# An Optimal Calibration Distance Leading to the Optimal Regression Estimator

**DANIEL THORBURN[1] and PER GÖSTA ANDERSSON[2]**

**Abstract**

Where there is auxiliary information in survey sampling, the design based "optimal (regression) estimator" of a finite population total/mean is known to be (at least asymptotically) more efficient than the corresponding GREG-estimator. The GREG-estimator was originally constructed using an assisting linear superpopulation model. It may also be seen as a calibration estimator; i.e. as a weighted linear estimator, where the weights obey the calibration equation and, with that restriction, are as close as possible to the original "Horvitz-Thompson weights" (according to a suitable distance). We show that also the optimal estimator can be seen as a calibration estimator in this respect, with a quadratic distance measure closely related to the one generating the GREG-estimator. Simple examples will also be given, revealing that this new measure is not always easily obtained though.

KEY WORDS: Horvitz-Thompson estimator; Regression estimator; Survey sampling theory.

[1]Daniel Thorburn, Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden..

[2]Per Gösta Andersson, Division of Mathematical Statistics, Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden

## 1. NOTATION AND BASICS

Consider a finite population $U$ consisting of $N$ objects labelled $1, \ldots, N$ with associated study values $y_1, \ldots, y_N$ and $J$-dimensional auxiliary (column-) vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. We want to estimate the population total $t_y = \sum_{i \in U} y_i$ by drawing a random sample $s$ of size $n$ (fixed or random) from $U$, with first and second order inclusion probabilities $\pi_i = P(i \in s)$, $\pi_{ij} = P(i, j \in s)$, $i, j = 1, \ldots, N$. The study values and the auxiliary vectors are recorded for the sampled objects and before the sample is drawn we assume that at least $\boldsymbol{t}_x = \sum_{i \in U} \boldsymbol{x}_i$ is known.

Finally some comments on matrix notation in this paper: Generally, the transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$ and if $\boldsymbol{A}$ is square, the inverse (a generalised inverse) is written $\boldsymbol{A}^{-1}(\boldsymbol{A}^-)$. We further let the column vectors $\boldsymbol{y} = (y_i)_{i \in s}$ and $\mathbf{w}_0 = (1/\pi_i)_{i \in s}$, $\boldsymbol{X}$ be the $J \times n$ "design" matrix of the auxiliary information from $s$ and finally $\boldsymbol{I}_n$ means a unit diagonal matrix of size $n$.

## 2. REGRESSION AND CALIBRATION ESTIMATORS

An unbiased simple estimator of $t_y$ is the Horvitz-Thompson estimator $\hat{t}_y = \sum_{i \in s} y_i/\pi_i = \boldsymbol{y}^T \boldsymbol{w}_0$. However, more efficient estimators may be obtained utilising the auxiliary information, e.g. the well-known model assisted GREG-estimator, see Särndal et al. (1992). Constructed from the assumption of e.g. a homoscedastic linear regression superpopulation model, the GREG estimator is

$$\hat{t}_{yr} = \boldsymbol{y}^T\boldsymbol{w}_0 + (\boldsymbol{y}^T\boldsymbol{R}_r\boldsymbol{X}^T)(\boldsymbol{X}\boldsymbol{R}_r\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x) \quad (1)$$

$$= \boldsymbol{y}^T\boldsymbol{g}, \quad (2)$$

where $\boldsymbol{R}_r = \boldsymbol{w}_0\boldsymbol{I}_n$, $\hat{\boldsymbol{t}}_x = \sum_{i\in s}\boldsymbol{x}_i/\pi_i$ and

$$\boldsymbol{g} = \left(\frac{1}{\pi_i}(1 + \boldsymbol{x}_i^T(\boldsymbol{X}\boldsymbol{R}_r\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x))\right)_{i\in s}.$$

Now, the expression (2) for the GREG-estimator is interesting since we also have that

$$\boldsymbol{x}^T\boldsymbol{g} = \boldsymbol{t}_x, \quad (3)$$

which is called the *calibration equation*. This brings us to an alternative possible derivation of the GREG-estimator according to Deville and Särndal (1992). Suppose that we seek an estimator $\boldsymbol{y}^T\boldsymbol{w}$ of $t_y$ with a vector $\boldsymbol{w}$ of sample-dependent weights $(w_i)_{i\in s}$, which respects the corresponding calibration equation, while also minimising the distance between $\boldsymbol{w}$ and $\boldsymbol{w}_0$ according to the quadratic distance measure

$$(\boldsymbol{w} - \boldsymbol{w}_0)^T\boldsymbol{R}(\boldsymbol{w} - \boldsymbol{w}_0),$$

where $\mathbf{R} = (\boldsymbol{w}_0\boldsymbol{I}_n)^{-1}$.

This results in

$$\boldsymbol{w} = \boldsymbol{w}_0 + \boldsymbol{R}^{-1}\boldsymbol{x}^T(\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x), \quad (4)$$

which means that $\boldsymbol{w} = \boldsymbol{g}$, since here $\boldsymbol{R} = \boldsymbol{R}_r^{-1}$.

