Calibration Adjustments When Not All Targets Can Be Met

Matthew Fetter, James Gentle, and Charles Perry
National Agricultural Statistics Service

Abstract

Sometimes benchmark constraints in a calibration problem cannot be met if there are range restrictions on the calibration weights. There are various approaches to this problem that involve either allowing the benchmark constraints to be adjusted within a specified tolerance, or to determine a minimal linear adjustment of the benchmark constraints. In this paper we propose an optimization problem that explicitly incorporates into the objective function a measure of the amount by which the benchmark constraints are missed.

Keywords: Data editing; Survey sampling; Benchmark constraints; Range restrictions; Optimization methods; Agricultural statistics.

1. Introduction

The basic calibration problem in survey sampling is to adjust data so that certain totals or other summary statistics match benchmark values. The data may suffer from incorrect coverage by the frame, from nonresponse, or from other nonsampling errors. The purpose of the calibration is usually to improve the accuracy of other quantities computed from the given data. The benchmark targets for the "calibration variables" are usually obtained from other sources that are believed to be more accurate for these specific values. Calibration is performed by adjusting weights of the individual records in the given data set. This adjustment of the weights is done in such a way that the new weights are close to the previous weights, by some measure of the distance of the change.

To state this more precisely, but without making assumptions about the nature of the sampling or of the analysis, let $[X|Y]$ be the $n \times m$ data matrix partitioned into an $n \times p$ matrix X containing observations on the calibration variables and an $n \times (m-p)$ matrix Y containing observations on the other variables. We assume all elements of X are real (that is, no missing values). Let d be the n-vector of weights or "expansion factors". We assume $d > 0$. We assume that for a quantity of interest, say, t_X, a p-vector of population totals for the calibration variables, a good estimate (in some sense) is

$$\widehat{t}_X = X^{\mathrm{T}} d. \tag{1}$$

Given a vector of targets for the calibration variables T_X, the requirements of calibration are expressed in the system of *calibration equations*,

$$T_X = X^{\mathrm{T}} w, \tag{2}$$

for some vector of weights w. These are the "benchmark constraints" (BC) to be satisfied. If X is of full column rank, it is a relatively simple matter to find a w so that the BC are satisfied. Generally, we seek a w that does not differ much from d.

There may also be "range restrictions" (RR) on the elements of w; for example, they may be required to be positive.

The problem becomes well-posed with the requirement that the extent of the adjustment, as measured by the differences in w and d, is minimized, subject of course to a formal definition of the "extent of the adjustment". Informally, a statement of the calibration problem is

$$\begin{array}{ll} \underset{w}{\text{minimize}} & \text{differences in } w \text{ and } d \\ \text{subject to} & \left\{ \begin{array}{l} \text{BC} \\ \text{RR} \end{array} \right. \end{array} \tag{3}$$

This problem was considered in detail by Deville and Särndal (1992), who proposed various measures of the differences in w and d. Singh and Mohl (1996) discussed the problem further, described computational approaches, and reported empirical results from using various measures and constraints.

If the minimum of the differences in w and d (according to some measure) occurs within the BC and the RR, the problem (3) has a simple solution and the approach is clearly appropriate.

In this paper we consider situations in which either the BC or the RR (or both) is not satisfied at the minimum of the differences (that is, when the extent of the adjustment is minimized). We also exhibit situations in which the BC and/or the RR *cannot* be satisfied. We formulate an optimization problem that is more appropriate than problem (3) in those cases.

We represent the (scalar) difference between w and d by $\delta(w, d)$. The distance measure $\delta(w, d)$ is nonnegative, but it may not be a metric. (It may not be symmetric in w and d.) Also, the summary function $\delta(w, d)$ may not give the same weights to all elements of w and d. One distance measure considered by Deville and Särndal (1992) is

$$\delta_Q(w, d) = (w - d)^{\mathrm{T}} (DQ)^{-1} (w - d), \tag{4}$$

where $D = \mathrm{diag}(d)$, the $n \times n$ diagonal matrix whose entries are the elements of d, and Q is an $n \times n$ diagonal matrix with positive entries, perhaps all equal to 1, or perhaps scaled so as to allow the weights of some records to change more than others. Minimizing the expression in equation (4) subject to $T_X = X^{\mathrm{T}} w$ is a linearly equality constrained least squares problem, and if a solution to $X^{\mathrm{T}} DQX\lambda = T_X - \widehat{t}_X$ exists, it is a solution to the minimization problem.

