Calibration Estimators by Penalty Function Method

Pius Nderitu Kihara
Department of Statistics and Actuarial Sciences, Technical University of Kenya

Abstract
Estimation of finite population total using calibration has been considered by several authors. A distance measure is minimized subject to some calibration constraints, usually by way of introducing Lagrange equation whose solution gives the design weights used in estimation of population total. Sometimes a solution to the Lagrange constants does not exist. In this paper, we have considered the calibration problem as a nonlinear constrained minimization problem, which we transform to an unconstrained optimization problem using penalty functions. The design weights are obtained iteratively in a numerical manner. We show that the resulting estimator is more accurate than the popular Horvitz Thompson design estimator.

Keywords: calibration, interior penalty function, exterior penalty function

1. Introduction
The notion of calibration was introduced by Deville and Sarndal [1] in the context of using auxiliary information from survey data. Suppose \( U = \{1, 2, \ldots, N\} \) is the set of labels for the finite population. Let \((y_i, x_i)\) be the respective values of the study variable \(y\) and the auxiliary variable \(x\) attached to the \(i^{th}\) unit. If we let \(s = \{1, 2, \ldots, n\}\) be the set of sampled units under a general sampling design \(p\), and let \(\pi_i = p(i \in s)\) be the first order inclusion probabilities, then the conventional calibration estimator for the population total \(\sum y_i\) is defined by \(\hat{Y}_c = \sum w_i y_i\) where \(w_i, s\) are design weights which are as close as possible to \(\pi_i^{-1}\) and are obtained by minimizing a given distance measure between \(w_i, s\) and \(d_i, s\) subject to some constraints. A common distance measure is the chi-square distance measure below.

\[
\Phi = \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{q d_i}
\]

(1)

where \(q, s\) are some constants unrelated to \(d_i, s\). Other distance functions were considered by Deville et al. [2], Singh and Mohl [7] as well as Stukel et al. [8]. Deville and Sarndal [1] considered the calibration constraint

\[
\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{N} x_i
\]

(2)

Minimizing (1) subject to (2) by way of Lagrange equation, they obtained the equation

\[
w_i = d_i = \left\{ \frac{d_i q x_i}{\sum_{i=1}^{n} d_i q x_i^i} \right\} \left\{ \sum_{i=1}^{N} x_i - \sum_{i=1}^{n} d_i x_i^i \right\}
\]

(3)

Wu and Sitter [10] introduced yet another calibration constraint.
\[
\sum_{i=1}^{n} w_i = N
\]  

and minimizing (1) subject to (2) and (4) and by way of Lagrange equation, they obtained

\[
w_i = d_i + (N - \sum_{i \in S} d_i) \left[ \frac{d_i q_i}{\sum_{i \in S} d_i q_i} - \frac{d_i q_i (x_i - \sum_{i \in S} d_i q_i x_i)}{\sum_{i \in S} d_i q_i (x_i - \sum_{i \in S} d_i q_i)^2} \right] + \left( \sum_{i \in S} x_i - \sum_{i \in S} d_i x_i \right) \frac{d_i q_i (x_i - \sum_{i \in S} d_i q_i x_i)}{\sum_{i \in S} d_i q_i (x_i - \sum_{i \in S} d_i q_i)^2}
\]

Obtaining the weights \( w_i \) as derived in (3) and (5) and hence obtaining the estimator \( \hat{y}_i = \sum_{i=1}^{n} w_i y_i \) is quite tedious and may not be feasible in day to day applications. Also, the solution for \( \lambda \) may not always exist in which case Deville and Sarndal [1] recommend that \( \lambda \) be set to 0. Ralf et al [5] considered transforming the calibration problem for general functions \( f \) into a nonlinear equation depending on the Lagrange multiplier \( \lambda \) and since the mapping \( w_i \) was no longer differentiable, they used, semismooth Newton method to solve the resulting equation numerically. We propose use of penalty function to obtain the design weights \( w_i \), a procedure that does not require introduction of Lagrange multipliers.