Turning to the optimal estimator, consider first the vector $(\hat{t}_y, \hat{\boldsymbol{t}}_x^T)$ and let $\boldsymbol{\Sigma}_{y,x}$ be the covariance (row) vector of $\hat{t}_y$ and $\hat{\boldsymbol{t}}_x$ and $\boldsymbol{\Sigma}_{x,x}$ the covariance matrix of $\hat{\boldsymbol{t}}_x$. Now, the minimum-variance, see Montanari (1987), unbiased linear estimator (in $\hat{t}_y$ and $\hat{\boldsymbol{t}}_x$) of $t_y$ is the difference estimator

$$\hat{t}_y + \boldsymbol{\Sigma}_{y,x}\boldsymbol{\Sigma}_{x,x}^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x). \tag{5}$$

Since $\boldsymbol{\Sigma}_{y,x}$ and $\boldsymbol{\Sigma}_{x,x}$ in practice are unknown, we let the optimal estimator be

$$\begin{aligned} \hat{t}_{y\,opt} &= \boldsymbol{y}^T\boldsymbol{w}_0 + \widehat{\boldsymbol{\Sigma}}_{y,x}\widehat{\boldsymbol{\Sigma}}_{x,x}^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x) \\ &= \hat{t}_y + (\boldsymbol{y}^T\boldsymbol{R}_{opt}\boldsymbol{X}^T)(\boldsymbol{X}\boldsymbol{R}_{opt}\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x), \end{aligned} \tag{6}$$

where $\boldsymbol{R}_{opt} = \Big((\pi_{ij} - \pi_i\pi_j)/(\pi_{ij}\pi_i\pi_j)\Big)_{i,j\in s}$.

In an asymptotic context, where $n \to \infty$ and $N \to \infty$, $\widehat{\boldsymbol{\Sigma}}_{x,y}$ and $\widehat{\boldsymbol{\Sigma}}_{x,x}$ may be viewed as components of the asymptotic covariance matrix of $(\hat{t}_y, \hat{\boldsymbol{t}}_x^T)$. Under the assumption of consistency of $\widehat{\boldsymbol{\Sigma}}_{x,y}$ and $\widehat{\boldsymbol{\Sigma}}_{x,x}$, which holds under very mild conditions, see Andersson et al. (1995), the optimal estimator has the same asymptotic variance as the difference estimator (5). In particular it follows that the optimal estimator is asymptotically better than the usual GREG estimator, see Rao (1994), Montanari (2000) and Andersson (2001), i.e. its asymptotic variance is never larger and usually smaller. However, one does not know anything about the efficiency for finite samples, since the covariance estimator may converge slowly. Note also that in some cases there exist asymptotically even better estimators which are not linear.

Now, the fact that the GREG-estimator is also a calibration estimator

using

$$(\boldsymbol{w}-\boldsymbol{w}_0)^T\boldsymbol{R}_r^{-1}(\boldsymbol{w}-\boldsymbol{w}_0) \tag{7}$$

as the distance measure and comparing (1) with (6), leads one to believe that replacing $\boldsymbol{R}_r$ by $\boldsymbol{R}_{opt}$ in (7) should imply that we instead derive the optimal regression estimator as a calibration estimator. That this actually holds is shown below.

## 3. THE MAIN RESULT

In order to show existence of a distance measure corresponding to the optimal estimator, we will first state and prove a result in the general case.

**Lemma**: With $\boldsymbol{R}$ denoting an arbitrary positive definite $n \times n$ matrix,

$$(\boldsymbol{w}-\boldsymbol{w}_0)^T\boldsymbol{R}(\boldsymbol{w}-\boldsymbol{w}_0) \tag{8}$$

is subject to the constraint $\boldsymbol{X}\boldsymbol{w}=\boldsymbol{t}_x$ minimised by

$$\boldsymbol{w}=\boldsymbol{w}_0+\boldsymbol{R}^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x-\hat{\boldsymbol{t}}_x).$$

**Proof**: Introducing the $J \times 1$ vector $\boldsymbol{\lambda}$ of Lagrange multipliers, we get after differentiation the equation system

$$2\boldsymbol{R}(\boldsymbol{w}-\boldsymbol{w}_0)+\boldsymbol{X}^T\boldsymbol{\lambda} = \boldsymbol{0} \tag{9}$$

$$\boldsymbol{X}\mathrm{W}-\boldsymbol{t}_x = \boldsymbol{0} \tag{10}$$

Multiplying (9) by $\boldsymbol{X}\boldsymbol{R}^{-1}$, using (10) and solving for $\boldsymbol{\lambda}$, yields with $\boldsymbol{X}\boldsymbol{w}_0=\hat{\boldsymbol{t}}_x$:

$$\boldsymbol{\lambda}=2(\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\hat{\boldsymbol{t}}_x-\boldsymbol{t}_x). \tag{11}$$

Putting this into (9) and solving for $\boldsymbol{w}$ finally leads to

$$\boldsymbol{w} = \boldsymbol{w}_0 + \boldsymbol{R}^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^T)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_x).$$

From the lemma we thus have the following main result:

**Theorem**: With $\boldsymbol{R}_{opt}$ being positive (semi-) definite and using the optimal calibration distance-measure, which we get by letting $\boldsymbol{R} = \boldsymbol{R}_{opt}^{-1}$ ($\boldsymbol{R}_{opt}^{-}$) in (8), the calibration estimator will become the optimal regression estimator.

