mathbf{y} = (y_1, ..., y_N)^T$ is the determination of a set of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)^T$ s.t.

$$\mu_{ki} = P(Y_k = i|\boldsymbol{x}_k) = E(Z_{ki}|\boldsymbol{x}_k) = \frac{\exp(\boldsymbol{x}_k \boldsymbol{\beta}_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_{kr} \boldsymbol{\beta}_r)}, \quad i = 1, .., m,$$

that is, to use the multinomial logistic model to relate the variables y and $\boldsymbol{x}$. Let $\boldsymbol{\beta}$ be the parameter vector $(\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_m^T)^T$, the first step is to estimate it using the information from the s.

In the *single-frame* framework, we propose to estimate $\boldsymbol{\beta}$ by maximizing the π-weighted likelihood (see e.g. Wu and Sitter, 2001) given by $L(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k \in s} d_k^\star \ln \mu_{ki}$. This usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work rather well. Given the estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we consider the following auxiliary variable

$$p_{ki} = \hat{\mu}_{ki} = \frac{\exp(\boldsymbol{x}_k \hat{\boldsymbol{\beta}}_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_{kr} \hat{\boldsymbol{\beta}}_r)}. \tag{4}$$

Since the vector $\boldsymbol{x}_k$ is known for all units of the population $\mathcal{U}$, the values p_{ki} are available $\forall k \in \mathcal{U}$ and we propose to use the values p_{ki} to obtain a new estimator for P_i,

$$\hat{P}_{MLRSi} = \frac{1}{N}\left(\sum_{k\in\mathcal{U}} p_{ki} + \sum_{k\in s} d^{\star}_k (z_{ki} - p_{ki})\right). \tag{5}$$

We observe that this estimator takes the same model-assisted form of a GREG estimator, but considering more complex models, as those considered, e.g., in Wu and Sitter (2001). Nonetheless, here it is adjusted to account for the dual frame sampling setting.

In a calibration setting, we propose to use an extension of the model calibration estimator considered in Wu and Sitter (2001), that also allows to include frame membership information. In particular, $\hat{P}_{MLcalSFi} = \frac{1}{N}\sum_{k\in s} w^{\star}_k z_{ki}$, where $w^{\star}_k$ minimizes $\sum_{k\in s} G(w^{\star}_k, d^{\star}_k)$ subject to:

$$\sum_{k\in s} w^{\star}_k \boldsymbol{r}_{ki} = \sum_{k\in U} \boldsymbol{r}_{ki}$$

where the elements of $\boldsymbol{r}_{ki}$ change according to the available auxiliary information. For example, if N_A, N_B N_{ab} are known, then $\boldsymbol{r}_{ki} = (\delta_k(a), \delta_k(ab) + \delta_k(ba), \delta_k(b), p_{ki})$, while, if only N_A and N_B are known, then $\boldsymbol{r}_{ki} = (\delta_k(a)+\delta_k(ab)+\delta_k(ba), \delta_k(b)+\delta_k(ba)+\delta_k(ab), p_{ki})$. Here, $\delta_k(a), \delta_k(ab), \delta_k(ba)$ and $\delta_k(b)$ are the indicator variables for domains a, ab, ba and b, respectively.

In the dual frame approach, parameter estimates and, then, probabilities μ_{ki} are obtained separately for each frame. That is, for each $k \in \mathcal{A}$, using data of sample s_A one can estimate μ_{ki} by

$$p^A_{ki} = \frac{\exp(\boldsymbol{x}_k \hat{\boldsymbol{\beta}}^A_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k \hat{\boldsymbol{\beta}}^A_r)},$$

where we estimate $\boldsymbol{\beta}^A$ by maximizing $L(\boldsymbol{\beta}^A) = \sum_{i=1,\ldots,m} \sum_{k\in s_A} d_{Ak} \ln \mu_{ki}$. Similarly we obtain p^B_{ki} for $k \in \mathcal{B}$. Then several alternative regression and model calibration estimators can be considered according to the different combinations of such estimates and frame level information. Note that the proposed estimators have the additional advantage that the estimates of proportions for each category add to 1. This is an important issue for statistical agencies because it grants internal consistency of estimates. For further details see Rueda et al. (2014).

4 Concluding remarks

The contribution deals with recent developments in estimation from dual frame surveys that try to incorporate available auxiliary information into the estimation procedure. The natural environment in a survey setting to achieve this is through the calibration framework. The latter allows to handle auxiliary information at different frame levels and for both categorical and continuous variables. In this sense, post-stratification, raking ratio and regression estimation can all be seen as particular cases of calibration.

The calibration framework is also very flexible because it allows to account using model calibration for the fact that the variable of interest can be categorical and, therefore, more suitable modeling could be performed. We have briefly reviewed the possibility of handling it using multinomial logistic models.

Calibration is also a well known tool to handle non-sampling errors, especially unit non-response. Extension of the calibration framework to handle estimation from dual frame surveys opens the possibility to use auxiliary information to reduce non-response bias also in this setting. This is particularly relevant when response propensity is different in different frames and calibration can allow for more flexible weight adjustments.

An R package `Frames2` is being developed for point and interval estimation in dual frame context. Functions composing the package implement the most important estimators in the literature for population totals and means under the dual-frame approach and also under the single-frame approach. The calibration approach is also included to incorporate auxiliary information about frame sizes and also about one or several auxiliary variables in one or two frames. Additional functions for confidence interval estimation based on the jackknife variance estimation are being included as well.

References

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

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020.

Hartley, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 203–206.

Kalton, G. and Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General)*, 149:65–82.

Lohr, S. L. (2009). Multiple-frame surveys. *Handbook of Statistics*, 29:71–88.

Ranalli, M. G., Arcos, A., Rueda, M. d. M., and Teodoro, A. (2014). Calibration estimation in dual frame surveys. *arXiv preprint arXiv:1312.0761v2*.

Rao, J. N. K. and Wu, C. (2010). Pseudo–empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105(492):1494–1503.

Rueda, M. d. M., Arcos, A., Molina, D., and Ranalli, M. G. (2014). Multinomial logistic estimation in dual frame surveys. *Proceedings of the 14th International Conference on*

Computational and Mathematical Methods in Science and Engineering, CMMSE 2014 37July, 2014, pages 1149–1160.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Skinner, C. J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86:779–784.

Skinner, C. J. and Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91:349–356.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.