Re-calibration of sample means.

E. Greenshtein* and Y. Ritov†

June 23, 2011

Abstract

We consider the problem of calibration and the GREG method as suggested and studied in Deville and Sarndal (1992). We show that a GREG type estimator is typically not minimal variance unbiased estimator even asymptotically. We suggest a similar estimator which is unbiased but is asymptotically with a minimal variance.

1 Introduction

The purpose of this note is to examine the popular calibration techniques, suggested, e.g., in Deville and Sarndal (1992), or Sarndal et.al. (1992) Chapter 6.4, those calibrated estimators are also known as GREG (the general regression estimator). Our development and criterion are elementary. We are interested in finding a minimum variance linear estimator.This leads lead to a very similar to the GREG estimator in form estimator, but with different constants. The difference between these two approaches as demonstrated in what follows. This demonstration is the main purpose of this note.

First we review the above mentioned calibration GREG approach, following Deville and Sarndal (1992). Consider a finite population $U = \{1, ..., N\}$, and a sample S, $S \subset U$. Denote $\pi_i = P(S \ni i)$, $\pi_{ij} = P(S \supseteq \{i, j\})$. Let $(y_i, \mathbf{x_i})$, be quantities associtaed with item i, $i \in U$, here y_i is a scalar and $\mathbf{x_i}$ is a vector. The quantity of interest is $t_Y = \sum_U y_i$, while $\mathbf{x_i}$ are considered as covariates. Suppose the total $\mathbf{t}_X = \sum_U \mathbf{x_i}$ is known, w.l.o.g., $\mathbf{t}_X = 0$. Then, it is suggested to utilize that information about the totals through the following reasoning.

*Israel Central Bureau of Statistics. eitan.greenshtein@gmail.com

†Department of Statistics, The Hebrew University of Jerusalem. yaacov.ritov@gmail.com

Define $\hat{\mathbf{t}}_X = \sum_{i\in S} \mathbf{x_i}/\pi_i \equiv \sum_{i\in S} d_i\mathbf{x_i}$ and $\hat{t}_Y = \sum_{i\in S} y_i/\pi_i \equiv \sum_{i\in S} d_i y_i$, where $d_i = 1/\pi_i$. The above are the Horowitz-Thompson estimators, hence we have $E\hat{t}_y = t_Y$ and $E\hat{\mathbf{t}}_X = \mathbf{t}_X = 0$. However, the value of $\hat{\mathbf{t}}_X$ is typically different than 0, which is unfortunate.

It is suggested to find "better" or "improved" weights $w_i,\ i \in S$ ("better" than d_i) and estimate t_y by $\sum_{i\in S} w_i y_i$. The heuristic derivation of the improved (random) weights $w_i,\ i \in S$ is the following. Given S denote by $\mathbf{w}$ the vector of improved weights. Then $\mathbf{w}$ is defined as the solution of the program:

$$\begin{aligned}\min_{\boldsymbol{\omega}} \sum_{i\in S}(\omega_i - d_i)^2/d_i q_i \\ s.t. \sum_{i\in S}\omega_i \mathbf{x}_i = 0;\end{aligned} \tag{1}$$

here, the q_i are selected parameters, which, as a default, suggested to be set to 1. The resulting estimator denoted $\hat{t}_{y|\mathbf{x}}$, may be written as: $\hat{t}_{y|\mathbf{x}} = \sum w_k y_k$.

The solution of (2) is simple. Using a vector of Lagrange multipliers λ we can find that

$$w_i = (1 + \lambda^{\mathsf{T}} q_i \mathbf{x}_i) d_i.$$

where λ is such that the constraint is satisfied, namely

$$\begin{aligned}\lambda &= -\Big(\sum_{i\in S} q_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}})\Big)^{-1} \sum_{i\in S} d_i \mathbf{x}_i \\ &\equiv -H_q^{-1}\hat{t}_X,\end{aligned}$$

where $H_q = \sum_{i\in S} q_i d_i \mathbf{x}_i \mathbf{x}_i$. Hence

$$\hat{t}_{y|\mathbf{x}} = \hat{t}_Y - \hat{\boldsymbol{\beta}}^{\mathsf{T}}\hat{t}_X,$$

where $\hat{\boldsymbol{\beta}} = H_q^{-1}\sum_{i\in S} d_i q_i y_i \mathbf{x}_i$.