1.1 Conditions for Satisfying the Calibration Equations

A necessary and sufficient condition that the calibration equations can be satisfied by some value of w is that the system $T_X = X^{\mathrm{T}} w$ is consistent, that is, that $\mathrm{rank}(X^{\mathrm{T}}) = \mathrm{rank}(X^{\mathrm{T}}|T_X)$.

If X is of full column rank this condition is always met.

A common way that a data matrix of less than full rank arises is in the case of categorical variables that have been coded as binary variables; for example, race, instead of being a single value with, say, three possible values is coded as three 0-1 variables.

In practice, a non-full-rank data matrix is rarely a problem. When it occurs, the statistician is usually aware of it and knows that any reasonable vector of targets must satisfy the same linear relationships as the data; therefore the calibration equations are consistent and a solution the exists. In some methods of calibration a system of equations with the Gramian $X^{\mathrm{T}} X$ coefficient matrix must be solved, and so there may be practical problems with a non-full-rank data matrix. The `solve` function in both R/S-Plus and SAS IML, for example, requires nonsingular coefficient matrices. (This lack of software is perhaps understandable, because the solution to a non-full-rank, consistent system is an infinite set of vectors.)

We will not address the problem of inconsistent calibration equations. If the targets cannot be met, there is likely some misspecification of the the targets or else there were major nonsampling errors. Although we will not consider this problem separately, if indeed it is desired to address this problem with the given data and targets, then the reformulation of Section 4. may be appropriate.

1.2 Conditions for Satisfying Both the Calibration Equations and the Range Restrictions

There are p calibration equations in n unknowns, so typically whether or not the data matrix is of full (column) rank, there are multiple vectors w that satisfy $T_X = X^{\mathrm{T}} w$.

A problem that may arise in applications is that none of the solutions to the calibration equations satisfy the range restrictions. If $\mathcal{W}$ is the set of all solutions of the calibration equations, and $\mathcal{R}$ is the set of all n-vectors that satisfy the range restrictions, then obviously

$$\mathcal{W} \cap \mathcal{R} \neq \emptyset \tag{5}$$

is a necessary and sufficient condition that both sets of constraints and restrictions be satisfied simultaneously.

In principle, this is an easy condition to check by determining whether the space of solutions to $T_X = X^{\mathrm{T}} w$ contains elements of $\mathcal{R}$. In practice, however, it is not always easy to determine whether two spaces intersect. Furthermore, because of the dearth of software that computes general solutions to $T_X = X^{\mathrm{T}} w$, this is not always a trivial problem.

2. Calibration When Not All Targets Can Be Met

Without range restrictions, the calibration problem (3) above can be formulated as an equality-constrained minimization problem:

$$\begin{aligned} \min_w \quad & \delta(w, d) \\ \text{s.t.} \quad & T_X = X^{\mathrm{T}} w. \end{aligned} \tag{6}$$

2.1 Meeting Calibration Targets without Range Restrictions

A common distance function is $\delta_Q(w, d)$ given in equation (4). The optimization problem with this distance function is a weighted linear least squares problem with linear equality constraints. The solution, which is easy to obtain by standard Lagrangian methods (see Deville and Särndal, 1992), is

$$w = d + DQX\lambda, \tag{7}$$

where λ is such that

$$X^{\mathrm{T}} DQX\lambda = T_X - \widehat{t}_X, \tag{8}$$

if a solution to these equations exists. It is not necessary that $X^{\mathrm{T}}DQ$ be of full rank.

If X is full column rank (and DQ is full rank), a solution exists, and it is given by equation (7). If X is not of full column rank, that is, if some columns of X can be expressed as linear combinations of others, a solution exists if and only if those same relationships exist within the targets. If the targets are inconsistent with the linear dependencies in $X^{\mathrm{T}}DQ$, it is likely that there is something wrong with the targets, and the statistician should examine them more closely.