2. Penalty Function Method

The penalty function methods transform the basic constrained optimization problem into an unconstrained optimization problem. Consider an optimization problem of the form

\[
\min \text{imize } f(X) \text{ subject to }
\]

\[
g_j(X) \leq 0, \quad j = 1, 2, \ldots, m \text{ and } \\
h_j(X) = 0, \quad j = 1, 2, \ldots, p
\]

(6)

By the interior penalty function method (also called barrier method), an unconstrained problem may be constructed as follows.

\[
\phi(X, r_k) = f(X) + \psi_1(r_k, g_j(X)) + \psi_2(r_k, h_j(X))
\]

(7)

where \( \psi_1(r_k, g_j(X)) \) and \( \psi_2(r_k, h_j(X)) \) are penalty functions and which are such that \( \psi_k^i(t = 1, 2) \) is continuous, \( \psi_k^i(r_k, t) \geq 0 \) for all \( r_k \) and \( t \in \mathbb{R}^n \) and \( \psi_k^i(r_k, t) \) is strictly increasing for \( r_k > 0 \) and \( t > 0 \). A common form similar to the one discussed in Rao [6] is given below

\[
\phi(X, r_k) = f(X) - r_k \sum_{j=1}^{m} \frac{1}{g_j(X)} + H(r_k) \sum_{j=1}^{p} l_j^2(X)
\]

(8)
where $H(r_k)$ is some function of the parameter $r_k$ tending to infinity as $r_k$ tends to zero and so that
\[
\sum_{j=1}^p l_j^f(X) \quad \text{also tend to zero. A common choice for value of } q \text{ is } 2. \text{ Also, the function } \phi \text{ will always be greater than } f \text{ since } g_j(X) \text{ is negative for all feasible points } X. \text{ The penalty terms are chosen such that their values will be small at points away from the constraint boundaries and will tend to infinity as the constraint boundaries are approached. Hence the value of } \phi \text{ will also blow up as the constraint boundaries are approached. Frank and Jorge [3] have discussed flexible ways of choosing the penalty. In an iterative process, the unconstrained minimization of } \phi \text{ is started from any feasible solution for the inequality constraint but not necessarily so for the equality constraints. The subsequent points generated will always lie within the feasible region since the constraint boundaries act as barriers during the minimization process. The rationale of the penalty terms as described by Ozgur [4] is that if the constraint is violated, that means $g_j(X) > 0 \text{ or } f_j(X) = 0 \text{, a big term will be added to } \phi \text{ function such that the solution is pushed back towards the feasible region. In the minimization of } \phi, \text{ for the solution to be the global minimum, we must have that } f(X) \cdot g_j(X) = 1,2,..., m, \text{ and } \sum_{j=1}^p l_j^f(X) \text{ being convex and we must also have one of the functions } f(X) \cdot g_j(X) = 1,2,..., m \text{ and } \sum_{j=1}^p l_j^f(X) \text{ being strictly convex. See Rao [6].}
\]

Using the exterior penalty function method, a solution to the constrained problem (6) would be given by
\[
\phi(X, r_k) = f(X) + r_k \sum_{j=1}^m \left(g_j(X)\right)^q + H(r_k) \sum_{j=1}^p l_j^f(X) \tag{9}
\]
where $\left(g_j(X)\right) = \max(g_j(X), 0)$. Also, as $k \to \infty, \ r_k \to \infty \text{ and } H(r_k) \to \infty$. For exterior penalty function method, in the iterative minimization of $\phi$, the starting point $X$ does not have to be feasible. Looking at the equations (8) and (9), we see that, when the optimization problem has only the equality constraints, both interior and exterior penalty functions yield a function of the form
\[
\phi(X, r_k) = f(X) + H(r_k) \sum_{j=1}^p l_j^f(X) \tag{10}
\]
Setting $H(r_k) = r_k$, where $r_k \to \infty \text{ as } k \to \infty$, and setting $q = 2$ we have from (10)
\[
\phi(X, r_k) = f(X) + r_k \sum_{j=1}^p l_j^f(X) \tag{11}
\]


