Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Relaxed calibration of survey weights

by Nicholas T. Longford

Release date: December 20, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2024

Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 261-285
Statistics Canada, Catalogue No. 12-001-X

Relaxed calibration of survey weights

Nicholas T. Longford[1]

Abstract

Population surveys are nowadays rarely analysed in isolation from any auxiliary information, often in the form of population counts, totals and other summaries. Calibration, or benchmarking, by which the weighted sample totals of auxiliary variables are matched to their (known) population totals, is widely applied. Methods for adjusting the weights to satisfy these constraints involve iterative procedures with unknown finite-sample properties. We develop an alternative method in which the weights are calibrated by minimising a quadratic function, requiring no iterations and yielding a unique solution. The relative priority of each constraint is represented by a tuning parameter. The properties of the weights and of the calibration estimator, as functions of these parameters, are explored analytically and by simulations. A connection of the proposed method with ridge calibration is established.

Key Words: Auxiliary information; Benchmarking; Priority; Ridge regression; Survey sampling; Weight adjustment.

1. Introduction

Calibration, or benchmarking, is generally regarded as an indispensable accompaniment of estimation of population summaries in large-scale surveys that are conducted in an environment in which other data sources provide auxiliary information. Such information has a potential to make estimation more efficient or for it to acquire some other valuable properties or attributes. Calibration has an important role in compensating for imperfections in the sampling design and its implementation, such as deficiencies in the sampling frame and nonresponse.

Calibration has an extensive literature; Deville and Särndal (1992) is widely regarded as a landmark, reinforced by Lundström and Särndal (1999) for its application in the context of modern official statistics. Estevao and Särndal (2006) and Särndal (2007) review subsequent developments. For more recent literature surveys, see Kim and Park (2010), Brick (2013), Wu and Lu (2016) and Lohr and Raghunathan (2017). The monograph of Tillé (2020) contains a comprehensive treatment of the subject. Devaud and Tillé (2019) is an appraisal of the impact of Deville and Särndal (1992) on survey sampling, and official statistics in particular. Davies (2018) reviews a wide range of methods for calibration. In his terminology, hard calibration refers to optimisation while satisfying a set of constraints with no scope for any discrepancy. We prefer soft calibration in which a compromise is sought among the constraints and objectives of weight adjustment and subsequent estimation.

Model-based and model-assisted approaches have found a fertile ground in survey sampling in general, and weight calibration in particular. Ordinary regression and its various generalizations have been widely applied; see Haziza and Beaumont (2017) and references therein. One such approach is motivated by ridge

1. Nicholas T. Longford, SNTL Statistics Research and Consulting, London, UK, 3 Badgers Walk, Whyteleafe CR3 0AS, Surrey, United Kingdom. E-mail: sntlnick@sntl.co.uk.

regression (Hoerl and Kennard, 1970), in which the imperative of satisfying the benchmark constraints is moderated to promote stability of the solution (calibration weights), to avoid large adjustments, unacceptable and extreme values, and reduce the dispersion of the weights. Beaumont and Bocci (2008) relate the original proposal by Chambers (1996) to the likelihood-based approach of Chen, Sitter and Wu (2002).

Calibration is especially challenging when a lot of auxiliary information is available, and many population totals are to be matched. Cardot, Goga and Shehzad (2017) and Vera, Sánchez Zuleta and Rueda (2023) address this problem by projecting the auxiliary data onto a manageable subspace for which established methods can be applied. See also Dagdoug, Goga and Haziza (2023) for a model-assisted approach.

We introduce a method of benchmarking in which the goal of matching the calibration constraints is relaxed to reducing the discrepancies between the sample summaries and their targets in line with specified priorities. The algorithm we develop caters for the same constituency of problems as the established methods of calibration but permits integration of the analyst's (or their client's) priorities in a flexible and transparent manner. The algorithm requires no model-related assumptions but is closely related to ridge calibration, in which a model is implied (Chambers, 1996; Rao and Singh, 2009). The priorities turn out to be very much like the reciprocals of the ridge coefficients. This approach has some commonality with Guggemos and Tillé (2010) who combine hard calibration with penalisation and regard it as a design-based alternative to procedures based on mixed models. Our approach is based entirely on penalisation, although by relating it to ridge calibration we establish a link with linear models.

The algorithm is computationally undemanding and some of its properties are derived analytically. Specification of the priorities as tuning parameters may seem like an additional burden; however, these parameters facilitate a control of the process of calibration that is not available with some established methods.

For a similar computational approach in an unrelated context, namely, balancing in causal inference, see Longford (2024). It replaces the goal of achieving a balance of specified quality for two treatment groups on a set of background variables with the best balance that can be achieved given the analyst's priorities – the relative urgency or importance of reducing the imbalance for each background variable.

The remainder of this section sets up the notation and presents the analytical context of the problem. The next section formulates the problem, derives its solution and relates it to ridge calibration. Section 3 discusses how the tuning parameters are to be set. Section 4 illustrates the method on examples. Section 5 presents a simulation study that demonstrates the efficiency of the proposed estimator and explores the value of auxiliary information and good implementation of the sampling design. The concluding section summarises the method, its strengths and full potential, and discusses some unresolved issues.

1.1 Notation and context

In a population survey with a specified sampling design, we consider an estimator of the population total of a variable y, linear in the vector of its observed values $\mathbf{y} = (y_1, \ldots, y_n)^\top$. For example, when the sample size n is fixed, this may be the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), $\hat{\theta}_{\text{HT}} = \mathbf{w}^\top \mathbf{y}$, where $\mathbf{w} = (w_1, \ldots, w_n)^\top$ is the vector of sampling weights, equal to the reciprocals of the probabilities of inclusion of the subjects in the sample. These probabilities are set by design and, possibly, adjusted after the sample is drawn.

Suppose the population totals t_k are known for one or several variables X_k, $k = 1, \ldots, K$. Denote by $\mathbf{X}^\circ$ their collection. We use $\mathbf{X}^\circ$ also for the $n \times K$ matrix formed by the values of these variables in the sample; $\mathbf{x}_k$ denotes column k of $\mathbf{X}^\circ$. To simplify the discourse, we assume that every variable X_k in $\mathbf{X}^\circ$ is either ordinal (continuous) or binary. A discrete variable with $H \geq 2$ categories is represented in $\mathbf{X}^\circ$ by H binary (dummy) variables. Singularity of $\mathbf{X}^\circ$ will raise no issues. We reserve the subscript 0 for the intercept, $\mathbf{x}_0 = (1, \ldots, 1)^\top$, and denote $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}^\circ)$.

Calibration is defined as a transformation of the weights, $\mathbf{u} = C(\mathbf{w})$, for which the weighted total $\mathbf{x}_k^\top \mathbf{u}$ matches the population total t_k for every variable k. That is, calibration arranges that $\mathbf{X}^{\circ\top} \mathbf{u} = \mathbf{t}^\circ$, where $\mathbf{t}^\circ = (t_1, \ldots, t_K)^\top$. Transformations of the original variables, including interactions (products), may be added to $\mathbf{X}^\circ$ when their population totals are known or are estimated with negligible error. Other population summaries, such as variances and quantiles, can also be matched.

We write $C(\mathbf{w}) = C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)$ to indicate the auxiliary information involved; this is useful when we consider which variables to include in $\mathbf{X}^\circ$. We may qualify the estimator $\hat{\theta}$ similarly, by writing $\hat{\theta}(\mathbf{u})$ and $\hat{\theta}\{C(\mathbf{w})\}$ or, more completely, $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}^\circ, \mathbf{t}^\circ)\}$. The outcomes $\mathbf{y}$ play no role in the search for $\mathbf{u} = C(\mathbf{w})$. Therefore, so long as $\mathbf{y}$ is not inspected until we settle on a particular calibration $C(\mathbf{w})$, the properties of estimator $\hat{\theta}(\mathbf{u})$ can be assessed without any regard for how $\mathbf{u}$ was derived. No concerns about capitalising on chance or data-snooping arise, even if calibration is explored using several matrices $\mathbf{X}^\circ$ and parameters involved in C. Of course, the properties of $\hat{\theta}(\mathbf{u})$ depend on $\mathbf{u}$.

In one perspective, when the estimation error $\hat{\theta} - \theta$ and its stochastic summary, such as the bias or mean squared error (MSE), are the sole concern, calibration has a distinctly cosmetic quality. In another, prevailing in practice, it is essential for the credibility of the estimates, even at the expense of some bias and inflation of MSE. Only when this inflation is substantial, or the weights $C(\mathbf{w})$ are much more dispersed than $\mathbf{w}$, some improvisation is called for. This often happens for relatively large K, when there are many auxiliary variables and constraints associated with them. This problem is usually resolved by removing some variables from $\mathbf{X}^\circ$.

We regard the dichotomy of including or excluding a variable in the calibration process as too rigid, and develop an approach in which the calibration constraints are assigned priorities that reflect the importance or urgency to match the weighted sample total of auxiliary variable X_k with its (population) target t_k. Priority is assigned also to other desirable properties: small alteration of the weights by the adjustment, preference for smaller dispersion of the elements of $\mathbf{u}$ and aversion to a change of the total of the weights.

In brief, the established constraints of exact match (no discrepancy) are replaced by penalties for the discrepancies. These penalties allow for some slack or leeway; they involve user-defined coefficients that quantify the relative priority of the constraints.

Our formulation of the problem leads to quadratic optimisation that has a closed-form solution. It involves no iterations and no large matrices have to be inverted (numerically), even with many margins being matched for large-scale data. For a set of tuning parameters, called *priorities*, the solution is unique, and its dependence on these priorities is easy to explore, dispensing with the need for any asymptotic theory and extensive experimentation with a black-box-like algorithm.

The priorities have a natural interpretation as the importance or urgency of each calibration constraint. More precisely, these constraints are not satisfied exactly, as stipulated in hard calibration, but each discrepancy $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$, an element of $\boldsymbol{\delta}^\circ = \mathbf{X}^{\circ\top}\mathbf{u} - \mathbf{t}^\circ$, may be made negligible, $\delta_k \doteq 0$, by setting the corresponding priority sufficiently high. At the other extreme, zero priority for variable k is equivalent to dropping column $\mathbf{x}_k$ from $\mathbf{X}^\circ$. Such flexibility may be seen as a distraction, imposing the burden of declaring the priorities, and having to justify the choice in a subsequent report. However, it offers an opportunity to incorporate the client's perspective, value judgements, insights and remits in the analysis. Also, one or a few outlying discrepancies $|\delta_k|$ can be reduced by increasing the corresponding priorities, possibly at the expense of increasing some other discrepancies.

2. Unconstrained optimisation

Denote by $\mathbf{0}_K$ the vector of zeros of length K. The subscript K is dropped when the length of the vector is obvious from the context. We use the symbols $\mathbf{1}_n$ and $\mathbf{1}$ similarly for the vector of ones, and $\mathbf{I}$ for the identity matrix. For a population of (finite) size N, we consider a sample of (fixed or random) size $n \ll N$ with the vectors of observations $\mathbf{y}$ and base weights $\mathbf{w}$.

