An Introduction to Calibration Estimators

Jennifer H. Nguyen
University of Waterloo
j6nguyen@uwaterloo.ca

Abstract: In survey sampling, the use of auxiliary information can greatly improve the precision of estimates of population total and/or means. In this paper, we explain the basic theory and use of calibration estimators proposed by Deville and Särndal, which incorporate the use of auxiliary data. Results of a simulation study conducted using real data from the 2008 Survey of Household Spending by Statistics Canada are presented, comparing the performance of two calibration estimators against the Horvitz-Thompson estimator. Limitations of calibration estimators and recent extensions made by other leading statisticians in this topic are also discussed.

1 Introduction

The technique of estimation by calibration was introduced by Deville and Särndal in 1992 [DS92]. The idea is to use auxiliary information to obtain a better estimate of a population statistic. First, consider a finite population U of size N with unit labels $1, 2, \ldots, N$. Let y_i, $i = 1, \ldots, N$ be the study variable and $\mathbf{x}_i$, $i = 1, \ldots, N$ be the k-dimensional vector of auxiliary variables associated with unit i.

Suppose we are interested in estimating the population total $t_y = \sum_{i=1}^{N} y_i$. We draw a sample $s = \{1, 2, \ldots, n\} \subset U$ using a probability sampling design P, where the first and second order inclusion probabilities are $\pi_i = Pr(i \in s)$ and $\pi_{ij} = Pr(i, j \in s)$ respectively. An estimate of t_y is the Horvitz-Thompson (HT) estimator

$$\widehat{t}_{HT} = \sum_{i \in s} d_i y_i,$$

where $d_i = 1/\pi_i$ is the sampling weight, defined as the inverse of the inclusion probability for unit i.[1] An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design P. Its variance under P is given as

$$V_p(\widehat{t}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{1.1}$$

Now let us suppose that $\{\mathbf{x}_i,\ i = 1, \ldots, N\}$ is available and $\mathbf{t}_x = \sum_{i=1}^{N} \mathbf{x}_i$, the population total for $\mathbf{x}$, is known. Ideally, we would like

$$\sum_{i \in s} d_i \mathbf{x}_i = \mathbf{t}_x,$$

but often times this is not true.

The idea behind calibration estimators is to find weights w_i, $i = 1, \ldots n$ close to d_i, based on a distance function, such that

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{t}_x. \tag{1.2}$$

[1] Please note that d_i and $\frac{1}{\pi_i}$ are used interchangeably in this paper.

We wish to find weights w_i similar to d_i so as to preserve the unbiased property of the HT estimator. Once w_i is found, the calibration estimator for t_y is

$$\widehat{t}_c = \sum_{i \in s} w_i y_i.$$

In Section 2, we discuss how to find w_i for a given sample s and the choice of distance function. The relationship of the calibration estimator to the generalized regression (GREG) estimator is also mentioned. In Section 3, we discuss the expectation and variance of $\widehat{t}_c$ and how to perform variance estimation. Section 4 presents the results of a simulation study to test the efficiency of $\widehat{t}_c$ against $\widehat{t}_{HT}$ using two different distance functions. In Section 5, we discuss advancements made by statisticians on the subject of calibration estimators.

2 Derivation of the Calibration Estimator

Given a sample s, we want to find w_i close to d_i based on a distance function $D(w, d)$ subject to the constraint in Equation 1.2. This is an optimization problem where we wish to minimize

$$Q(w_1, \ldots, w_n, \boldsymbol{\lambda}) = \sum_{i \in s} D(w_i, d_i) - \boldsymbol{\lambda} \left(\sum_{i \in s} w_i \mathbf{x}_i - \mathbf{t}_x \right) \tag{2.1}$$

using the method of Lagrange multipliers.

Examples of distance functions are presented in Table 2.1. We will derive the calibration weights using the Chi-squared distance $(w-d)^2/2qd$ (see Table 2.1), where q is a tuning parameter that can be manipulated to achieve the optimal minimum of Equation 2.1. Note that in practice, the choice of distance function depends on the statistician and the problem.