If there are no restrictions on the weights, if the calibration equations are consistent, a solution exists, and it is given by equation (7). Solutions for other distance functions, $\delta(w,d)$, may be somewhat harder to obtain, but the existence of solutions is determined by the consistency of the calibration equations.

2.2 Meeting Calibration Targets with Range Restrictions

The solution to the calibration problem (6) may have either negative or very large elements in w, and neither of these cases may be acceptable. We therefore often impose range restrictions on the weights.

We reformulate the minimization problem as the weighted least squares problem with both linear equality constraints and other constraints,

$$\begin{aligned} \min_{w} \quad & \delta(w,d) \\ \text{s.t.} \quad & T_X = X^{\mathrm{T}}w \\ & w \in \mathcal{R}, \end{aligned} \tag{9}$$

where $\mathcal{R}$ is some set of n-vectors, perhaps $\Re^n_+$, the positive real numbers. More often, $\mathcal{R}$ is a set of positive numbers less than some given value. The set may even be restricted to positive integers less than some value.

The restrictions on the weights is the main practical problem in implementing calibration methodology.

The simple Lagrange multiplier solution of problem (6) no longer is a solution. Both Deville and Särndal (1992) and Singh and Mohl (1996) suggest iterating to the solution for problem (9) by sequential solutions for problem (6) with d replaced in the j^{th} iteration by $w_*^{(j-1)}$, the solution at the $(j-1)^{\text{th}}$ iteration and beginning with $w_*^{(0)} = d$. This fails, of course, if a solution to the equations (8) does not exist, but more perniciously it fails if $\mathcal{W} \cap \mathcal{R} = \emptyset$.

This approach may be slow to converge even if there is a solution. It is perhaps a more serious problem that the existence of a solution is difficult to establish.

If there is no solution, the problem must be changed in some reasonable way. Rao and Singh (1997) propose a "ridge-shrinkage" method to relax the benchmark constraints within a prespecified tolerance so that the range restrictions can be satisfied. Chen, Sitter, and Wu (2002) use bisection to find the smallest adjustments to the targets in the benchmark constraints that will allow the range restrictions to be met. The bisection occurs between the solution to the problem without the range restrictions and a solution to the problem without the benchmark constraints.

One way that depends on a relaxation of the BC within a prespecified tolerance is to replace the point targets with intervals; that is, the calibration equations are replaced by the *calibration intervals*,

$$T_X - L_T \ \le\ X^{\mathrm{T}}w \ \le\ T_X + U_T, \tag{10}$$

where L_T and U_T are p-vectors with nonnegative elements. This has the effect of enlarging $\mathcal{W}$, thus making it less likely that $\mathcal{W} \cap \mathcal{R} = \emptyset$. One can show by toy counterexamples that so long as the intervals on the targets and on the weights are finite, there are cases in which no solution exists.

This approach was used by the National Agricultural Statistics Service (NASS) of USDA in calibration of data arising from the 2002 Census of Agriculture. Even by relaxing the calibration targets to reasonable prespecified intervals, however, there were situations in which solutions could not be found.

3. Measuring the Extent to Which Targets Are Missed

If the calibration targets cannot be met, we need a scalar measure, say $\phi(w,T_X)$, of the extent to which they are missed whether the targets are points or intervals.

An obvious candidate for $\phi(w,T_X)$ is the sum of squares of differences. The sum of absolute differences or any other norm applied to $(T_X - X^{\mathrm{T}}w)$ would also be obvious candidates. Any such $\phi(w,T_X)$ would of course be a metric, but there is no reason to require that it be a metric. Just as with $\delta(w,d)$ we may scale the differences based on the magnitude T_X. It may also be appropriate to include in $\phi(w,T_X)$ the variances and covariances associated with each observational variable and/or the variances and covariances associated with each element of T_X. If T_X comes from a previous survey, it may be possible to have a good measure of its variance.