A typical approach to calibration imposes the constraint $\delta_k = 0$ or specifies an upper bound Δ_k on $|\delta_k|$ for each $k = 1, \ldots, K$. These constraints can be replaced by a single upper bound for the sum of squares $\boldsymbol{\delta}^{\circ\top}\boldsymbol{\delta}^\circ = \delta_1^2 + \cdots + \delta_K^2$. For larger K we may distinguish variables k for which the match, $\delta_k = 0$, is more important than for the rest. Further, a separate priority coefficient may be assigned to each variable, or the variables may be partitioned to sets with constant coefficients within these sets. The squares in the sum $\sum_k \delta_k^2$ may be associated with weights, imposing an upper bound on $\boldsymbol{\delta}^{\circ\top}\mathbf{P}^\circ\boldsymbol{\delta}^\circ = \sum_k p_k^\circ \delta_k^2$, with the priorities $p_k^\circ > 0$ set by the analyst; $\mathbf{P}^\circ$ is the diagonal matrix with $p_1^\circ, \ldots, p_K^\circ$ on its diagonal.

These ways of relaxing the calibration constraints and introducing priorities for reducing the discrepancies motivate our proposal. We assign a nonnegative priority p_k° for each variable k and consider first finding the minimum of the function

$$F^\circ(\mathbf{u}; \mathbf{w}) = \sum_{k=1}^{K} p_k^\circ \delta_k^2,$$

subject to the constraints of small deviation of $\mathbf{u}$ from $\mathbf{w}$ and of the weight total matching the population size N; $\mathbf{u}^\top \mathbf{1}_n = N$. When N is not known we replace it by its estimate $\mathbf{w}^\top \mathbf{1}_n$. Parameter p_k° is confounded with the scale of $\mathbf{x}_k$; replacing $\mathbf{x}_k$ with $c^{-1}\mathbf{x}_k$ is compensated by replacing p_k° with $c^2 p_k^\circ$. To avoid the associated ambiguity, we assume that each $\mathbf{x}_k$ is standardised, that is, linearly transformed to have zero mean and unit variance; t_k is adjusted accordingly. As an alternative the values $\mathbf{x}_k$ and t_k may be retained and the corresponding priority rescaled.

We relax the constraint $\mathbf{u}^\top \mathbf{1} = t_0$, where $t_0 = N$ or $t_0 = \mathbf{w}^\top \mathbf{1}$, and incorporate it in the objective function F°. We define priority p_0° by the importance assigned to δ_0^2 attaining a small value. The adapted function F° includes the new term $p_0^\circ \delta_0^2$; that is, its summation is now from $k = 0$ to K. It corresponds to attaching $\mathbf{x}_0 = \mathbf{1}_n$ to $\mathbf{X}^\circ$ as another column, forming the matrix $\mathbf{X}$.

We assign priority S to small dispersion of the weights $\mathbf{u}$, motivated by the desire for small variance of $\hat{\theta}(\mathbf{u}) = \mathbf{u}^\top \mathbf{y}$, and R to the desire for small alteration of the weights, which is indirectly related to bias reduction, preferring $\hat{\theta}(\mathbf{u})$ to remain close to estimator $\hat{\theta}(\mathbf{w})$ which would be unbiased if the weights $\mathbf{w}$ were correct and n fixed. Note that small variance cannot be equated with efficiency.

Instead of $F^\circ(\mathbf{u})$ we find the (unconstrained) minimum of the function

$$\begin{aligned} F(\mathbf{u}; \mathbf{w}) &= \sum_{k=0}^{K} p_k^\circ \delta_k^2 + R(\mathbf{u}-\mathbf{w})^\top(\mathbf{u}-\mathbf{w}) + S\left(\mathbf{u}^\top\mathbf{u} - \frac{1}{n}\mathbf{u}^\top \mathbf{1}\mathbf{1}^\top \mathbf{u}\right). \\ &= (R+S)\,\mathbf{u}^\top\mathbf{u} - \frac{1}{n} S\,\mathbf{u}^\top \mathbf{1}\mathbf{1}^\top \mathbf{u} + \sum_{k=0}^{K} p_k^\circ \delta_k^2 - 2R\,\mathbf{u}^\top\mathbf{w} + R\,\mathbf{w}^\top\mathbf{w}. \end{aligned}$$

Since optimising cF for positive constants c constitutes identical problems, no generality is lost by assuming that $R + S = 1$. With this convention, and by expanding

$$p_k^\circ \delta_k^2 = p_k^\circ \mathbf{u}^\top \mathbf{x}_k \mathbf{x}_k^\top \mathbf{u} - 2 p_k^\circ t_k \mathbf{u}^\top \mathbf{x}_k + p_k^\circ t_k^2$$

for $k = 0, \ldots, K$, we can express F compactly as

$$F(\mathbf{u}) = \mathbf{u}^\top \mathbf{H}\mathbf{u} - 2\mathbf{u}^\top \mathbf{s} + D,$$

where

$$\begin{aligned} \mathbf{H} &= \mathbf{I}_n + \sum_{k=0}^{K} p_k \mathbf{x}_k \mathbf{x}_k^\top = \mathbf{I}_n + \mathbf{X}\mathbf{P}\mathbf{X}^\top \\ \mathbf{s} &= R\mathbf{w} + (1-R)\,\frac{t_0}{n}\,\mathbf{1}_n + \sum_{k=0}^{K} p_k t_k \mathbf{x}_k = \mathbf{w}_R + \mathbf{X}\mathbf{P}\mathbf{t}; \end{aligned} \tag{2.1}$$

$\mathbf{w}_R = R\mathbf{w} + (1-R)\,n^{-1} t_0 \mathbf{1}_n$, $p_0 = p_0^\circ - \frac{1}{n} S$, $p_k = p_k^\circ$ for $k = 1, \ldots, K$, $\mathbf{P} = \operatorname{diag}(\mathbf{p})$, where $\mathbf{p} = (p_0, p_1, \ldots, p_K)$, and $D = \mathbf{t}^\top \mathbf{P}\mathbf{t} + R\mathbf{w}^\top\mathbf{w} + (1-R)\,t_0^2/n$ is a positive scalar not relevant to what follows. The minimum of $F(\mathbf{u})$ is attained for $\mathbf{u} = \mathbf{H}^{-1}\mathbf{s}$, and the minimum attained is $F(\mathbf{u}) = D - \mathbf{s}^\top \mathbf{H}^{-1}\mathbf{s}$. The calibration estimator of θ is $\hat{\theta}(\mathbf{u}) = \mathbf{y}^\top \mathbf{u} = \mathbf{y}^\top \mathbf{H}^{-1}\mathbf{s}$. Section 3 discusses how the values of the priorities p_k and R should be set.

2.1 Inversion of H

For large sample size n, numerical inversion of $\mathbf{H}$ might seem like a nontrivial computational hurdle. We apply a recursive algorithm that exploits the form of $\mathbf{H}$ as the sum of a matrix that is easy to invert, $\mathbf{I}$, and one, $\mathbf{XPX}^\top$, of (relatively) low rank, at most $K+1 \ll n$. We note in passing that singularity of the matrix $\mathbf{X} = (\mathbf{x}_0, \mathbf{X}^\circ)$ raises no computational issues, although it may have some consequences on the interpretation of the priorities in $\mathbf{p}$. For example, if $\mathbf{x}_k = \mathbf{x}_l$ for some $k \neq l$, then the common variable is more appropriately associated with priority $p_k + p_l$. Also, a conflict arises when $\mathbf{x}_0 = \mathbf{1}_n$ is equal to the total of the set of columns of $\mathbf{X}^\circ$ that correspond to the indicators for a categorical variable (prior to standardisation), but the total of the corresponding targets (population counts) t_k differs from t_0. The algorithm is not affected by such a conflict.

Denote $\mathbf{H}_{-1} = \mathbf{I}_n$ and $\mathbf{H}_k = \mathbf{H}_{k-1} + p_k \mathbf{x}_k \mathbf{x}_k^\top$, $k = 0, \ldots, K$, so that $\mathbf{H} = \mathbf{H}_K$. We have the identity

$$\mathbf{H}_k^{-1} \;=\; \mathbf{H}_{k-1}^{-1} \;-\; \frac{p_k}{1 + p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1} \mathbf{x}_k}\; \mathbf{H}_{k-1}^{-1} \mathbf{x}_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1}.$$

Its validity is easy to check by evaluating the product of the expressions for $\mathbf{H}_k$ and $\mathbf{H}_k^{-1}$. In the recursive evaluation of $\mathbf{u} = \mathbf{H}^{-1}\mathbf{s}$ we do not have to form any matrices $\mathbf{H}_k$ or $\mathbf{H}_k^{-1}$ because we require only the vectors $\mathbf{h}_{kl} = \mathbf{H}_k^{-1}\mathbf{x}_l$ and $\mathbf{h}_{k,\mathrm{w}} = \mathbf{H}_k^{-1}\mathbf{w}$. For the former we have the identities

$$\mathbf{h}_{kl} \;=\; \mathbf{h}_{k-1,l} \;-\; \frac{p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,l}}{1 + p_k \mathbf{x}_k^\top \mathbf{h}_{k-1,k}}\; \mathbf{h}_{k-1,k},$$

and for the latter the same identities, with index l replaced by w. Note that every denominator $1 + p_k \mathbf{x}_k^\top \mathbf{H}_{k-1}^{-1} \mathbf{x}_k$ is positive. The solution,

$$\mathbf{u} \;=\; R\mathbf{h}_{K,\mathrm{w}} \;+\; (1-R)\,\frac{t_0}{n}\,\mathbf{h}_{K0} \;+\; \sum_{k=0}^{K} p_k t_k \mathbf{h}_{Kk},$$

is a linear combination of the vectors $\mathbf{h}_{Kk} = \mathbf{H}^{-1}\mathbf{x}_k$ and $\mathbf{h}_{K,\mathrm{w}} = \mathbf{H}^{-1}\mathbf{w}$. In summary, there is a unique solution $\mathbf{u}$ and it is evaluated only by operations on vectors of length n. Matrix $\mathbf{H}$ involves neither R nor $\mathbf{w}$. The vector $\mathbf{s}$ is a linear function of both R and $\mathbf{w}$, and therefore so is the solution $\mathbf{u}$. For $R = 0$, $\mathbf{u}$ does not depend on $\mathbf{w}$.

2.2 Relation to ridge calibration

In this section we show that the priorities $\mathbf{p}$ have a role similar to the ridging costs in Chambers (1996), equation (10), although function F incorporates all constraints and aims of calibration, and therefore involves additional parameters. Just like ridging can be motivated as a compromise between applying no ridging and infinite ridging, our proposal is a compromise of $\mathbf{p} = \mathbf{0}$ and infinitely large $\mathbf{p}$. Both approaches yield estimators that can be interpreted as shrinkage estimators.