Let there be a population of size $N$ for our variable of interest $y$ from which we draw a sample of size $n$. Let the auxiliary value $x_i$ be available for every element of the population of variable $y$. We wish to estimate the population total $y = \sum_{i=1}^N y_i$ from a sample of size $n$ and incorporating the auxiliary information present. To obtain design weights, we reduce the chi-square distance measure (1) subject to the constraints (2) considered by Deville and Sarndal [1]. Using the penalty function method we obtain the penalty function
\[
\phi_1(w, r_k, x) = \sum_{i=1}^n \left(\frac{w_i - d_i}{q d_i}\right)^2 + r_k \left[\sum_{i=1}^n w_i x_i - \sum_{i=1}^N x_i\right]^2 \tag{12}
\]
where $r_k$ is some penalty. We need to find the weights $w_i$ that minimize the penalty function (12) above.

Differentiating (12) partially with respect to $w_i$ we have
\[
\phi_1^i(w, r_k, x) = \frac{2(w_i - d_i)}{q d_i} + 2r_k x_i \left[\sum_{j=1}^n w_j x_j - \sum_{j=1}^N x_j\right] \tag{13}
\]
We equate (13) to zero and solve for \( w_i \) to obtain

\[
d_i - r_k x_i q_i d_i \left( \sum_{j=1}^{n} w_j x_j - \sum_{j=1}^{N} x_j \right) = \frac{w_i}{1 + r_k (x_i^2 q_i d_i)}
\]

We have the following estimator of population total

\[
\hat{y}_{n} = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} \frac{y_i d_i}{1 + r_k (x_i^2 q_i d_i)} - \sum_{i=1}^{n} \frac{r_k x_i q_i d_i y_i}{1 + r_k (x_i^2 q_i d_i)} \left( \sum_{j=1}^{n} w_j x_j - \sum_{j=1}^{N} x_j \right)
\]  

Minimizing (1) subject to both (2) and (4) as considered by Wu and Sitter [10], we have the penalty function

\[
\phi_k(w, r_k, x) = \sum_{i=1}^{n} \left( \frac{w_i - d_i}{q_i d_i} \right)^2 + r_k \left[ \sum_{i=1}^{n} w_j x_j - \sum_{j=1}^{N} x_j \right]^2 + r_k \left[ \sum_{i=1}^{n} w_i - N \right]^2
\]

Differentiating (16) partially with respect to \( w_i \) we have

\[
\phi_k^i(w, r_k, x) = \frac{2(w_i - d_i)}{q_i d_i} + 2r_k x_i \left[ \sum_{j=1}^{n} w_j x_j - \sum_{j=1}^{N} x_j \right] + 2r_k \left[ \sum_{i=1}^{n} w_i - N \right]
\]

Equating (17) to zero and solving for \( w_i \) we have

\[
w_i = \frac{d_i - r_k q_i d_i \left( \sum_{j=1}^{n} w_j (x_j x_j + 1) - \sum_{j=1}^{N} (x_i x_j - 1) \right)}{1 + r_k (x_i^2 + 1) q_i d_i}
\]

We therefore have the following estimator of population total
\[ \hat{y}_{12} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} 1 + r_i (x_i^2 + 1) q_i d_i} - \frac{r_i q_i d_i y_i}{(x_i + 1) q_i d_i} \left( \sum_{j=1}^{n} w_j (x_i x_j + 1) - \sum_{j=1}^{n} (x_i x_j - 1) \right) \]

(19)

The beauty with this approach is that to obtain the weights \( w_i \) (\( i = 1, 2, ..., n \)) we solve the penalty functions (12) and (16) as unconstrained minimization problems in which case we only require to start with some initial guess for \( w_i \) and \( r_i \) and then iteratively improve on the initial values until we have optimal values. Since the constraints (2) and (4) are equality constraints, we need not start with a feasible guess for \( w_i \). We appeal to Newton method of unconstrained optimization. See Rao [6].