In the following we consider weights $q_i \equiv 1$, and denote then H_q simply by H.

Note that for any (pre-determined) $\boldsymbol{\beta}$, $\hat{t}_Y - \boldsymbol{\beta}^{\mathsf{T}}\hat{t}_X$ is an unbiased estimator of t_Y. Hence we may look for the minimal variance estimator of this type. One may restrict himself to a linear estimator (linear in $Y_i,\ i \in S$). That is, an estimator of the form $\sum w_i y_i$, with a sequence of weight that could simultaneously be used for getting an estimator for any functional. Still one may look for such weights that would ensure that the estimator has a minimal variance. We will argue that the weights given by (2) are generally speaking, far from being optimal.

Similar problem were discussed in Bickel, Klaassen, Ritov, and Wellenr (1998) in the context of i.i.d. observations and semiparametric models. The question there, was defined as the semiparametric efficient estimation of parameter, when other parameters are known (e.g., estimation of the joint distribution, when the marginal distributions are known). Our solution is similar to the examples analyzed in that literature.

2 Minimum variance linear unbiased estimator

Consider estimators of t_Y which are linear in $\hat{t}_Y$ and $\hat{t}_X$, i.e., of the form

$$T(\boldsymbol{\beta}) = \hat{t}_Y - \boldsymbol{\beta}^{\mathsf{T}}\hat{\mathbf{t}}_X,$$

where $\boldsymbol{\beta}$ is non-random. The above class is unbiased since $E\hat{\mathbf{t}}_X = 0$. Consider the estimator $T(\boldsymbol{\beta}_o)$ in the above class with minimal variance. Clearly,

$$\boldsymbol{\beta}_o = \Sigma_{\hat{\mathbf{t}}_X}^{-1}\Sigma_{\hat{\mathbf{t}}_X,\hat{t}_Y}, \tag{2}$$

where $\Sigma_{\hat{\mathbf{t}}_X}$ and $\Sigma_{\hat{\mathbf{t}}_X,\hat{t}_Y}$ are the variance-covariance matrix of $\hat{\mathbf{t}}_X$, and the covariance vector of $\hat{\mathbf{t}}_X$ and $\hat{t}_Y$, respectively.

First we argue that $\hat{\boldsymbol{\beta}}$ is not a consistent estimator of $\boldsymbol{\beta}_0$. The following example, while being extreme, is enlightening.

Example 2.1 Consider a population divided into two stratas of equal sizes. For each $i \in U$ there is a corresponding y_i and x_i, i.e., we have one dimensional covariates. Suppose we randomly sample n units from each strata, i.e., a total of $M = 2n$ where $\pi_i \equiv M/N$.

Assume the mean of x_i in stratum 1 is -1 and their mean in stratum 2 is 1. The variance of x_i within each stratum is σ^2. Now assume in stratum 1, $y_i \equiv -1$, while in stratum 2, $y_i \equiv 1$. Therefore, $\mathrm{Var}(\hat{t}_Y) = 0$, and hence $\boldsymbol{\beta}_o = 0$. In fact, the optimal estimator in this case is simply $\hat{t}_Y \equiv 0$, on the other hand, $\hat{\beta} = H^{-1}\sum_{i\in S} d_i y_i \mathbf{x}_i \xrightarrow{\mathrm{p}} (1+\sigma^2)^{-1}$. Asymptotically (as $n \to \infty$) the GREG estimator $T(\hat{\beta}) \approx -(1+\sigma^2)^{-1}\hat{\mathbf{t}}_X$ has therefor variance of order N^2/n, while the optimal estimator for this case is exact with zero variance.