In equation (2.1), $\mathbf{w}_R$ is a convex combination of $\mathbf{w}$ and t_0/n, interpreted as shrinking $\mathbf{w}$ towards its mean or expectation. Write $\mathbf{w}_R = \mathbf{X}\mathbf{v}_R + \boldsymbol{\varepsilon}_R$, where $\mathbf{v}_R = (\mathbf{X}^\top\mathbf{X})^- \mathbf{X}^\top\mathbf{w}_R$ is a projecting vector and $\boldsymbol{\varepsilon}_R$ is such that $\mathbf{X}^\top\boldsymbol{\varepsilon}_R = \mathbf{0}$. Here $(\mathbf{X}^\top\mathbf{X})^-$ is a generalised inverse of $\mathbf{X}^\top\mathbf{X}$. Its non-uniqueness is resolved below. We have the analogous decomposition $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\omega}$, where $\boldsymbol{\gamma} = (\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{y}$, so that $\mathbf{X}^\top\boldsymbol{\omega} = \mathbf{0}$. The vectors $\mathbf{v}_R$ and $\boldsymbol{\gamma}$ are respective least-squares fits for $\mathbf{w}_R$ and $\mathbf{y}$ in terms of $\mathbf{X}$. We do not assume any aspects of validity of the implied ordinary regressions $(\mathbf{w}_R \mid \mathbf{X})$ and $(\mathbf{y} \mid \mathbf{X})$. Note that the residuals $\boldsymbol{\varepsilon}_R$ and $\boldsymbol{\omega}$ are unique, and $\boldsymbol{\varepsilon}_R^\top\mathbf{y} = \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$.

Let $\mathbf{e}_0 = (1, 0, \ldots, 0)^\top$ be the vector comprising one unity and K zeros. For a matrix $\mathbf{A}$ with $K+1$ columns, $\mathbf{A}\mathbf{e}_0$ is equal to its first column. Owing to orthogonality, $\mathbf{X}^{\circ\top}\mathbf{x}_0 = \mathbf{0}$, arranged by standardisation, we have the identity $(\mathbf{X}^\top\mathbf{X})^-\mathbf{X}^\top\mathbf{1}_n = n(\mathbf{X}^\top\mathbf{X})^-\mathbf{e}_0 = \mathbf{e}_0$. Hence

$$\mathbf{v}_R \;=\; R\left(\mathbf{X}^\top\mathbf{X}\right)^- \mathbf{X}^\top\mathbf{w} \;+\; \frac{(1-R)\,t_0}{n}\;\mathbf{e}_0, \tag{2.2}$$

and by its substitution in (2.1),

$$\mathbf{u} \;=\; \mathbf{H}^{-1}\mathbf{s} \;=\; \left(\mathbf{I} \;+\; \mathbf{X}\mathbf{P}\mathbf{X}^\top\right)^{-1}\left\{\mathbf{X}\left(\mathbf{P}\mathbf{t} \;+\; \mathbf{v}_R\right) \;+\; \boldsymbol{\varepsilon}_R\right\}.$$

The identity $(\mathbf{I} + \mathbf{X}\mathbf{P}\mathbf{X}^\top)^{-1} = \mathbf{I} - \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ implies that

$$\begin{aligned}\mathbf{u} \;&=\; \mathbf{X}\left\{\mathbf{I} - \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\right\}\left(\mathbf{P}\mathbf{t} + \mathbf{v}_R\right) \;+\; \left\{\mathbf{I} - \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}\boldsymbol{\varepsilon}_R \\ &=\; \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t} + \mathbf{P}^{-1}\mathbf{v}_R\right) \;+\; \boldsymbol{\varepsilon}_R.\end{aligned}$$

Therefore the calibration estimator is $\hat{\theta}(\mathbf{u}) = \hat{\theta}_1 + \hat{\theta}_2(\mathbf{w}) + \hat{\theta}_3(\mathbf{w})$, where

$$\begin{aligned}\hat{\theta}_1 \;&=\; \mathbf{t}^\top\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \hat{\theta}_2(\mathbf{w}) \;&=\; \mathbf{v}_R^\top\mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \hat{\theta}_3(\mathbf{w}) \;&=\; \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}.\end{aligned} \tag{2.3}$$

Here $\hat{\theta}_1$ is an estimator of the population total θ based on the ridge regression prediction by the linear model $(\mathbf{y} \mid \mathbf{X})$. Not involving the weights $\mathbf{w}$ and based on ridging, $\hat{\theta}_1$ is unbiased only when the biases due to no weighting and ridging happen to cancel out. When $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega} = 0$, $\hat{\theta}(\mathbf{u})$ has the form of ridge calibration (Chambers, 1996; Goga and Shehzad, 2010), with $\mathbf{P}$ in the role of the (diagonal) cost matrix and $\mathbf{t}$ adjusted by $\mathbf{P}^{-1}\mathbf{v}_R$. However, the crossproducts $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top\mathbf{y}$ are evaluated without the weights $\mathbf{w}$.

As $\mathbf{P}^{-1}$ converges to the zero matrix, which corresponds to diminishing interest in the deviation of $\mathbf{u}$ from $\mathbf{w}$ as well as in the dispersion of $\mathbf{u}$, $\hat{\theta}(\mathbf{u})$ approaches $\mathbf{t}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega}$ when $\mathbf{X}$ is nonsingular. This reduces to the least-squares predictor when $\boldsymbol{\varepsilon}_R^\top\boldsymbol{\omega} = 0$. If $\mathbf{X}$ has deficient rank and all priorities in $\mathbf{p}$ are large, then $\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}$ has one or more small eigenvalues, and so $\hat{\theta}(\mathbf{u})$ is unstable.

As each priority in $\mathbf{p}$ converges to zero, $\hat{\theta}(\mathbf{u})$ converges to the shrinkage estimator $\mathbf{w}_R^\top \mathbf{y} = R\mathbf{w}^\top \mathbf{y} + (1-R)\, t_0\, \bar{y}$, where $\bar{y}$ is the sample (unweighted) mean of y. For $R = 0$ this is the trivial (unweighted) estimator $t_0 \bar{y}$ and for $R = 1$ it is the weighted mean $\mathbf{w}^\top \mathbf{y}$. Note that the elements of $\mathbf{p}$ (the diagonal of $\mathbf{P}$) are not set by the usual considerations of ridge regression to reduce the sampling variance in exchange for a small bias. Introducing these considerations is not straightforward because of the contribution of $\hat{\theta}_2 + \hat{\theta}_3$ to $\hat{\theta}$.

Denote by $\hat{\mathbf{y}}$ the projection vector $\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-}\mathbf{X}^\top\mathbf{y}$ and by $\hat{\mathbf{y}}_{\mathrm{P}}$ its "ridged" version $\mathbf{X}(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Then $\hat{\theta}_3(\mathbf{w}) = R\boldsymbol{\varepsilon}_1^\top \boldsymbol{\omega} = R\mathbf{w}^\top(\mathbf{y} - \hat{\mathbf{y}})$ and, owing to (2.2),

$$
\begin{aligned}
\hat{\theta}_2(\mathbf{w}) &= R\mathbf{w}^\top \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-} \mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&\quad + \frac{(1-R)\, t_0}{n}\, \mathbf{e}_0^\top \mathbf{P}^{-1}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= R\mathbf{w}^\top \mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-} - \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\right\} \mathbf{X}^\top\mathbf{y} + \frac{1-R}{1+np_0}\, \frac{t_0}{n}\, \mathbf{y}^\top \mathbf{1}_n \\
&= R\mathbf{w}^\top\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + (1-R)\, \frac{t_0 \bar{y}}{1+np_0}.
\end{aligned}
$$

Hence the total $\hat{\theta}_2(\mathbf{w}) + \hat{\theta}_3(\mathbf{w})$ is

$$
\hat{\theta}_{23} = R\mathbf{w}^\top\left(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + (1-R)\, \frac{t_0 \bar{y}}{1+np_0}.
$$

It is a linear function of both $\mathbf{w}$ and R, shrinking the weighted total of the prediction errors, $\mathbf{w}^\top(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}})$, toward a quantity that does not involve $\mathbf{X}$. Column 0 of $\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}$ is equal to $\left(1/p_0 + n\right)\mathbf{e}_0$, and so, recalling the notation $\mathbf{X}^\circ$, $\mathbf{P}^\circ$ and $\mathbf{t}^\circ$, $\hat{\theta}(\mathbf{u})$ can be expressed as

$$
\mathbf{t}^{\circ\top}\left(\mathbf{P}^{\circ -1} + \mathbf{X}^{\circ\top}\mathbf{X}^\circ\right)^{-1}\mathbf{X}^{\circ\top}\mathbf{y} + R\mathbf{w}^\top\left(\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{P}}\right) + \left(1 - \frac{R}{1+np_0}\right) t_0 \bar{y}.
$$

It is useful to present $\hat{\theta}(\mathbf{u}; R)$ as the linear interpolation of

$$
\begin{aligned}
\hat{\theta}(\mathbf{u}; R=0) &= \frac{t_0\, \bar{y}}{1+np_0} + \mathbf{t}^\top \hat{\boldsymbol{\beta}} \\
&= t_0\, \bar{y} + \mathbf{t}^{\circ\top}\hat{\boldsymbol{\beta}}^\circ \\
\hat{\theta}(\mathbf{u}; R=1) &= \mathbf{w}^\top\mathbf{y} - \boldsymbol{\delta}_{\mathrm{w}}^\top \hat{\boldsymbol{\beta}},
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, $\hat{\boldsymbol{\beta}}^\circ = (\mathbf{P}^{\circ -1} + \mathbf{X}^{\circ\top}\mathbf{X}^\circ)^{-1}\mathbf{X}^{\circ\top}\mathbf{y}$, and $\boldsymbol{\delta}_{\mathrm{w}} = \mathbf{X}^\top\mathbf{w} - \mathbf{t}$ is the discrepancy vector evaluated with weights $\mathbf{w}$. The $K \times 1$ vectors $\mathbf{t}^\circ$ and $\hat{\boldsymbol{\beta}}^\circ$ are formed from $\mathbf{t}$ and $\hat{\boldsymbol{\beta}}$ by dropping their respective first elements. Thus, $\hat{\theta}(\mathbf{u}; R=0)$ can be described as the trivial estimator $t_0\, \bar{y}$ adjusted by ridge prediction at $\mathbf{t}^\circ$. For $R \geq 0$,

$$\hat{\theta}(\mathbf{u}; R=r) = \hat{\theta}(\mathbf{u}; R=0) + r\left\{\mathbf{w}^\top\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) - \frac{t_0\,\bar{y}}{1+np_0}\right\}; \tag{2.4}$$

that is, $\hat{\theta}(\mathbf{u}; R=0)$ is adjusted by R-multiple of the estimator of the population-total error in the prediction of $\mathbf{y}$ and a fraction of $t_0\,\bar{y}$. Setting the value of R matters more when the prediction errors $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}$ are large and are correlated with the base weights $\mathbf{w}$. Note that $\hat{\theta}$ is evaluated without using $(\mathbf{X}^\top\mathbf{X})^-$.