**Remark**: $\boldsymbol{R}_{opt}$ may in some cases be indefinite (see below). The only thing we know is that it is an unbiased estimator of a covariance matrix. If it is not positive semi-definite there also exist $x$-values such that $\boldsymbol{X}\boldsymbol{R}_{opt}\boldsymbol{X}^T$ is not positive semi-definite, but the probability of such $x$-values goes to zero as the population and sample sizes increase (and if $\boldsymbol{\Sigma}_{x,x}$ is positive definite). A strict minimisation of a distance with "a negative component" would lead to infinitely large corrections. This problem of the optimal estimator has, to our knowledge, not been pointed out previously.

The simplest way to find a distance which gives the optimal estimator as a calibration estimator is to find a matrix $\boldsymbol{R}_{dist}$ which has the same eigenvectors as $\boldsymbol{R}_{opt}$ but where the eigenvalues are replaced by their absolute values. (This result can be shown along the same lines as the proof of the lemma above. The distance can be seen as the sum of the products of the eigenvalues and the squared eigenvectors. Putting the derivatives equal to zero means that in the proposition we found the extremes i.e. the minima for positive eigenvalues and the maxima for negative eigenvalues. By changing all negative signs the extremes will all be minima).

## 4. EXAMPLES

*Positive-definite* $\boldsymbol{R}_{opt}$ : Suppose that the objects in $U$ are independently drawn with inclusion probabilities $\pi_1, \ldots, \pi_N$ (Poisson sampling); thus implying a random sample size $n$, where $E[n] = \sum_{i \in U} \pi_i$. Due to the independence of drawings, $\boldsymbol{R}_{opt}$ is diagonal and specifically

$$\boldsymbol{R}_{opt}^{-1} = \boldsymbol{I}_n \Big( \frac{\pi_i^2}{1 - \pi_i} \Big)_{i \in s}.$$

*Positive-semidefinite* $\boldsymbol{R}_{opt}$ : Suppose $n$ objects are drawn according to simple random sampling, i.e. each object has inclusion probability $\pi_i = n/N$. The elements of $\boldsymbol{R}_{opt}$ then are

$$\begin{aligned} i = j: &\quad \Big(\tfrac{N}{n}\Big)^2 \tfrac{N-n}{N} \\ i \neq j: &\quad \Big(\tfrac{N}{n}\Big)^2 \tfrac{n-N}{N(n-1)}. \end{aligned}$$

This means that $\boldsymbol{R}_{opt}$ is singular with rank $n - 1$.

*Non-positive semi-definite* $\boldsymbol{R}_{opt}$ : Let U consist of four elements and s of two elements. Suppose that a systematic sample is taken with probability 0.94 and a simple random sample with probability 0.06, i.e. $\pi_{13} = \pi_{24} = 0.48$ and $\pi_{12} = \pi_{14} = \pi_{23} = \pi_{34} = 0.01$. In that case

$$\boldsymbol{R}_{opt} = \begin{pmatrix} 2 & 23/12 \\ 23/12 & 2 \end{pmatrix} \tag{12}$$

with probability 0.96 and

$$\boldsymbol{R}_{opt} = \begin{pmatrix} 2 & -96 \\ -96 & 2 \end{pmatrix} \tag{13}$$

with probability 0.04. The second matrix has a negative eigenvalue.

The problem does not necessarily disappear if $N$ is large. Consider instead a population consisting of N/4 strata with four elements each. Suppose that the above sampling procedure is used independently in each stratum. In that case $\boldsymbol{R}_{opt}$ will consist of a matrix with the above $2 \times 2$-matrices along the diagonal and zeroes elsewhere.

## REFERENCES


ANDERSSON, P.G. (2001). Improving estimation quality in large sample surveys. *Ph.D. Thesis*, Department of Mathematics, Chalmers University of Technology and Göteborg University.

ANDERSSON, P.G., NERMAN, O., and WESTHALL J. (1995). Auxiliary information in survey sampling. *Technical Report NO 1995:3*, Department of Mathematics, Chalmers University of Technology and Göteborg University.

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

MONTANARI, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

MONTANARI, G.E. (2000). Conditioning on auxiliary variable means in finite population inference. *Australian & New Zeeland Journal of Statistics*, 42, 407-421.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model assisted survey sampling*. New York: Springer Verlag.