Table 2.1: Examples of distance functions $D(w, d)$ adapted from Deville and Särndal [DS92]

	$D(w, d)$
1. Chi-squared distance	$(w-d)^2/2qd$
2. Modified minimum entropy distance	$q^{-1}(w \log(w/d) - w - d)$
3. Hellinger distance	$2(\sqrt{w} - \sqrt{d})^2/q$
4. Minimum entropy distance	$q^{-1}(-d \log(w/d) + w - d)$
5. Modified chi-squared distance	$(w-d)^2/2qw$

Letting $D(w_i, d_i) = (w_i - d_i)^2/2q_i d_i$ in Equation 2.1 and differentiating with respect to w_i, we get

$$\frac{\partial Q}{\partial w_i} = \frac{(w_i - d_i)}{q_i d_i} - \boldsymbol{\lambda} \mathbf{x}_i. \tag{2.2}$$

Setting Equation 2.2 to zero and solving for w_i we get

$$w_i = d_i(1 + q_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\lambda}). \tag{2.3}$$

Using the constraint in Equation 1.2 we also get,

$$\boldsymbol{\lambda} = \mathbf{T}_s^{-1}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}),$$

where $\mathbf{T}_s = \sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\widehat{\mathbf{t}}_{x_{HT}}$ is the HT estimator for the population total with respect to the auxiliary variable $\mathbf{x}$.

The resulting calibration estimator of t_y is then

$$\begin{aligned}\widehat{t}_c &= \sum_{i\in s} w_i y_i \\ &= \widehat{t}_{y_{HT}} + \sum_{i\in s} d_i q_i \mathbf{x}_i^{\mathrm{T}} \mathbf{T}_s^{-1} (\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}) y_i \\ &= \widehat{t}_{y_{HT}} + \widehat{\mathbf{B}}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}), \end{aligned} \tag{2.4}$$

where $\widehat{\mathbf{B}} = \mathbf{T}_s^{-1} \sum_{i\in s} d_i q_i \mathbf{x}_i y_i$.

Written in this form, we see that $\widehat{t}_c$ is the same as the GREG estimator [CSW76]. In fact, the GREG estimator is a special case of the calibration estimator when the chosen distance function is the Chi-square distance [DS92].

It is important to note that depending on the chosen distance function $D(w, d)$, there may not exist an analytical solution to Equation 2.2 and an approximation of w_i using the Newton-Raphson or a similar method may be required. Furthermore, the solution to Equation 2.2 may yield positive and/or negative weights or extremely large weights, which may be undesirable in a survey sampling context. In terms of efficiency, Deville and Särndal showed that for medium to large samples, the choice of $D(w, d)$ does not make a large impact on the variance of $\widehat{t}_c$ [DS92]. Deville and Särndal also showed that under certain conditions, $\widehat{t}_c$ is asymptotically equivalent to $\widehat{t}_{GREG}$ for any distance function $D(w, d)$ [DS92]. Thus, the choice of distance function is unimportant for large samples, but rather depends on the computational effort of solving Equation 2.2.

3 Expectation and Variance Estimation

To find the expectation and variance of $\widehat{t}_c$, we use the linearization technique to find an approximation of $E_p(\widehat{t}_c)$ and $V_p(\widehat{t}_c)$ with respect to a probability sampling design P. Let $\mathbf{B}$ be the population-level version of $\widehat{\mathbf{B}}$. Then a linear approximation of $\widehat{t}_c$ is

$$\widehat{t}_c \doteq \underbrace{\widehat{t}_{y_{HT}}}_{O_p(1)} + \underbrace{\mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})}_{O_p(n^{-1/2})} + \underbrace{(\widehat{\mathbf{B}} - \mathbf{B})(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})}_{O_p(n^{-1})}, \tag{3.1}$$

where the second term is of order $O_p(n^{-1/2})$ and the last term is of order $O_p(n^{-1})$ as shown by Deville and Särndal [DS92]. Consequently, the last term can be omitted since it is of order $O_p(n^{-1})$. Thus, we can rewrite Equation 3.1 as

$$\widehat{t}_c \doteq \widehat{t}_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}}). \tag{3.2}$$

Using Equation 3.2, the design-based expectation of $\widehat{t}_c$ is

$$E_p(\widehat{t}_c) \doteq E_p\left(\widehat{t}_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})\right) = t_y.$$

Thus, $\widehat{t}_c$ is an approximately designed-unbiased estimator of t_y.