The bias of $\hat{\theta}$ is a linear function of R, so it is minimised either for $R=0$ or 1. Even though $\hat{\theta}$ does not involve $\mathbf{w}$ when $R=0$, it would have the smallest bias when the effect of the weights is estimated well by the prediction model. This is the case when the weights are constructed using only variables in $\mathbf{X}$, for which the population totals are available. We confirm this by simulations in Section 5.

In Appendix A we derive and discuss an expression for $\hat{\theta}$ in (2.4) when calibration is on a single categorical variable (and the population size). The estimator has a decomposition $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$, where $\hat{\theta}^{(1)}$ depends on $\boldsymbol{\mu}$, $\mathbf{p}$ and $\mathbf{t}$ only through p_0 and t_0, and $\hat{\theta}^{(Z)}$ does not depend on p_0. Also, $\hat{\theta}^{(Z)} = 0$ when $p_1 = \ldots = p_K = 0$. In this case, calibrating only on the population size,

$$\hat{\theta} = \hat{\theta}^{(1)} = R\left(\mathbf{w}^\top\mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,\mathbf{w}^\top\mathbf{1}_n\,\bar{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right)t_0\,\bar{y}.$$

For $R=0$, we have $\hat{\theta} = t_o\,\bar{y}$ for any p_0. For $R=1$, $\hat{\theta} = \mathbf{w}^\top\mathbf{y} + np_0/(1+np_0) \times (t_0 - \mathbf{w}^\top\mathbf{1}_n)\,\bar{y}$, which converges to the "obvious" adjustment $\mathbf{w}^\top\mathbf{y} + (t_0 - \mathbf{w}^\top\mathbf{1}_n)\,\bar{y}$ as $p_0 \to +\infty$.

3. Setting the priorities

In some applications, the constraints are equally important for all the variables, but only after taking into account the dispersions of these variables. As stated earlier, we assume that every variable in $\mathbf{X}^\circ$ is standardised. Setting the priorities $p_1, \ldots, p_K$ to a common constant is a reasonable starting point or default. For large K we may define a small number of groups of variables and assign a common priority within each group. For example, these groups of variables may be associated with distinct categorical variables. The intercept $\mathbf{x}_0 = \mathbf{1}$ has a special status among the variables in $\mathbf{X}$. High priority p_0 corresponds to the desire for small $|\delta_0|$. When $t_0 = N$ and p_0 is set sufficiently large, $\mathbf{u}^\top\mathbf{1} \doteq N$.

By construction, $0 < R < 1$, so $R = 0.5$ might be a default. A more profound approach would weigh the relative importance of small change (R) and small variance (S) of the calibration weights. With greater n, concerns about bias become dominant, so smaller S and greater $R = 1 - S$ are appropriate. Setting R to a very small value is usually unwise because then $\hat{\theta}$ depends only weakly on $\mathbf{w}$. Reducing the difference $\mathbf{u} - \mathbf{w}$ and the dispersion of the weights are both devices to avoid extreme (very large and negative) weights. These goals are given greater prominence by reducing the priorities, e.g., from $\mathbf{p}$ to $c\,\mathbf{p}$ for $0 < c < 1$.

These guidelines are admittedly rather vague and incomplete. However, the computational simplicity of the solution allows us to explore a range of plausible settings of $\mathbf{p}$ and R, especially when $\mathbf{p}$ involves only

a few distinct values. For example, when $\mathbf{X}^\circ$ is based on a single discrete variable with H categories, then $p_1 = \ldots = p_H$, and so we have only three tuning parameters, p_0, p_1 and R. When there are two discrete variables four parameters are involved. When the entire two-way table of the population margins for these two variables is available, matching the univariate margins is usually more important than matching the two-way subtotals.

Dependence of the fitted discrepancy $\delta_k = \mathbf{x}_k^\top \mathbf{u} - t_k$ on p_k can be explored analytically. By matrix differentiation we obtain the identity

$$\begin{aligned} \frac{\partial \delta_l^2}{\partial p_k} &= -2\delta_l \left(\mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{H}}{\partial p_k} \mathbf{H}^{-1} \mathbf{s} - \mathbf{x}_l^\top \mathbf{H}^{-1} \frac{\partial \mathbf{s}}{\partial p_k} \right) \\ &= -2\delta_k \delta_l \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_l, \end{aligned} \tag{3.1}$$

and its special case,

$$\frac{\partial \delta_k^2}{\partial p_k} = -2\delta_k^2 \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k < 0,$$

so that

$$\frac{\partial \log\left(\delta_k^2\right)}{\partial p_k} = -2\, \mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k.$$

Hence, $|\delta_k|$ is a decreasing function of p_k, in accord with the motivation of p_k as the relative importance of reducing $|\delta_k|$. However, the decrease is slower when $|\delta_k|$ is smaller;

$$\frac{\partial^2 \log\left(\delta_k^2\right)}{\partial p_k^2} = 2\left(\mathbf{x}_k^\top \mathbf{H}^{-1} \mathbf{x}_k\right)^2,$$

so δ_k^2 is a log-convex function. Interpreted loosely, trying to wipe out a small discrepancy $|\delta_k|$ requires a sizeable change of p_k, and that could inflate some other discrepancies $|\delta_l|$.

The solution $\mathbf{u}$ is a linear function of R, which is involved only in $\mathbf{s}$;

$$\frac{\partial \mathbf{u}}{\partial R} = \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\mathbf{1}_n\right),$$

and

$$\frac{\partial \delta_k}{\partial R} = \mathbf{x}_k^\top \mathbf{H}^{-1}\left(\mathbf{w} - \frac{t_0}{n}\,\mathbf{1}_n\right). \tag{3.2}$$

The expression for $\mathbf{s}$ in equation (2.1), as well as the original intent, imply that R acts as a shrinkage factor, reducing the deviation of the weights $\mathbf{u}$ from $\mathbf{w}$ and $1-R$ as reducing their dispersion. Indeed, when no auxiliary information is available and $\mathbf{u}^\top \mathbf{1}_n$ is matched to $\mathbf{w}^\top \mathbf{1}_n$, $\mathbf{H} = \mathbf{I}_n$ and the solution is the convex

combination $\mathbf{w}_R$ introduced in equation (2.1). In this case, we have a simple expression, $\mathrm{E}(\mathbf{w}_R^\top \mathbf{y}) - \theta = -(1-R)\,\mathrm{cov}(w, y)$, for the bias of the calibrated estimator when the weights $\mathbf{w}$ are correct and $\mathbf{w}^\top \mathbf{y}$ is unbiased.

The bias of $\hat{\theta}(\mathbf{u})$ is likely to be largest when R is set to zero, when $\mathbf{u}$ does not depend on $\mathbf{w}$. At the same time, $\delta_0 = 0$ only when $R = 0$, unless $\mathbf{w}^\top \mathbf{1} = t_0$; see Appendix B for proof. So, focus on one goal of calibration to the detriment of others may be ill-advised.

4. Examples

In this section we demonstrate the control of the discrepancies $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_K)^\top$ and dispersion of the calibrated weights $\mathbf{u}$ by the priorities $\mathbf{p}$ and parameter R. We use a synthetic population of size $N =$ 120,000 from which we draw a sample of expected size $\mathrm{E}(n) \doteq$ 1,000 according to a design with unequal probabilities and independent inclusions in the sample, without replacement. In the population, we have a discrete variable Z with six categories and a continuous variable X.

The six subpopulations defined by Z have sizes between 13,000 and 26,000. The values of the sole continuous background variable X are generated as random samples, one within each category k, from gamma distributions with shapes ξ_k and common rate 5.0, where $\boldsymbol{\xi} =$ (6.4, 6.7, 6.1, 6.6, 6.9, 6.4), so that the within-category means of X are in the range 1.22 − 1.28 and their standard deviations in the range 0.244 − 0.276. The outcome Y is generated as $\frac{1}{2}X + \exp(\zeta)$, where ζ comprises independent random samples from the normal distributions with means ξ_k in category k and standard deviations (0.05, 0.07, 0.04, 0.06, 0.08, 0.05). The entire setting has been selected arbitrarily, to generate a dataset in which the distributions of X and Y differ across the categories of Z, are skewed to the right, and X and Y are moderately correlated both overall (their correlation is 0.30), and within the six categories. The relevant summaries of X, Y and Z are given in Table 4.1.

Table 4.1
Population summaries of the variables X, Y and Z.

	Category (k) of Z						
	1	2	3	4	5	6	All
$t_1 \ldots t_6$	25,000	15,000	22,000	26,000	13,000	19,000	120,000
$\sum X$	31,949	20,053	26,710	34,329	17,890	24,421	155,352
$\bar{X}$ (sd)	1.28 (0.51)	1.34 (0.52)	1.21 (0.49)	1.32 (0.51)	1.38 (0.52)	1.29 (0.51)	1.29 (0.51)
$\sum Y$	117,449	92,376	79,480	146,142	96,201	89,336	620,984
$\bar{Y}$ (sd)	4.70 (0.32)	6.16 (0.47)	3.61 (0.27)	5.62 (0.39)	7.40 (0.60)	4.70 (0.33)	5.17 (1.18)

4.1 Sampling design

For drawing a sample from this population we consider a design that would be planned, and from which the base weights are calculated. In this design, the probabilities of inclusion are constant within the six categories, set to 0.0075, 0.0093, 0.0070, 0.0090, 0.0110 and 0.0075. The inclusions are mutually independent.

We inject realism into the simulated survey data collection by generating the "real" inclusion probabilities that reflect the imperfect implementation of the survey. The design probabilities are multiplied by a random sample from the log-normal distribution based on $\mathcal{N}(0, \sigma^2)$. No imperfection corresponds to $\sigma = 0$ and larger σ, with greater deviations from the base weights, to greater impact of the imperfections in the conduct of the survey. That is, $\hat{\theta}(\mathbf{w})$ is biased because the base weights $\mathbf{w}$ have been distorted in the implementation of the design, which has led to an altered (perturbed) set of weights $\mathbf{w}^\dagger$. By applying $\hat{\theta}\{C(\mathbf{w})\}$ we aim to remedy this problem. In practice, the altered weights $\mathbf{w}^\dagger$ cannot be recovered, nor estimated, but there may be some information (or well-founded opinion) about the extent of such perturbation. In our model, this perturbation is characterised by the variance σ^2 of the changes of the log-weights, or equivalently, of the log-probabilities.