Let \( W = \{ w_1, w_2, ..., w_n \} \) be the set of the weights. We need to obtain \( W^- \) such that

\[ g(W^-) = [\phi'(w_1, r_1, x), ..., \phi'(w_n, r_n, x)]^T = 0 \]

(20)

We first start with some initial approximation \( W_i \) of \( W^- \) so that \( W^- = W_i + Z \). The Taylor’s series expansion of \( g(W^-) \) gives

\[ g(W^-) = g(W_i + Z) = g(W_i) + J_{W_i} Z + ... \]

(21)

By neglecting the higher order terms in (21) and setting \( g(W^-) = 0 \) we obtain

\[ g(W_i) + J_{W_i} Z = 0 \]

(22)

Where \( J_{W_i} \) is the matrix of second derivatives evaluated at \( W_i \). In general, when we consider the constraint (2) alone, then \( J \) is a \( n \times n \) matrix with \( i = 1, 2, ..., n \) rows and \( j = 1, 2, ..., n \) columns. It has diagonal elements \( \frac{2}{q_i d_i} + 2 r_i x_i^2 \) and elements \( 2 r_i x_i x_j \) elsewhere. If we consider both constraints (2) and (4), then \( J \) has diagonal elements \( \frac{2}{q_i d_i} + 2 r_i (x_i^2 + 1) \) and elements \( 2 r_i (x_i x_j + 1) \) elsewhere. If \( J_{W_i} \) is nonsingular, then, from the set of linear equations (22) we have for vector \( Z \)

\[ Z = J_{W_i}^+ g(W_i) \]

(23)

The following iterative procedure is used to find the improved approximations of \( W^- \).
The sequence of the points \( W_1, W_2, \ldots, W_{r+1} \) eventually converges to the actual solution \( W^* \). Since our penalty functions (12) and (16) are quadratic, we find the minimum in a single step using equation (24) since the Taylor’s series expansion is exact.

Now, if we let \( W^*_r \) be the minimum of \( W^* \) obtained for a particular penalty \( r_k \), we obtain a sequence of minimum points \( W^*_r, W^*_{r+1}, \ldots, W^*_{k+1} \) for the penalties \( r_1, r_2, \ldots, r_{k+1} \) until \( W^*_{k+1} = W^*_{k+1} \) or \( \phi(w, r_k, x) = \phi(w, r_{k+1}, x) \) for some specified accuracy level. The accuracy level may for example be, to certain decimal points or significance level. The penalty values are set such that the starting point \( r_1 > 0 \) and \( r_{k+1} = c r_k \), where \( c > 1 \). We can now generalize our estimator for the population total as

\[
\hat{y}_v = \sum_{i=1}^{n} w_i y_i = W^* Y_v, \quad v = 1, 2
\]

where \( Y_v = (y_1, y_2, \ldots, y_n) \) is the sample from the population of \( y \) and \( v = 1 \) if \( w_i^* \)'s are obtained as defined in (14) and \( v = 2 \) if \( w_i^* \)'s are obtained as defined in (18).

4. Empirical Analysis

Using R program, we simulated a population of independent and identically distributed variable \( x \) using uniform \((0, 1)\). Using \( x \) as the auxiliary variable we generated the populations of size 300 for random variable \( y \) as a linear function \( y = 2 + 5x \) and quadratic function \( y = (2 + 5x)^2 \). For both populations, the estimators exhibited same properties. We will therefore report the results for the linear function \( y = 2 + 5x \). For each of different sample sizes \( n \), 5 samples were generated. Our initial penalty constant was set at \( r_0 = 0.00010 \). The convergence criteria considered was \( W^*_r = W^*_r \) and \( \phi(w, r_k, x) = \phi(w, r_{k+1}, x) \) to six decimal places. In section 4.1, we report on the performance of estimator \( y_{11} \) and compare its performance with that of Horvitz Thompson estimator \( y_{10} = \sum_{i=1}^{n} y_i d_i \) discussed in Thompson [9], while in section 4.2, we report on the results for estimator \( y_{12} \) and again compare with Horvitz Thompson estimator.