Again using Equation 3.2, the designed-based asymptotic variance of $\widehat{t}_c$ is

$$\begin{aligned} V_p(\widehat{t}_c) &\doteq V_p\left(t_{y_{HT}} + \mathbf{B}(\mathbf{t}_x - \widehat{\mathbf{t}}_{x_{HT}})\right) \\ &= V_p(t_{y_{HT}} - \mathbf{B}\widehat{\mathbf{t}}_{x_{HT}}) \\ &= V_p\left(\sum_{i\in s} d_i(y_i - \mathbf{B}\mathbf{x}_i)\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} (\pi_{ij} - \pi_i\pi_j)\left(d_i(y_i - \mathbf{B}\mathbf{x}_i)\right)\left(d_j(y_j - \mathbf{B}\mathbf{x}_j)\right) \text{ by Equation 1.1.} \end{aligned}$$

Note that since $\mathbf{B}\mathbf{t}_x$ is the true population parameter, $V(\mathbf{B}\mathbf{t}_x) = 0$. The corresponding variance estimator is given by

$$v(\widehat{t}_c) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(d_i (y_i - \widehat{\mathbf{B}} \mathbf{x}_i) \right) \left(d_j (y_j - \widehat{\mathbf{B}} \mathbf{x}_j) \right). \tag{3.3}$$

It is acceptable to use the design weights d_i in the variance estimation but Deville and Särndal suggest that the calibration weights w_i be used in Equation 3.3 as this makes the variance estimator both design-consistent and nearly model-unbiased [DS92]. Moreover, since the calibration estimator is asymptotically equivalent to the GREG estimator, it can be inferred that calibration estimators are more efficient compared to the HT estimator if there is a strong correlation between y and $\mathbf{x}$ [CSW76].

4 Simulation Study

In this section we test the performance of the calibration estimator using distance functions one and two from Table 2.1 against the HT estimator.

4.1 Background and simulation set-up

The data used is obtained from the 2008 Survey of Household Spending conducted by Statistics Canada. There are $N = 9787$ cases. The study variable, y, represents the cost of food purchased from restaurants and the auxiliary variable, x represents the household income before taxes. Figure 4.1 shows a plot of y against x, indicating some positive relationship between restaurant spending and household income, but not a linear relationship.

The statistic of interest is the mean cost of food purchased from restaurants, $\mu_y = t_y/N$, with corresponding estimator $\widehat{\mu}_y = \widehat{t}_y/N$. We treat all $N = 9787$ cases as the finite population. Thus, we know the true population means for y and x are $\mu_y = 1545.74$ and $\mu_x = 71195.52$ respectively.

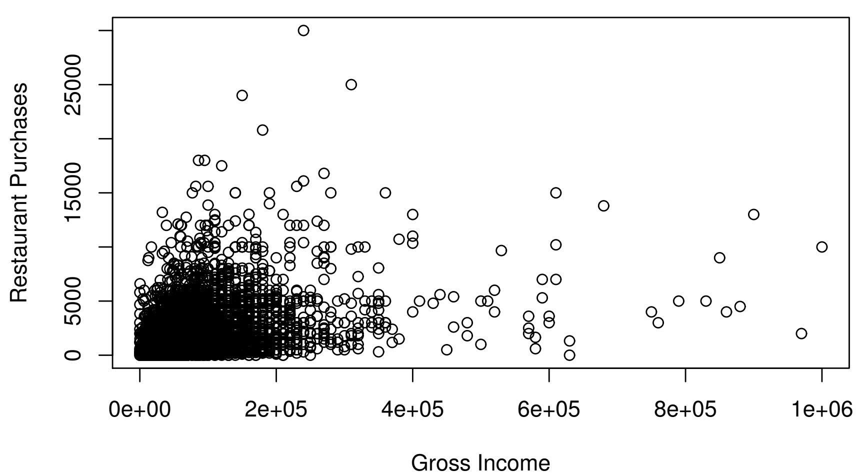

Figure 4.1: Scatter plot of restaurant spending vs. household income [Can10].

A simple regression of the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

was done to see if the model is a good fit of the data. Residual diagnostics show that the residuals meet the basic ordinary least squares assumptions. The R^2 value is only 0.149, indicating that the model explains only some of the variance. Even though the model is poor, we are still assured that $\widehat{t}_c$ is unbiased with respect to the sampling design as demonstrated in Section 3. The correlation between x and y is $\rho_{xy} = 0.387$, which is not strong, but still sufficient to imply that the calibration estimators would provide a better estimate of the total.