In this section we use a sample drawn with $\sigma = 0.1$. This amounts to a substantial perturbation; the standard deviations of the perturbed weights in this sample are 13.3, 10.4, 14.4, 11.5, 9.7 and 12.3 within the respective categories 1, …, 6, whereas the base weights are constant within the categories. The average base weight is $\bar{w} = 122.44$. The average of the perturbed weights is $\bar{w}^\dagger = 123.22$.

4.2 Calibration

We calibrate a single sample and we are concerned solely with the discrepancies $\boldsymbol{\delta}$. We use the tuning parameter values

$$p_0 = 1.5,\ p_1 = \ldots = p_6 = 0.15,\ p_7 = 0.5,\ R = 0.9, \tag{4.1}$$

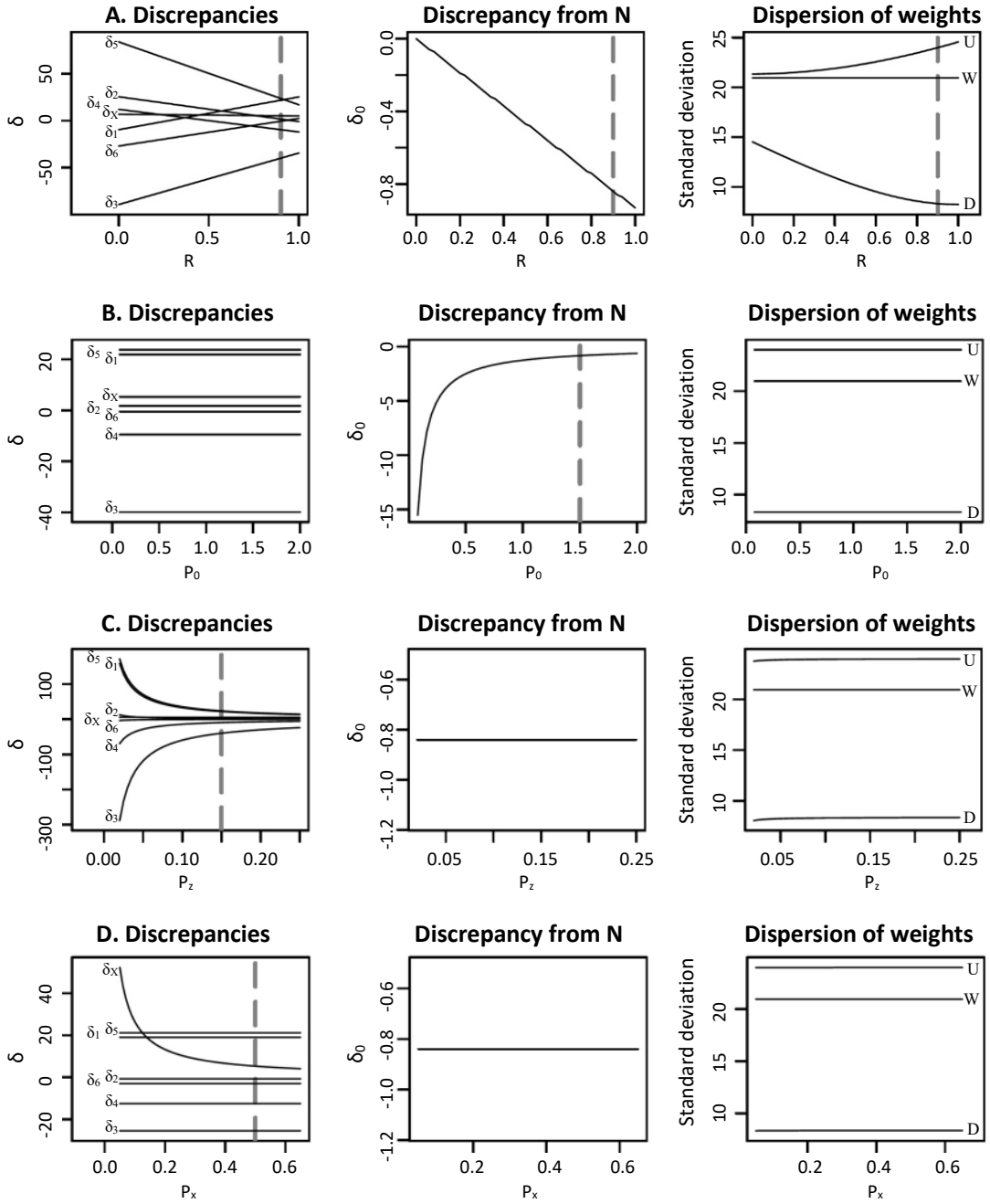
and call them the reference setting. We use the notation p_X for p_7 and δ_X for δ_7. When $p_1 = \ldots = p_6$, we denote the common value by p_Z.

We apply first calibration with values of R in the range $(0,1)$ and the reference setting for $\mathbf{p}$. The discrepancies, as functions of R, are plotted in row A at the top of Figure 4.1. In the left-hand panel, the discrepancies are plotted for X and the six categories of Z. The plot confirms that these functions $\delta_k(R)$ are linear; see equation (2.4). The discrepancies are dispersed more for $R = 0$ than for $R = 1$, although some of the functions $\delta_k(R)$ cross zero at $R \in (0,1)$.

Function $\delta_0(R)$ is plotted in a separate panel because a much narrower scale is required for the vertical axis. The match is perfect, $\delta_0 = 0$, for $R = 0$, in accord with the proof in Appendix B. The standard deviations of $\mathbf{u}$ and $\mathbf{u} - \mathbf{w}$, marked by the respective symbols U and D, are drawn in the right-hand panel, together with the standard deviation of $\mathbf{w}$ (W), which, of course, is constant. The standard deviation of $\mathbf{u}$

increases and the standard deviation of $\mathbf{u} - \mathbf{w}$ decreases with R. Based on these three plots we would choose a large value of R, to reduce the values of δ_k overall, even at a small sacrifice of δ_0. We settle on $R = 0.9$, indicated by vertical dashes.

Figure 4.1 Discrepancies δ and standard deviations of the weights as functions of the parameter $R \in (0,1)$ in row A (top), of $p_0 \in (0,2)$ in row B, of $p_Z \in (0.02, 0.25)$ in row C and of $p_X \in (0.05, 0.65)$ in row D, all adapted from the reference setting given by (4.1), indicated in the plots by long vertical dashes.



In row B, the discrepancies are plotted as functions of p_0 in the range $(0, 2)$ with the other parameters given by the reference setting. Apart from $\delta_0(p_0)$, each discrepancy and standard deviation is very close to a constant. Function $|\delta_0(p_0)|$ decreases rapidly for small values of p_0 and converges to zero very slowly for large values of p_0, as discussed following equation (3.1). Based on this plot, we set p_0 to 1.5.

In row C, the value of $p_Z = p_1 = \ldots = p_6$ is varied in the range (0.02, 0.25), and the other parameters are held at their reference values. The discrepancies $\delta_1, \ldots, \delta_6$ approach zero rapidly for small values of p_Z and converge to zero slowly for large values of p_Z. The other functions of p_Z, δ_0 and δ_X, as well as the two standard deviations in the right-hand panel, are each in a very narrow range, and their curvature is appreciable only for very small values of p_Z. We set $p_Z = 0.15$. Since $|\delta_3|$, equal to 39.9 at $p_Z = 0.15$, is larger than the discrepancies for the other five categories, we increase p_3 to 0.25. Now $\delta_3 = -25.4$, much smaller in absolute value. The second largest discrepancy is for category 5, $\delta_5 = 21.2$, reduced from 23.7 with the original setting.

The panels in row D (bottom) present the setting with p_X being varied while the other parameters are held at their reference values, except for $p_3 = 0.25$. The diagram confirms that $|\delta_X(p_X)|$ decreases rapidly for small values of p_X and more slowly as it approaches zero for large values of p_X. The other discrepancies and the two standard deviations depend on p_X very weakly. We set $p_X = 0.5$.

In summary, Figure 4.1 illustrates that we can reduce any one of the absolute discrepancies $|\delta_k|$ by increasing the corresponding priority p_k. Also, by altering R, we can trade off small dispersion of $\mathbf{u}$, essential for efficiency, for small alteration of the weights (small dispersion of $\mathbf{u} - \mathbf{w}$), indirectly related to bias. Figure 4.1 confirms that eradicating a small discrepancy requires substantial increase of the corresponding priority. It suggests that such an increase would affect the other discrepancies only slightly. Equation (3.1) and its discussion imply that this is not the case in general, especially with many variables in $\mathbf{X}$, some of them highly correlated, when several discrepancies are nontrivial.

5. Simulations

In this section we study the empirical bias and root-MSE (rMSE) of the calibration estimator $\hat{\theta}\{C(\mathbf{w}; \mathbf{X}, \mathbf{t})\} = \mathbf{u}^\top \mathbf{y}$ for several settings ($\mathbf{p}$ and R) and levels of imperfect implementation of the sampling design, characterised by the perturbation parameter σ. We also assess the contribution a variable used in calibration makes to the reduction of rMSE. The simulations we describe involve sets of 1,000 replications. We checked that this number is sufficient by comparing the results with their re-runs using 2,000 replications for a selection of cases. We use the same population ($N = 120{,}000$), planned design with $\mathrm{E}(n) \doteq 1{,}000$ and the process of perturbing the base weights to represent the imperfect implementation of the design, as described in Section 4. The bias and rMSE are presented as functions of R, using the interpolation based on (2.4). These functions are evaluated for $\sigma = 0, 0.02, \ldots, 0.10$ and a selection of vectors $\mathbf{p}$.