The difference between $\boldsymbol{\beta}_o$ and $\hat{\boldsymbol{\beta}}$ would be large, when there is more than a scale difference between the second moments of $\hat{t}_Y, \hat{t}_X$ and of those of Y, X. This precludes the simple random sample, but is typical for other sampling scheme. The following example is less extreme than the first one, but describes a practical situation.

Example 2.2 Suppose we sample clusters, the units in the sample are indexed by j and k, where all units in cluster j, refer to the same central value s_j, and satisfy $x_{jk} = s_j + \varepsilon_{jk}$ and $y_{jk} = s_j + \gamma\nu_{jk}$, where the correlation between ε_{jk} and ν_{jk} is 0. Suppose that K units are sampled in each cluster. It is clear that if the number of clusters is large, then with obvious notation: $\boldsymbol{\beta}_o = \Sigma_s/(\Sigma_s + \Sigma_\varepsilon/K)$ while $\hat{\boldsymbol{\beta}} \xrightarrow{\text{p}} \Sigma_s/(\Sigma_s + \Sigma_\varepsilon)$. In the simple case where $\Sigma_s = \Sigma_\varepsilon = \Sigma_\nu$, if $K = 5$ then the estimator with β_0 would have a variance smaller by approximately 25% than the variance of the estimator using $\hat{\beta}$. The difference is approximately 50% when $K = 10$.

In order to estimate the $\Sigma_{\hat{\mathbf{t}}_X}$ and $\Sigma_{\hat{t}_Y,\hat{\mathbf{t}}_\mathbf{X}}$, we may use the classical variance estimators for Horovitz-Thompson estimator, see, e.g., Cochran (1977) or Sharon (1999). Those estimators are typically given in the literature for one dimensional variance rather than to a covariance matrix, however the same reasoning applies. Since $\mathbf{t}_X = 0$,

$$\begin{aligned}
\Sigma_{\hat{t}_Y,\hat{\mathbf{t}}_\mathbf{X}} &= \mathrm{E}\sum_{i,j\in S} \frac{1}{\pi_i\pi_j} y_i\mathbf{x}_j \\
&= \sum_{i,j\in U} \frac{\pi_{ij}}{\pi_i\pi_j} y_i\mathbf{x}_j \\
&= \sum_{i,j\in U} \Big(\frac{\pi_{ij}}{\pi_i\pi_j} - c\Big) y_i\mathbf{x}_j, \quad \forall c.
\end{aligned}$$

Similarly,

$$\Sigma_{\hat{t}_X} = \sum_{i,j\in U} \Big(\frac{\pi_{ij}}{\pi_i\pi_j} - c\Big) \mathbf{x}_i\mathbf{x}_j^{\mathsf{T}}, \quad \forall c.$$

Hence, the following are unbiased estimators:

$$\begin{aligned}
\hat{\Sigma}_{\hat{t}_X,\hat{t}_Y} &= \sum_{i,j\in S} \frac{1}{\pi_{ij}} \Big(\frac{\pi_{ij}}{\pi_i\pi_j} - c\Big) y_i\mathbf{x}_j, \quad \forall c. \\
\hat{\Sigma}_{\hat{t}_X} &= \sum_{i,j\in S} \frac{1}{\pi_{ij}} \Big(\frac{\pi_{ij}}{\pi_i\pi_j} - c\Big) \mathbf{x}_i\mathbf{x}_j^{\mathsf{T}}, \quad \forall c.
\end{aligned} \tag{3}$$

We assume that we consider a sequence of populations and designs such that the estimators in (3) are consistent. Typically, taking $c = 1$ in (3), would suffice to make most of the terms of lower order than the diagonal. In a simple random sample without replacement, $c = (n-1)N/n(N-1)$ leaves only the diagonal.