The simulation was conducted using the R statistical package. There were $B = 1000$ simulation runs in total. For the b-th run ($b = 1, \ldots, B$), a Bernoulli sample is drawn where each unit is selected into the sample independently with inclusion probability $\pi_i = n/N$. Here we fix $n = 100$. The corresponding HT and calibration estimators of μ_y are computed: $\widehat{\mu}^{(b)}_{y_{HT}}$, $\widehat{\mu}^{(b)}_{y_{c1}}$, and $\widehat{\mu}^{(b)}_{y_{c2}}$. For simplicity, we set the tuning parameter $q_i = 1$. The weights for $\widehat{\mu}_{y_{c1}}$ are given in Equation 2.3. For $\widehat{\mu}_{y_{c2}}$, the weights are of the form

$$w_i = d_i e^{\lambda x_i},$$

where λ is approximated using the constraint in Equation 1.2 with the Newton-Raphson method since there is no closed-form solution.

4.2 Simulation Evaluation

Since each unit is drawn independently, the variance estimators simplify to

$$\begin{aligned} v(\widehat{\mu}_{y_{HT}}) &= N^{-2} \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} y_i^2, \\ v(\widehat{\mu}_{y_{c1}}) &= N^{-2} \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} (y_i - \widehat{B} x_i)^2, \text{ and} \\ v(\widehat{\mu}_{y_{c2}}) &= N^{-2} \sum_{i \in s} (1 - \pi_i)(w_i y_i)^2. \end{aligned}$$

For each estimator of $\widehat{\mu}_y$, a 95% confidence interval $(\widehat{\mu}_L, \widehat{\mu}_U)$ is constructed, where

$$\begin{aligned} \widehat{\mu}_L &= \widehat{\mu}_y - 1.96 \sqrt{v(\widehat{\mu}_y)} \text{ and} \\ \widehat{\mu}_U &= \widehat{\mu}_y + 1.96 \sqrt{v(\widehat{\mu}_y)}. \end{aligned}$$

To compare the performance of each estimator, we look at four metrics: relative bias (RB), mean square error (MSE), average length of the confidence interval (AL), and the coverage probability (CP) of $\widehat{\mu}_y$. Each measure is calculated as follows:

$$\begin{aligned} RB(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} \frac{\widehat{\mu}_y^{(b)} - \mu}{\mu} \\ MSE(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_y^{(b)} - \mu)^2 \\ AL(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_U^{(b)} - \widehat{\mu}_L^{(b)}) \\ CP(\widehat{\mu}_y) &= \frac{1}{B} \sum_{b=1}^{B} I(\widehat{\mu}_L^{(b)} < \mu < \widehat{\mu}_U^{(b)}). \end{aligned}$$

4.3 Results

The results of the simulation are presented in Table 4.1. We see the relative bias for all three estimators are relatively small, but the variance for the HT estimator is significantly larger than the variances for both calibration estimators, as indicated by their respective mean squared errors. The average length of the confidence interval for calibration estimator number one is also smaller than the HT estimator, but the average length of the confidence interval for calibration estimator two is comparable to the HT estimator. The coverage probabilities for all three confidence intervals are above the 0.90 mark with the coverage probability of calibration estimator number two close to one. These results are reflected in Figure 4.2, which show greater variation in the estimates made by the HT estimator than either calibration estimators.

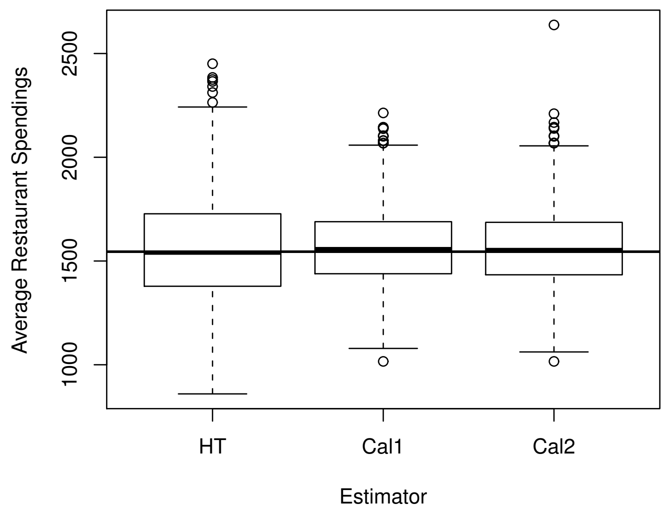

Figure 4.2: Distribution of estimates made by each estimator (the horizontal line indicates the true mean).