Figure 5.1 presents by solid black lines the bias and the root-MSE of $\hat{\theta}$, as functions of $R \in (0, 1)$ for the reference setting of $\mathbf{p}$ and $\sigma = 0, 0.02, \ldots, 0.1$, as indicated at the right-hand margin. For each value

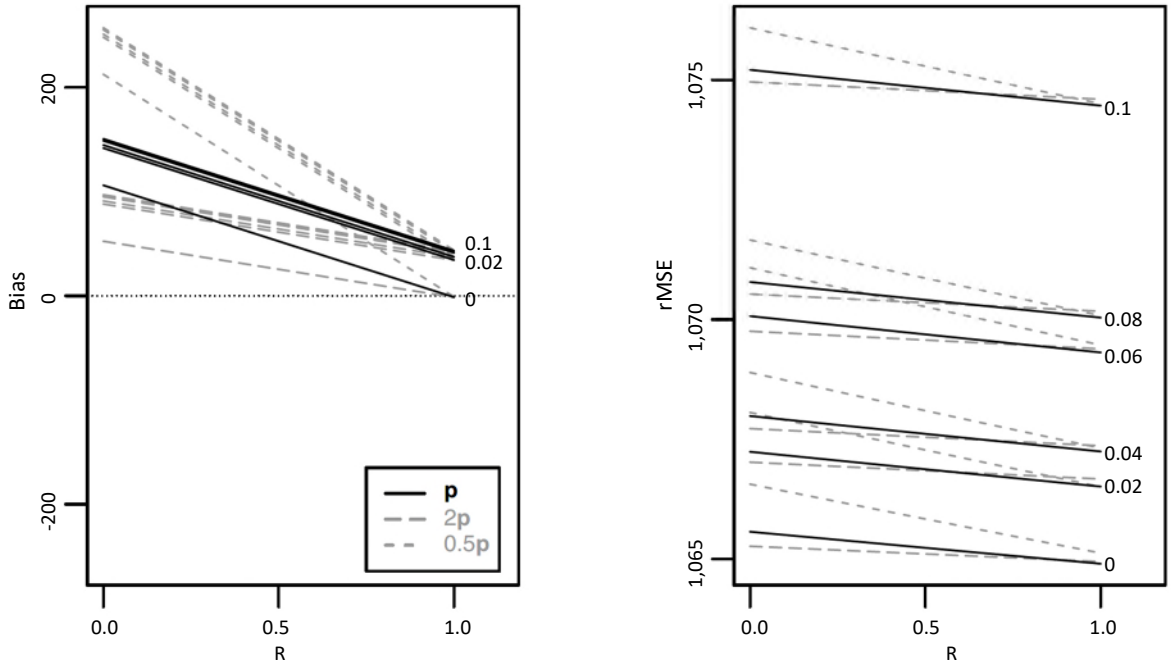
of σ, the bias is smallest for $R = 1$. The bias functions are parallel and for $\sigma \geq 0.02$ differ only slightly. The rMSE functions have very little curvature, are nearly parallel and also attain their smallest values for $R = 1$. The rMSE increases with σ for all R, although not evenly. The bias and rMSE functions for priorities $2\mathbf{p}$ and $\frac{1}{2}\mathbf{p}$ are drawn by grey lines with long and short dashes, respectively. All the functions have steeper gradients for $\frac{1}{2}\mathbf{p}$ and lower gradients for $2\mathbf{p}$, and differ from their counterparts for the reference value of $\mathbf{p}$ at $R = 1$ only slightly. Table 5.1 presents the values of rMSE at $R = 0$ and 1 in a tabular form, together with some results discussed below.

Table 5.1
Root-MSEs of the calibrated estimators of the population total of Y. The first column indicates the variables on which the design probabilities are based and, in brackets underneath, the variables used in calibration.

Design [calibration]		σ					
		0	**0.02**	**0.04**	**0.06**	**0.08**	**0.10**
		$\mathbf{p}$ (reference)					
	$R = 0$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
	$R = 1$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
		$2\mathbf{p}$					
	$R = 0$	1,065.3	1,067.0	1,067.7	1,069.8	1,070.5	1,075.0
	$R = 1$	1,064.9	1,066.7	1,067.4	1,069.4	1,070.2	1,074.6
		$\frac{1}{2}\mathbf{p}$					
	$R = 0$	1,066.6	1,068.1	1,068.9	1,071.1	1,071.7	1,076.1
	$R = 1$	1,065.1	1,066.5	1,067.3	1,069.5	1,070.1	1,074.5
Z, X		$p_3 = 0.25$					
$[Z, X]$	$R = 0$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
	$R = 1$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
		$p_0 = 15$					
	$R = 0$	1,064.9	1,066.5	1,067.2	1,069.3	1,070.0	1,074.5
	$R = 1$	1,065.6	1,067.2	1,068.0	1,070.1	1,070.8	1,075.2
		Ridge estimators					
	$\hat{\theta}_1$	1,082.5	1,100.4	1,103.7	1,103.1	1,103.1	1,105.4
	$\tilde{\theta}_W$	1,082.3	1,100.6	1,103.8	1,103.1	1,103.3	1,105.5
		$\mathbf{p}$ (reference)					
Z, X	$R = 0$	1,431.8	1,440.9	1,443.0	1,445.7	1,449.0	1,449.1
$[Z]$	$R = 1$	1,431.3	1,440.3	1,442.3	1,445.1	1,448.4	1,448.4
		$\mathbf{p}$ (reference)					
	$R = 0$	1,397.2	1,403.1	1,403.9	1,405.5	1,408.2	1,408.1
	$R = 1$	1,378.6	1,384.9	1,385.5	1,386.9	1,389.4	1,389.3
		$2\mathbf{p}$					
	$R = 0$	1,396.7	1,402.6	1,403.3	1,404.9	1,407.6	1,407.4
	$R = 1$	1,378.5	1,384.8	1,385.4	1,386.7	1,389.2	1,389.1
Z, X, U		$\frac{1}{2}\mathbf{p}$					
$[Z, X]$	$R = 0$	1,398.5	1,404.7	1,405.5	1,407.2	1,409.9	1,409.9
	$R = 1$	1,379.1	1,385.6	1,386.2	1,387.6	1,390.1	1,390.1
		Ridge estimators					
	$\hat{\theta}_1$	1,445.4	1,461.8	1,463.9	1,468.6	1,473.5	1,479.5
	$\tilde{\theta}_W$	1,426.6	1,446.8	1,448.7	1,452.7	1,457.8	1,463.6

Figure 5.1 shows that rMSE depends on R and the factor c in the priorities $c\mathbf{p}$ only weakly, much less than it depends on σ. For R much smaller than unity, the bias is a nontrivial fraction of the rMSE but the gradient of the bias is to a large extent ameliorated in rMSE. For example, if the sampling variance of $\hat{\theta}$ were equal to $1{,}065^2$ for all $R \in (0,1)$ and the bias decreased from 100 at $R = 0$ to zero at $R = 1$, then the rMSE would attain values 1,069.7 and 1,065.0 at $R = 0$ and 1, respectively, that is, 4.7 apart. The corresponding empirical values, 1,065.6 and 1,064.9, are only 0.7 apart.

Figure 5.1 Bias and root-MSE as functions of the parameter R for the reference setting and perturbation σ = 0, 0.02, …, 0.1, indicated at the right-hand margin; settings with the reference values of p (solid black lines), 2p (long grey dashes) and $\frac{1}{2}$p (short grey dashes). Calibration on Z and X.



We compare our calibration estimators with two alternatives based on ridge regression,

$$\hat{\theta}_1 = \mathbf{t}^\top \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$
$$\tilde{\theta}_W = \mathbf{t}^\top \left(\lambda \mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y},$$

where $\mathbf{W}$ is the diagonal matrix with $\mathbf{w}$ on its diagonal and $\lambda = n/N$. We selected $\hat{\theta}_1$ because it is a component in the decomposition of $\hat{\theta}$ in (2.3); $\tilde{\theta}_W$ is motivated by Chambers (1996).

These ridge estimators have much greater biases, in the range 300 – 500, but their rMSE's are greater than the rMSE's of our calibration estimators by much smaller margins; the rMSE's of $\hat{\theta}_1$ and $\tilde{\theta}_W$ are in the range 1,082.3 – 1,105.5 for $\sigma \in$ (0, 0.1); details are given at the bottom of the first block of Table 5.1.

For any given setting, the rMSE's of $\hat{\theta}_1$ and $\tilde{\theta}_W$ differ by less than 0.1. The weighted estimator $\mathbf{w}^\top \mathbf{y}$ and the Hájek estimator $N\mathbf{w}^\top \mathbf{y} \,/ \mathbf{w}^\top \mathbf{1}$ have small biases but their rMSE's are far greater, exceeding 2,000.

By increasing p_3 from 0.15 to 0.25, as suggested in the discussion of Figure 4.1, the bias and rMSE functions are altered only slightly. The bias is reduced, by up to 0.05 but, contrary to expectation, the rMSE is increased for all values of σ, although by no more than 0.1. When p_0 is increased from 1.5 to 15, with p_3 set to 0.15, the bias and rMSE are altered imperceptibly, by less than 0.01.

In conclusion, Figure 5.1 suggests that calibration increases the efficiency of $\hat{\theta}$, and the estimator is not very sensitive to the choice of the tuning parameters, even though the discrepancies $\boldsymbol{\delta}$ are sensitive to the choice. In all settings we used, $R = 1$ is the optimal choice; that is, the criterion of small change $\|\mathbf{u} - \mathbf{w}\|$ should be ignored. It also results in small bias, much smaller than for $R = 0$. This does not apply in general. Positive S, that is, $R < 1$, is useful when the weights are much more dispersed and not highly correlated with the outcome. In small samples, variance reduction is relatively more important than bias reduction, and then S should be set to a larger positive value.

The value of a margin t_k can be quantified by calibration with the corresponding variable X_k omitted, or assigned very small priority. By dropping the margin t_X from calibration the rMSE functions are inflated substantially, to the range (1,431.8, 1,449.1) for $\sigma \in (0, 0.1)$ at $R = 0$ and are smaller by between 0.5 and 0.6 at $R = 1$. Details are given in the middle block of Table 5.1.

The sampling design and calibration setting explored so far are unrealistically congenial in that the base weights $\mathbf{w}$ depend only on the variables for which the population totals in $\mathbf{t}$ are available and therefore $\hat{\theta}_3 = 0$ in equation (2.3). We generate a more realistic setting by including a variable U correlated with the outcome variable Y, and define design probabilities that depend, apart from Z, also on U. In particular, $\mathbf{U}$ is generated as a random sample from the log-normal distribution based on $\mathcal{N}(1, 0.4)$, so that its mean is 2.94 and standard deviation is 1.23. Further, the design probabilities are set to

$$\boldsymbol{\pi}_U = \boldsymbol{\pi}_0 + \frac{0.004\,(\mathbf{U} - \bar{U})}{\max(\mathbf{U}) - \min(\mathbf{U})},$$