Theorem 2.1 *Let* $\hat{\boldsymbol{\beta}}_o = \hat{\Sigma}^{-1}_{\hat{\mathbf{t}}_X}\hat{\Sigma}_{\hat{t}_Y,\hat{\mathbf{t}}_X}$*, where the terms in the RHS are given by* (3) *with a given* c*. Then*

$$T(\hat{\boldsymbol{\beta}}_o) = \sum_{i\in S} w_i y_i,$$

where

$$w_i = \frac{1}{\pi_i} - \hat{\mathbf{t}}_X^{\mathsf{T}}\hat{\Sigma}^{-1}_{\hat{\mathbf{t}}_X}\sum_{j\in S}\frac{1}{\pi_{ij}}\Big(\frac{\pi_{ij}}{\pi_i\pi_j} - c\Big)\mathbf{x}_j, \quad i \in S.$$

Thus the weights are a function of $\mathbf{x}_i$*,* π_i*,* π_{ij}*,* $i,j \in S$ *only. In particular* $\sum_{i\in S} w_i\mathbf{x}_i = 0$*.*

3 Examining $\hat{\boldsymbol{\beta}}$ under linear model assumptions.

In this section we will examine the rational in the estimator $\hat{\boldsymbol{\beta}}$ under the convenient and (too) often assumed super-population model under which $Y_k = \boldsymbol{\beta}\mathbf{x}_k + \epsilon_k$, where $E\epsilon_k = 0$ and for simplicity assume that ϵ_k, $k \in U$ have equal variance.

Under this model it is easy to check that $\Sigma_{XY} = \boldsymbol{\beta}\Sigma_{XX}$, and $\Sigma_{\mathbf{t}_X,t_Y} = \boldsymbol{\beta}\Sigma_{\mathbf{t}_\mathbf{X}}$. Hence $\hat{\boldsymbol{\beta}}$ is a possible estimator of $\boldsymbol{\beta}_o = \boldsymbol{\beta}$. However, if this model is assumed, it is still not clear why $\hat{\boldsymbol{\beta}}$ should be used. We have here a standard regression problem. Elementary regression theory (namely the Gauss-Markov Theorem) implies that the optimal estimator is not $\hat{\boldsymbol{\beta}}$, but the standard un-weighted linear regression of $Y_1, \dots, Y_n$ on $\mathbf{x}_1, \dots, \mathbf{x}_n$.

It might be argued that in fact we are taking the linear model super-population assumption with a grain of salt, and thus we are using the estimator for

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} \sum_{i\in U}(y_i - \mathbf{b}^{\mathsf{T}}\mathbf{x})^2, \tag{4}$$

which is defined under no linear model assumptions. However, since in this case we have no interest in that population parameter per se, but just in a tool for construction a good estimator for t_Y, than $\boldsymbol{\beta}_o$ should be our target.

To summarize. If we are interested in the super-population parameter, than $\hat{\boldsymbol{\beta}}$ is not efficient, and if we are interested in good estimator of t_Y, than $\hat{\boldsymbol{\beta}}$ is not consistent under complex sampling schemes.

4 A partial knowledge of $\mathbf{t}_X$

In many cases $\mathbf{t}_X$ is not really known. However, it might be that there is an additional independent sample with information about $\mathbf{t}_X$ but not about t_Y. Thus we have three unbiased estimators $\hat{t}_Y^1, \hat{\mathbf{t}}_X^1$ based on one sample, and $\hat{\mathbf{t}}_X^2$ based on another independent sample. The best estimator of t_Y would be based on these three. Following the same argument as before we should consider estimator of the form

$$\hat{\mathbf{t}}_Y^1 - \boldsymbol{\beta}^\mathsf{T}(\hat{\mathbf{t}}_X^1 - \hat{\mathbf{t}}_X^2).$$

Note that this estimator yields an unbiased estimate of t_Y for any $\boldsymbol{\beta}$. The optimal value, however, is given by

$$\boldsymbol{\beta}_{o2} = -\big(\Sigma_{\hat{\mathbf{t}}_X^1} + \Sigma_{\mathbf{t}_X^2}\big)^{-1}\Sigma_{\hat{\mathbf{t}}_X^1,\hat{\mathbf{t}}_Y^1}.$$

Note, that if $\hat{\mathbf{t}}_X^2$ is based on the all universe U, then $\Sigma_{\hat{\mathbf{t}}_X^2} = 0$, and $\beta_{o2} = \beta_o$.