Let V_X and V_T be the $p \times p$ variance-covariance matrices, respectively, of the calibration variables and of the calibration targets (or realized consistent positive definite estimators of them), and P be a $p \times p$ diagonal matrix with positive entries, perhaps all equal to 1, or perhaps scaled so as to attach differential weights of importance to the various calibration variables. A reasonable measure of how much the targets are missed is

$$\begin{aligned}&\phi_Q(w, T_X) = \\ &(T_X - X^{\mathrm{T}}w)^{\mathrm{T}}\left((V_X + V_T)P\right)^{-1}(T_X - X^{\mathrm{T}}w).\end{aligned} \tag{11}$$

If the calibration equations are replaced by calibration intervals, the measure of how much the targets are missed would be of similar form, but would be based on how far outside of the interval the computed values lie:

$$\phi_Q(w, T_X) = e^{\mathrm{T}}\left((V_X + V_T)P\right)^{-1} e, \tag{12}$$

where $e_j = \min(0,\ \max((T_X)_i - (L_T)_i - y_i^{\mathrm{T}}w,\ y_i^{\mathrm{T}}w - (T_X)_i - (U_T)_i))$ with y_j the j^{th} column of X.

The distances in equations (11) and (13) are formulated in terms of the extent to which each target variable contributes to the measure of the overall deviation from the targets. More precisely, in equation (11), the contribution of the j^{th} calibration variable to the measure is the j^{th} diagonal element of the matrix

$$(T_X - X^{\mathrm{T}}w)\,(T_X - X^{\mathrm{T}}w)^{\mathrm{T}}\left((V_X + V_T)P\right)^{-1}.$$

The sum of the p measures for the individual target variables is $\phi_Q(w, T_X)$.

We can also formulate these distances in terms of squares of scaled differences of a given target and the corresponding computed value. Let $V^{-\frac{1}{2}}$ be such that $(V^{-\frac{1}{2}})^2 = ((V_X + V_T)P)^{-1}$. (Such a matrix exists because the matrix $((V_X + V_T)P)^{-1}$ is positive definite. Furthermore such a matrix is symmetric.) Now write $\phi_Q(w, T_X)$ in equation (11) as

$$\operatorname{trace}\left(\left(V^{-\frac{1}{2}}(T_X - X^{\mathrm{T}}w)\right)\,\left(V^{-\frac{1}{2}}(T_X - X^{\mathrm{T}}w)\right)^{\mathrm{T}}\right).$$

The j^{th} diagonal of this matrix is a measure, say r_j, of the extent to which the j^{th} target is missed. It is square of the j^{th} element in the vector that forms the outer product, that is,

$$r_j = \left(V^{-\frac{1}{2}}(T_X - X^{\mathrm{T}}w)\right)_j^2. \tag{13}$$

The sum of the p measures for the individual targets is $\phi_Q(w, T_X)$.

4. A Reformulation of the Problem

If no solution to the problem exists, because of either of the situations we have described, we must modify the problem in a reasonable way (or just quit!). We are faced with the situation that the targets cannot be met while satisfying range restrictions. Note that this nonexistence of a solution has nothing to do with the objective function $\delta(w, d)$.

If both $T_X = X^{\mathrm{T}}w$ and $w \in \mathcal{R}$ cannot be satisfied simultaneously, a reasonable approach might be to relax the calibration equations, and formulate an objective function that measures how much we miss the targets. Let $\phi(w, T_X)$ be such a function. We still have the objective of making a minimum adjustment to the weights, so we might form an objective function that is a weighted average of the two separate objective functions. An appropriate optimization problem would be of the form

$$\begin{array}{ll}\displaystyle\min_{w} & \alpha\delta(w, d) + (1-\alpha)\phi(w, T_X) \\ \text{s.t.} & \qquad w \in \mathcal{R},\end{array} \tag{14}$$

where α is some number between 0 and 1.

An important difference in this problem and the optimization problem (9) is that this problem always has a solution under the usual conditions that δ and ϕ are continuous, positive functions, and $\mathcal{R}$ is compact.