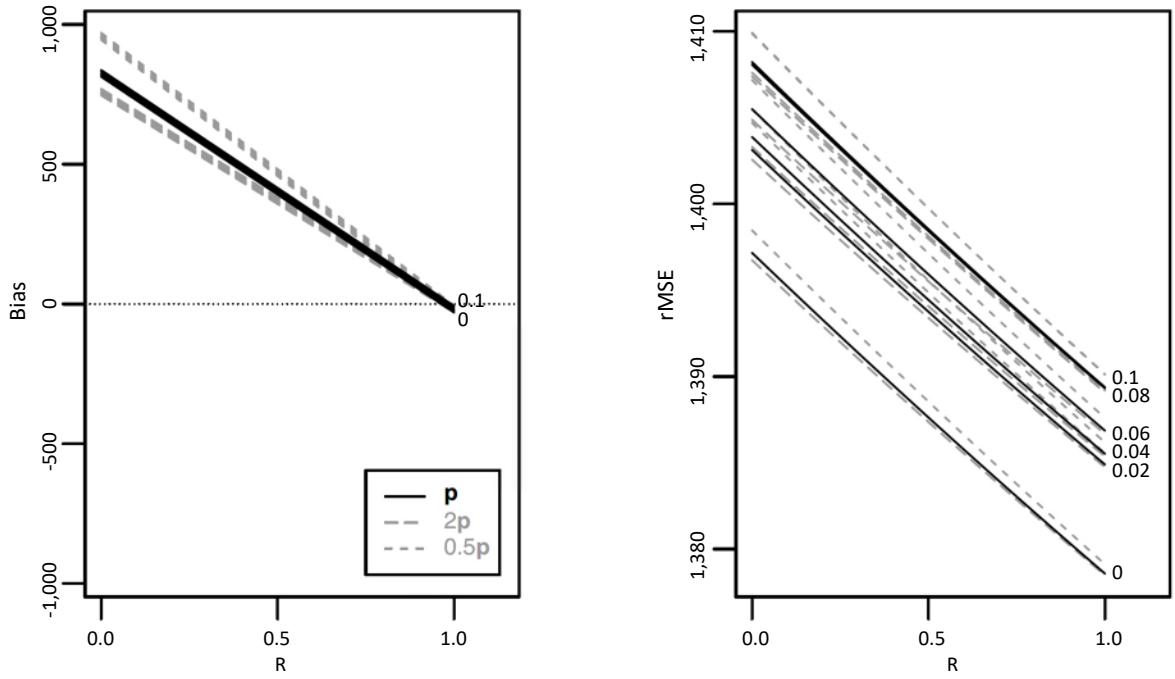
where $\boldsymbol{\pi}_0$ are the design probabilities used thus far. The probabilities $\boldsymbol{\pi}_U$ are standardised to have population total equal to 1,000. The "new" outcome variable is set to $Y + \frac{1}{5}U$. The corresponding target is $(\mathbf{Y} + \frac{1}{5}\mathbf{U})^\top \mathbf{1}_N$ = 691,516.8. The design probabilities $\boldsymbol{\pi}_U$ scaled to yield a sample of size $n \doteq$ 1,000, have standard deviations around 2.0×10^{-4} within the six categories of Z. (The average probability is $n/N = 8.33 \times 10^{-4}$).

Figure 5.2 presents the results using the same layout as Figure 5.1. It shows that the rMSE functions attain far greater values than in the setting without variable U, the rMSE functions attain their minima for $R = 1$ but the slopes of these functions are much steeper than with the original setting.

The ridge estimators $\hat{\theta}_1$ and $\tilde{\theta}_W$ are less efficient than $\hat{\theta}$, although $\tilde{\theta}_W$ is now discernibly more efficient than $\hat{\theta}_1$; their rMSE's differ by between 14 (for $\sigma = 0.1$) and 19 ($\sigma = 0$); see the bottom of the last block in Table 5.1. The bias of $\hat{\theta}_1$ is in the range (330, 370), comparable to the bias of $\hat{\theta}$ at $R = 0.6$. The bias of

$\tilde{\theta}_{\mathrm{W}}$ is much smaller, in the range $(-15, 24)$. Clearly, not involving the weights $\mathbf{w}$ is a handicap of $\hat{\theta}_{1}$ when the regression model $(y \mid X)$ is deficient.

Figure 5.2 Bias and root-MSE as functions of the parameter R for the reference setting and perturbation $\sigma = 0, 0.02, \ldots, 0.1$, indicated at the right-hand margin. Settings with the reference values of p (solid black lines), 2p (long grey dashes) and $\frac{1}{2}$p (short grey dashes). Calibration on Z, X and U.



The simulations provide evidence that our calibration estimator $\hat{\theta}$ is more efficient than the two ridge estimators $\hat{\theta}_{1}$ and $\tilde{\theta}_{\mathrm{W}}$, but the gains in efficiency are rather modest. Although the priorities $\mathbf{p}$ can be set with a view to minimising the discrepancies in $\boldsymbol{\delta}$, a careful choice is not always rewarded by greater efficiency. However, bias and efficiency are fairly insensitive to the setting of $\mathbf{p}$. We obtained very similar results using designs with fixed sample size and stratification on Z.

We have identified two factors that have a strong impact on the efficiency of $\hat{\theta}$ – the imperfection in how the sampling design is implemented, governed in the simulations by the parameter σ, and the completeness of vector $\mathbf{t}$. That is, efficiency is enhanced by good implementation of the sampling design and by calibrating on all the variables on which the sampling design is based, or by constructing the base weights using only variables with known population totals. Conclusions based on our simulations do not warrant a generalisation to settings with much more extensive calibration vector $\mathbf{t}$ and complex sampling designs.

6. Discussion and conclusion

We have introduced a method of calibration that matches the prescribed margins (population and subpopulation sizes and totals) subject to small discrepancies, while attending to the concern of efficiency. The imperative of an exact match and the dichotomy of matching or ignoring each available margin is replaced by a set of priority coefficients that quantify the importance or urgency of matching each margin, together with limiting the adjustment of the base weights and preference for less dispersed weights. The algorithm for calibrating the weights entails no iterations, nor handling any large matrices, and so it can be applied multiple times, searching for the best compromise of the competing constraints.

The calibration estimator $\hat{\theta}$ is related to a ridge regression predictor. We derived two decompositions of this estimator, one to the sum of a ridge regression predictor that does not involve the sampling weights, a bias adjustment that also has the form of ridge prediction, and a residual term that vanishes in some congenial settings. The other decomposition leads to a linear interpolation between the values of $\hat{\theta}$ for $R = 0$ and 1, both of which involve simple expressions.

Our exploration, analytically and by simulations, indicates that reduction of an absolute discrepancy $|\delta_k|$ is achieved by a small alteration of the priority p_k while the discrepancy is large, but p_k has to be increased substantially when δ_k is close to zero. This suggests that relaxed calibration applied after a careful exploration of the settings for the priority coefficients may be much more constructive than the binary choice of matching exactly or ignoring the available margin. In our experiments with relatively simple settings, we showed that the discrepancies $\boldsymbol{\delta}$ are easy to control but refined control is not necessarily rewarded by greater efficiency of the calibration estimator. However, the efficiency is quite insensitive to the detailed setting of the vector of priorities $\mathbf{p}$ and R.

We identified three factors that influence efficiency of $\hat{\theta}$ strongly. The first, the level or extent of the imperfection in how the sampling design is implemented comes as no surprise, even though calibration is meant to address this problem. It may do so to some extent, but does not compensate for it fully. The second is the availability of the population totals for the variables on which the design weights $\mathbf{w}$ are based. The third is the residual variance of the linear model relating the outcome y to the variables in $\mathbf{X}$. Of course, these conclusions have to be confirmed in more complex settings, sampling designs and information about population totals.

Our method of calibration is entirely model-free but it does not preclude adaptations that involve models. For example, the base weights $\mathbf{w}$ may be first adjusted by a model-based or model-assisted method, and the resulting weights subjected to relaxed calibration. The priorities $\mathbf{p}$ (and maybe also the coefficient R) may be set by considerations related to models, in addition to the expert judgement that we presumed originally.

The decompositions we derived involve ridge regression coefficients and predictors based on implied models. An avenue to explore in the future is whether and when it is advantageous to replace these terms using model-based considerations. Another challenge is to devise ways of incorporating known or

conjectured properties of the outcome variable(s) in the calibration. In regularly conducted surveys, they may be based on insights gained from their previous rounds. In some simple settings we have identified the correlation of the weights and outcomes as an important factor. We conjecture that these correlations are important also in more complex settings.

Adaptation of the presented method to settings with imprecise auxiliary information (West and Little, 2012; Opsomer and Erciulescu, 2021) is another outstanding challenge. The chi-squared distance for the weights used in optimisation may be replaced by alternatives, possibly at the price of a more complex (iterative) algorithm. One exception is the chi-squared distance for the log-weights (or another transformation of the weights), which may be motivated by reference to multiplicative adjustments of the weights.

We set out with no intention to posit a model for the outcome variable but derived an estimator that is related to a model. This suggests that we should reduce the emphasis on model validity in a model-based approach and focus on exploiting all the available information. The offset in the ridge regression to which the calibration estimator is related can be considered similarly, and set by the considerations of the trade-off of bias and variance. Maybe the committment to an approach or paradigm is inferior to their well-devised combination or compromise, exploiting the strengths of each and ameliorating their weaknesses.

All the computing described in this article was accomplished in R. The code developed can be obtained from the author on request.

Appendix

A. Calibration on a single categorical variable

This appendix derives an expression for the estimator $\hat{\theta}$ when calibrating only on the population size t_0 and the subpopulation sizes t_k of the categories $k = 1, \ldots, K$ of a discrete variable Z. We assume that the probability that a category does not appear in the sample is negligible. The starting point is equation (2.4), according to which we have to evaluate $\mathbf{w}^\top \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{t}^\top\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}$ and $\mathbf{X}$ is matrix of auxiliary variables after its columns $1, \ldots, K$ are standardised.

The original $n \times (K+1)$ matrix of auxiliary data, denoted by $\mathbf{X}_{\text{or}}$, comprises $\mathbf{1}_n$ in column 0 and the indicator of category k in column $k = 1, \ldots, K$. Denote by $\boldsymbol{\mu} = (0, \mu_1, \ldots, \mu_K)^\top$ the vector of sample proportions of the categories of Z, preceded by a zero, $s_0^2 = 1$ and $s_k^2 = \mu_k(1-\mu_k)$ for $k = 1, \ldots, K$, and $\mathbf{S} = \text{diag}(1, s_1, \ldots, s_K)$. Then standardisation is the transformation $\mathbf{X} = (\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top)\,\mathbf{S}^{-1}$; it leaves column 0 intact. Since the other columns of $\mathbf{X}_{\text{or}}$ are pairwise orthogonal indicators, $\hat{\boldsymbol{\beta}} = (\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is expressed in terms of $\mathbf{X}_{\text{or}}$, $\boldsymbol{\mu}$ and $\mathbf{S}$ as

$$\hat{\boldsymbol{\beta}} = \mathbf{S}\left\{\mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\left(\boldsymbol{\mu}^{(1)}\right) - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right\}^{-1}\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)^\top\mathbf{y},$$

where $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu} + \mathbf{e}_0$; element 0 of $\boldsymbol{\mu}^{(1)}$ is equal to 1 and its other elements are μ_k. Denote $\boldsymbol{\Pi} = \mathbf{S}\mathbf{P}^{-1}\mathbf{S} + n\,\text{diag}\,(\boldsymbol{\mu}^{(1)})$; its diagonal elements are $1/p_0 + n$ and $s_k^2/p_k + n\mu_k$, $k = 1, \ldots, K$. The inverse involved in $\hat{\boldsymbol{\beta}}$ is

$$\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1} = \boldsymbol{\Pi}^{-1} + \frac{n}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}$$

where, since $\boldsymbol{\mu}^\top\mathbf{1} = 1$,

$$\begin{aligned} G &= 1 - n\,\boldsymbol{\mu}^\top\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \sum_{k=1}^{K}\left(\mu_k - \frac{n\,p_k\mu_k^2}{\mu_k(1-\mu_k) + n\,p_k\mu_k}\right) = \sum_{k=1}^{K} g_k s_k^2, \end{aligned}$$

and $g_k = 1/(1 - \mu_k + n\,p_k)$, $k = 1, \ldots, K$.