5 Extensions and Discussion

We have presented the concept of calibration estimators proposed by Deville and Särndal, which are simply a class of linearly weighted estimators, of which the GREG is a special member. Furthermore, it has been shown that all calibration estimators are asymptotically equivalent to the GREG [DS92].

Consequently, a limitation of the calibration estimator is that it relies on an implicit linear relationship between the study variable, y, and the auxiliary variable $\mathbf{x}$. Thus, if there exists a non-linear relationship

Table 4.1: Performance of estimators from simulation study.

Estimator	RB	MSE	AL	CP
$\widehat{\mu}_{y_{HT}}$	0.011	64669	958.5	0.938
$\widehat{\mu}_{y_{c1}}$	0.014	35615	686.8	0.920
$\widehat{\mu}_{y_{c2}}$	0.011	36269	969.9	0.988

between y and $\mathbf{x}$, the calibration estimator does not perform as well as the HT estimator, that is, if we ignore the auxiliary variable altogether [WS01].

To address this shortcoming, Wu and Sitter developed a model-assisted framework for a model-calibration technique [WS01]. The idea behind model-calibration is to rely on the predicted values, $\widehat{y}_i$, provided by a model ξ, of either linear or non-linear form. As with the original calibration method proposed by Deville and Särndal, the weights w_i are found by minimizing $D(w_i, d_i)$. Instead of using the original calibration constraint in Equation 1.2, however, the minimization is done subject to the constraints

$$\sum_{i \in s} w_i = N \qquad \text{and} \qquad \sum_{i \in s} w_i \widehat{y}_i = \sum_{i=1}^{N} \widehat{y}_i.$$

Wu and Sitter showed that the model-calibration estimator using this technique is more efficient in terms of variance reduction than the simple calibration estimator. It is guaranteed to perform better than the HT estimator, which is sometimes not the case for the original calibration estimator.

This framework paves the way for the use of a variety of models for estimation assistance and generalized the work of Briedt and Opsomer [BO00], who introduced the use of local polynomial regression for estimation. Furthermore, non-parametric models can also be used. Using this framework, Montanari and Ranalli presented a neural network model-calibrated approach [MR05].

Another limitation of the calibration estimator previously mentioned is that the weights can take on negative and/or extremely large values. Deville and Särndal recognized this issue and showed how to restrict the weights to fall within a certain range. Since then, many other methods have been developed to remedy this issue. One such method proposed by Rao and Singh uses a ridge shrinkage method to readjust the weights in an iterative fashion to meet the range restriction [RS97].

6 Acknowledgements

I would like to thank Professor Changbao Wu for introducing me to the topic of calibration estimators and for his guidance in the writing of this paper. I would also like to thank the reviewers for their constructive feedback and advice in helping to improve the clarity of this paper.

References

[BO00] F.J. Breidt and J.D. Opsomer, *Local polynomial regression estimators in survey sampling*, Annals of Statistics **28** (2000), 1026–1053.

[Can10] Statistics Canada, *Survey of household spending 2008*, 2010.

[CSW76] C.M. Cassel, C.E. Särndal, and Wretman, *Some results on generalized difference estimation and generalized estimation for finite populations*, Biometrika **63** (1976), 615–620.

[DS92] J.C. Deville and C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association **87** (1992), 376–382.

[MR05] G.E. Montanari and M.G. Ranalli, *Nonparametric model estimation in survey sampling*, Journal of the American Statistical Association **100** (2005), 1429–1442.

[RS97] J.N.K. Rao and A.C. Sing, *A ridge shrinkage method for range restricted weight calibration in survey sampling*, Proceedings of the Survey Research Methods Section, The American Statistical Association, 1997, pp. 57–64.

[WS01] C. Wu and R.R. Stitter, *A model-calibration approach to using complete auxiliary information from survey data*, Journal of the American Statistical Association **96** (2001), 185–193.