A possible form of $\phi(w, T_X)$, analogous to $\delta_Q(w, d)$ in equation (4), is given by equation (11). With this measure of the amount that the targets are missed, and $\delta_Q(w, d)$ as the measure of how much the weights are changed, the objective function to be minimized in problem (14) is

$$\begin{aligned}&\alpha(w-d)^{\mathrm{T}}(DQ)^{-1}(w-d) \quad + \\ &(1-\alpha)(T_X - X^{\mathrm{T}}w)^{\mathrm{T}}\left((V_X + V_T)P\right)^{-1}(T_X - X^{\mathrm{T}}w)\end{aligned}$$

or

$$\begin{aligned}&w^{\mathrm{T}}\left(\alpha(DQ)^{-1} + (1-\alpha)X((V_X + V_T)P)^{-1})X^{\mathrm{T}}\right)w \\ &-2w^{\mathrm{T}}\left(\alpha(DQ)^{-1}d + (1-\alpha)X((V_X + V_T)P)^{-1}\right)T_X \\ &+\text{constant}.\end{aligned} \tag{15}$$

At a minimum within $\Re_+^n$, we have

$$\begin{gathered}\left((\alpha(DQ)^{-1} + (1-\alpha)X((V_X + V_T)P)^{-1})X^{\mathrm{T}}\right)w \\ = \\ (\alpha(DQ)^{-1}d + (1-\alpha)X((V_X + V_T)P)^{-1})T_X.\end{gathered} \tag{16}$$

The second derivative is nonnegative definite; insuring that a solution to equation (16) is indeed a minimum.

The second component of the coefficient matrix in equation (16) is singular. (This component is exactly the same matrix we would have in the simple problem of determining weights so as to satisfy the calibration equations without any constraints at all.) The whole coefficient matrix may or may not be singular. If the matrix is singular, any solution to equation (16) is a solution to the unconstrained minimization problem.

A minimum within $\Re^n_+$ may or may not be a minimum within $\mathcal{R}$, however, and some iterative method must be used to satisfy the range restrictions.

In calibration problems, $\mathcal{R}$ is generally a product of intervals, usually equal intervals. Indeterminancy in equation (16) may help to insure that some unrestricted minimum satisfies the range restrictions.

There are various possible approaches for obtaining a solution that satisfies interval range restrictions. An exact solution solution may require computations over all combinations of weights at boundary values.

5. Computational Issues and Further Discussion

While the optimization problem (14) always has a solution, it may not be easy to obtain. The problem, of course, is not the objective function, but rather the restrictions on the weights.

It is interesting to observe that without the restrictions on the weights, the problem of determining the calibration weights was a problem in p dimensions. With restrictions on the weights the problem is in n dimensions. This is because each one of the n weights must be within a given range.

The iterative scheme for problem (9) is likely to be even less effective for problem (14) because one component of the objective function does not pull the iterations toward convergence.

The difficulties in solving the problem with the range restrictions suggests a decomposition of the problem into subproblems with the hope that the sequence of solutions leads to a solution to the overall problem. For example the r_j in equation (13) may be used to suggest individual targets or groups of targets to begin to address. There is, of course, no guarantee that a sequential process will be any simpler, because the restriction $w \in \mathcal{R}$ is still present, and it is this restriction that presents the computational challenge.

A stochastic method such as simulated annealing or a genetic algorithm may be used. The steps in the stochastic algorithm would be constrained to lie within $\mathcal{R}$. An efficient stochastic algorithm would need to have a good method for moving from $w^{(k)}$ to a good candidate point for $w^{(k+1)}$.

When both components of the objective function are sums of squares, the problem is a quadratic programming problem, and standard software is available for its solution. Although the formulation and solution of the problem is straightforward, our preliminary studies indicate that the method is not sufficiently fast enough to address the problems with multiple tens of thousands of records, as in the NASS Census of Agriculture. We are currently continuing experimenting with a quadratic programming approach for various subproblems.

References

Chen, J., R. R. Sitter, C. Wu (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika* **89**, 230–237.

Deville, Jean-Claude, and Carl-Erik Särndal (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**, 376–382.

Rao, J. N. K., and A. C. Singh (1997), A ridge-shrinkage method for range restricted weight calibration in survey sampling, *Proceedings of the Section on Survey Sampling*, American Statistical Association 57–65.

Singh, A. C., and C. A. Mohl (1996), Understanding calibration estimators in survey sampling, *Survey Methodology* **22**, 107–115.