Denote $\bar{\mathbf{Y}} = (\bar{y}, \bar{y}_1\,\mu_1, \ldots, \bar{y}_K\mu_K)^\top$, where $\bar{y}$ is the sample mean of y and $\bar{y}_k$ the subsample mean of y in category k of Z. Define $\bar{\mathbf{W}} = (\bar{w}, \bar{w}_1\,\mu_1, \ldots, \bar{w}_K\mu_K)^\top$ similarly. We have the identities $(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top)^{-1}\boldsymbol{\mu} = \frac{1}{G}\,\boldsymbol{\Pi}^{-1}\boldsymbol{\mu}$ and $\mathbf{a}^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} = \sum_{k=1}^{K} a_k p_k g_k$ for any vector $\mathbf{a} = (a_0, a_1, \ldots, a_K)^\top$. Therefore

$$\begin{aligned} \mathbf{w}^\top\mathbf{X}\hat{\boldsymbol{\beta}} &= n\,\mathbf{w}^\top\left(\mathbf{X}_{\text{or}} - \mathbf{1}_n\boldsymbol{\mu}^\top\right)\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &= n^2\left(\bar{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &\quad + \frac{n^3}{G}\left(\bar{\mathbf{W}} - \bar{w}\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \times \left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right)^\top\boldsymbol{\Pi}^{-1}\boldsymbol{\mu} \\ &= \frac{n^2 p_0}{1 + n\,p_0}\bar{w}\,\bar{y} + n^2\sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{w}_k - \bar{w}\right)\left(\bar{y}_k - \bar{y}\right) \\ &\quad + \frac{n^3}{G}\sum_{k=1}^{K} g_k p_k\,\mu_k\left(\bar{w}_k - \bar{w}\right) \times \sum_{k=1}^{K} g_k p_k \mu_k\left(\bar{y}_k - \bar{y}\right). \end{aligned}$$

(The sign $\times$ is added to emphasise that the multiplication is applied to two scalars.) The same sequence of operations yields the identity

$$\begin{aligned} \mathbf{t}^\top\hat{\boldsymbol{\beta}} &= n\left(\mathbf{t}_{\text{or}} - t_0\boldsymbol{\mu}\right)^\top\left(\boldsymbol{\Pi} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)^{-1}\left(\bar{\mathbf{Y}} - \bar{y}\,\boldsymbol{\mu}\right) \\ &= \frac{n\,p_0}{1 + n\,p_0}t_0\,\bar{y} + n t_0\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right)\left(\bar{y}_k - \bar{y}\right) \\ &\quad + \frac{n^2 t_0}{G}\sum_{k=1}^{K} g_k p_k\left(\frac{t_k}{t_0} - \mu_k\right) \times \sum_{k=1}^{K} g_k p_k\left(\bar{y}_k - \bar{y}\right). \end{aligned}$$

Hence, after substituting to (2.4), $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$, where

$$\begin{aligned}\hat{\theta}^{(1)} &= R\left(\mathbf{w}^\top\mathbf{y} - \frac{n\,p_0}{1+n\,p_0}\,n\,\overline{w}\,\overline{y}\right) + \left(1 - \frac{R}{1+n\,p_0}\right) t_0\,\overline{y} \\ \hat{\theta}^{(Z)} &= n\sum_{k=1}^{K} g_k p_k \Lambda_k \left(\overline{y}_k - \overline{y}\right) + \frac{n^2}{G}\sum_{k=1}^{K} g_k p_k \Lambda_k \times \sum_{k=1}^{K} g_k p_k \left(\overline{y}_k - \overline{y}\right)\end{aligned} \tag{6.1}$$

and $\Lambda_k = t_k - n\,R\,\overline{w}_k - \mu_k(t_0 - n\,R\,\overline{w}) = t_k - t_0\mu_k - nR(\overline{w}_k - \overline{w}\mu_k)$. Note that $\hat{\theta}^{(1)}$ depends on $\boldsymbol{\mu}$, $\mathbf{p}$ and $\mathbf{t}$ only through p_0 and t_0, whereas $\hat{\theta}^{(Z)}$ does not depend on p_0. In fact, $\hat{\theta}^{(Z)} = 0$ when calibrating only on the population size, when $p_1 = \ldots = p_K = 0$. We have $\hat{\theta}^{(Z)} = 0$ also when $\Lambda_1 = \ldots = \Lambda_K = 0$. An example of such a reduction arises when the weights are constant, $\mathbf{w} = \overline{w}\,\mathbf{1}_n$ and the subsample sizes within the K categories are fixed, $t_k = t_0\,\mu_k$. In general, $\hat{\theta}^{(Z)}$ can be regarded as an adjustment of $\hat{\theta}^{(1)}$ for the priorities associated with Z.

Estimator $\hat{\theta} = \hat{\theta}^{(1)} + \hat{\theta}^{(Z)}$ depends on $\mathbf{w}$ and $\mathbf{y}$ only through $\mathbf{w}^\top\mathbf{y}$ and the K pairs of within-category means $\overline{w}_k$ and $\overline{y}_k$. The product $g_k p_k = p_k/(1-\mu_k + np_k)$ is an increasing function of p_k with its respective limits of 0 at $p_k = 0$ and $1/n$ as $p_k \to +\infty$. The denominator G is a decreasing function of each p_k. It attains its maximum of 1 for $p_1 = \ldots = p_K = 0$ and its limit as all K priorities p_k diverge to $+\infty$ is zero. In the latter case $\hat{\theta}$ becomes unstable. However, in practice p_0 is set higher than the other priorities, so this case is of little practical relevance.

As p_0 diverges to $+\infty$, $\hat{\theta}^{(1)}$ converges to

$$R\,(\mathbf{w}^\top\mathbf{y} - n\,\overline{w}\,\overline{y}) + t_0\,\overline{y} = (n-1)\,R\,\mathrm{cov}\,(w, y) + t_0\,\overline{y}\,.$$

This confirms that the weights are especially important when they are highly correlated with the outcome variable.

B. $R = 0$ implies $\delta_0 = 0$

In the decomposition $\mathbf{w}_R = \mathbf{X}\mathbf{v}_R + \boldsymbol{\varepsilon}_R$ for $\mathbf{w}_0$, we have $\mathbf{v}_0 = \mathbf{e}_0$ and $\boldsymbol{\varepsilon}_R = \mathbf{0}$. Hence

$$\mathbf{u} = \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{t} + \frac{t_0}{n}\,\mathbf{P}^{-1}\mathbf{e}_0\right).$$

Owing to standardisation, $\mathbf{X}^\top\mathbf{1}_n = n\,\mathbf{e}_0$, and column 0 of $\mathbf{X}^\top\mathbf{X}$ is also equal to $n\,\mathbf{e}_0$. Further, $\mathbf{P}^{-1}\mathbf{e}_0 = p_0^{-1}\mathbf{e}_0$, and so

$$\begin{aligned}\mathbf{u}^\top\mathbf{1} &= \left(\mathbf{t} + \frac{t_0}{np_0}\,\mathbf{e}_0\right)^\top\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{1} \\ &= \frac{n}{p_0^{-1} + n}\left(\mathbf{t} + \frac{t_0}{np_0}\,\mathbf{e}_0\right)^\top\mathbf{e}_0 = \mathbf{t}_0.\end{aligned}$$

Therefore $\delta_0 = \mathbf{u}^\top \mathbf{1} - t_0 = 0$. We explore when the slope of δ_0, a linear function of R, vanishes. According to equation (3.2), this slope is

$$\begin{aligned}
\frac{\partial \delta_0}{\partial R} &= \mathbf{e}_0^\top \mathbf{X}^\top \left\{ \mathbf{I}_n - \mathbf{X} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \right\} \left(\mathbf{w} - \frac{t_0}{n} \mathbf{1}_n \right) \\
&= \mathbf{e}_0^\top \left\{ \mathbf{I} - \mathbf{X}^\top \mathbf{X} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \right\} \mathbf{X}^\top \left(\mathbf{w} - \frac{t_0}{n} \mathbf{1}_n \right) \\
&= \mathbf{e}_0^\top \mathbf{P}^{-1} \left(\mathbf{P}^{-1} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \left(\mathbf{X}^\top \mathbf{w} - \frac{t_0}{n} \mathbf{X}^\top \mathbf{1}_n \right) \\
&= \frac{1}{p_0} \frac{1}{p_0^{-1} + n} \mathbf{e}_0^\top \left(\mathbf{X}^\top \mathbf{w} - t_0 \mathbf{e}_0 \right) \\
&= \frac{1}{1 + n p_0} \left(\mathbf{w}^\top \mathbf{1} - t_0 \right).
\end{aligned}$$

Hence $\delta_0 = 0$ when $R = 0$ or $\mathbf{w}^\top \mathbf{1} = t_0$, as stated at the end of Section 3. Note that in most surveys with complex sampling design and nontrivial (random) sample size the probability that $\mathbf{w}^\top \mathbf{1} = t_0$ is very small or zero.

References

Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 5-20.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 323-353.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

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

Dagdoug, M., Goga, C. and Haziza, D. (2023). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*, 50, 761-785.

Davies, G. (2018). *Examination of Approaches to Calibration in Survey Sampling.* PhD. thesis, Cardiff University, UK.

Devaud, D., and Tillé, Y. (2019). Deville and Särndal's calibration: Revisiting a 25-years-old successful optimization problem. *Test*, 28, 1033-1065.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association*, 87, 1013-1020.

Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.

Goga, C., and Shehzad, M.A. (2010). Overview of ridge estimators in survey sampling. Université de Bourgogne, Dijon, France.

Guggemos, F., and Tillé, Y. (2010). Penalized calibration in survey sampleing: Design based estimation assisted by mixed model. *Journal of Statistical Planning and Inference*, 140, 3199-3212.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Lohr, S., and Raghunathan, T. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Longford, N.T. (2024). Statistical balancing as an unconstrained optimisation problem. Submitted.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Opsomer, J.D., and Erciulescu, A.L. (2021). Replication variance estimation after sample-based calibration. *Survey Methodology*, 47, 2, 265-277. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00006-eng.pdf.

Rao, J.N.K., and Singh, A.C. (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics*, 25, 371-383.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 2, 99-119. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10488-eng.pdf.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: John Wiley & Sons, Inc.

Vera, J.F., Sánchez Zuleta, C.C. and Rueda, M. (2023). A unified approach based on multidimensional scaling for calibration estimation in surveys with qualitative auxiliary information. *Statistical Methods in Medical Research*, 32, 760-772.

West, B.T., and Little, R.J.A. (2012). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society*, Series A, 176, 211-225.

Wu, C., and Lu, W.W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, 84, 21-39.