4.1 Results for Estimator \( y_{11} \)

We let \( y_i = \sum_{i=1}^{N} y_i \) be the actual population total, \( r_k \) be the penalty parameter, and \( y_r - y_{11} \) and \( y_r - y_{10} \) be the errors in the estimation.
Looking at table (1), we see that the estimators $y_{t1}$ and $y_{ht}$ have almost equal error margins, but consistently, $y_{t1}$ has a smaller error margin. For all the samples, convergence is achieved at the same penalty value of 0.00010 and which was the initial penalty value. For different sample sizes, we observed that the penalty value ranged between 0.0001 and 0.0013 with no particular pattern that could be attributed to the sample size. In most of the cases however, the penalty was 0.00010.

![Fig 1: Variance for Estimator $y_{t1}$](image-url)
In Fig (1) and Fig (2), the variances for $y_{t1}$ and $y_{ht}$ have a similar pattern. As the sample size increases, the variance decreases. From Fig (3), the ratio $\frac{\text{var}(y_{t1})}{\text{var}(y_{ht})}$ settles almost to a constant as the sample size increases. The constant is found to be about 0.97, which indicates that $y_{t1}$ has a smaller variance than $y_{ht}$, and which is consistent with the smaller error margin for $y_{t1}$ as seen in table (1).
4.2 Results for Estimator $y_{r_2}$

<table>
<thead>
<tr>
<th>sample number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample size n</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$y_t$</td>
<td>1384.49498</td>
<td>1384.49498</td>
<td>1384.49498</td>
<td>1384.49498</td>
<td>1384.49498</td>
</tr>
<tr>
<td>$y_{r_2}$</td>
<td>1400.01439</td>
<td>1406.98567</td>
<td>1398.03903</td>
<td>1321.13222</td>
<td>1330.37056</td>
</tr>
<tr>
<td>$y_{ht}$</td>
<td>1400.27413</td>
<td>1407.37208</td>
<td>1398.26738</td>
<td>1320.20986</td>
<td>1329.55309</td>
</tr>
<tr>
<td>$y_t - y_{r_2}$</td>
<td>-15.51940</td>
<td>-22.49068</td>
<td>-13.54405</td>
<td>63.36276</td>
<td>54.12442</td>
</tr>
<tr>
<td>$y_t - y_{ht}$</td>
<td>-15.77915</td>
<td>-22.87709</td>
<td>-13.77240</td>
<td>64.28512</td>
<td>54.94189</td>
</tr>
<tr>
<td>$r_k$</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
<td>0.00010</td>
</tr>
</tbody>
</table>

In table (2) $y_t - y_{r_2}$ and $y_t - y_{ht}$ are the errors in the estimation. From the table, $y_{r_2}$ and $y_{ht}$ error margins are quite close, but with $y_{r_2}$ consistently, having the smaller error margin. Also the penalty value is 0.00010 for all the samples.

Fig 4: Variance for Estimator $y_{r_2}$
Fig 5: Variance for Horvitz Thompson Estimator $y_{ht}$

Fig (4) and Fig (5), show similar patterns for the variances of $y_{t_2}$ and $y_{ht}$. As the sample size increases, the variances are decreasing. From Fig (6), the ratio $\frac{\text{var}(y_{t_2})}{\text{var}(y_{ht})}$ tends to a constant, estimated to about 0.97 and which indicates that $y_{t_2}$ has a smaller variance than $y_{ht}$.

Fig 6: Variance Ratio $\frac{\text{var}(y_{t_2})}{\text{var}(y_{ht})}$

5. Conclusion

We conclude that both estimators $y_{t_1}$ and $y_{t_2}$ are more accurate than the Horvitz Thompson design estimator $y_{ht}$ since they both have smaller margin of errors and smaller variance than $y_{ht}$. From the variance ratios $\frac{\text{var}(y_{t_1})}{\text{var}(y_{ht})}$ and $\frac{\text{var}(y_{t_2})}{\text{var}(y_{ht})}$ both of which are about 0.97, we conclude that $\text{var}(y_{t_1})$ and $\text{var}(y_{t_2})$ are not significantly different and that estimators $y_{t_1}$ and $y_{t_2}$ are not different in terms of the
We conclude that the estimators $y_{1i}$ and $y_{2i}$ are consistent in the sense that as the sample size increases, their variances tend to zero.

References