Even more generally, we can consider a situation in which $\mathbf{x}$ is measured for all units in the a super sample S_2, $S_1 \subseteq S_2 \subseteq U$, while the y values are measured only for units in the smaller sample S_1. For example, y is measured only for one unit in a cluster, while the $\mathbf{x}$ is measured for all units. Let $\hat{\delta}_X = \mathbf{t}_X^1 - \mathbf{t}_X^2$. It may be natural to assume that $\hat{\delta}_X$ is correlated with $\hat{t}_Y$ while having a mean 0. We consider the natural extension $\hat{t}_Y^1 - \boldsymbol{\beta}_{o3}^\mathsf{T}\hat{\boldsymbol{\delta}}_X$, with $\boldsymbol{\beta}_{o3} = -\Sigma_{\hat{\boldsymbol{\delta}}_X}^{-1}\Sigma_{\hat{\boldsymbol{\delta}}_X,\hat{t}_Y^1}$.

Example 4.1 Consider the super-population model in which it is assumed that $y_{j,k} = \beta x_{j,k} + \varepsilon_{j,k}$, $j = 1,\ldots,M$, $k = 1,\ldots,K$ where $\varepsilon_{j,k}$ are i.i.d., independent of $x_{j,k}$, while $x_{j',k'}$ and $x_{j,k}$ are independent if $j \neq j'$, and have correlation ρ if $j = j'$ and $k \neq k'$. Let $\mathrm{Var}(x_{j,k}) = \sigma^2$. Consider the sample $C \subset \{1,\ldots,M\}$ of n clusters. Suppose that for each $j \in C$, $x_{j,k}$, $k = 1,\ldots,K$ are obtained, while only $y_{j,1}$ is measured, assume also that $M \gg n$. The universe size is $N = MK$. Hence, we assume for simplicity a simple random sample (with replacement) of clusters. Then

$$\hat{\delta}_X = \frac{N}{n}\sum_{j\in C}\big(x_{j,1} - K^{-1}\sum_{k=1}^{K} x_{j,k}\big)$$

It is easy to verify

$$\mathrm{Var}(\hat{t}_Y^1) = \frac{N^2}{n}\beta^2\sigma^2$$

$$\text{Var}(\hat{\delta}_X) = \frac{N^2}{n}\frac{K-1}{K}(1-\rho)\sigma^2$$
$$\text{cov}(\hat{\delta}_X, \hat{t}^1_Y) = \frac{N^2}{n}\frac{K-1}{K}(1-\rho)\beta\sigma^2.$$

Hence

$$\frac{\text{Var}(\hat{t}^2_Y - \beta_{o3}\hat{\delta}_X)}{\text{Var}(\hat{t}^2_Y)} = 1 - \frac{K-1}{K}(1-\rho).$$

The efficiency of the scheme increases as K increases and ρ decreases. Note that the case of a simple random sample of units in which the y value is measured only for a small random sub-sample, corresponds to $\rho = 0$.

References

Bickel, P.J., Klassen, C.A.J., Ritov, Y., Wellner, J.A. (1998). *Efficient and adaptive estimation for semiparametric models.* Springer Verlag New York Inc.

Cochran, W.G (1977). *Sampling techniques.* 3d ed. New York: Wiley.

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *JASA*, **87**, 376-382.

Sahron L. Lohr (1999). *Sampling Design and Analysis.* Brooks/Cole publishing company.

Sarndal, C. E., Swenson, B., Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag.