Zhang, Li-Chun

Working Paper — Digitized Version

Post-Stratification and Calibration - A Synthesis

Discussion Papers, No. 216

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Zhang, Li-Chun (1998) : Post-Stratification and Calibration - A Synthesis, Discussion Papers, No. 216, Statistics Norway, Research Department, Oslo

This Version is available at:
http://hdl.handle.net/10419/192200

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.
Statistics Norway
Research Department

Li-Chun Zhang

Post-Stratification and Calibration
- A Synthesis
Li-Chun Zhang

Post-Stratification and Calibration
– A Synthesis

Abstract:
The paper offers a synthesis of several widely used estimation methods in survey sampling from a rather personal point of view. The methods which will be discussed include post-stratification estimation, generalized regression estimation and calibration estimation. The presentation puts emphasis on understanding, with as little mathematics as possible. It is hoped that in this way anybody, with however varied background of experience with these methods, may find something useful here. The appendix introduces a program package called CALWGT for calibration, which is available on contacting the author.

Keywords: Post-stratification, generalized regression, calibration.

Acknowledgement: Thanks to Leiv Solheim who has taken the initiative for this work, and Ib Thomsen for inspiring discussions. Special thanks to Jan F. Bjørnstad who has read several earlier versions and contributed with detailed and constructive comments.

Address: Statistics Norway, Division of Statistical Methods and Standards, P.O.Box 8131 Dep., N-0033 Oslo. E-mail: lcz@ssb.no
Discussion Papers comprises research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborated than a usual article by including intermediate calculation and background material etc.

Abstracts with downloadable postscript files of Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
P.O. Box 1260
N-2201 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Saig-abonnement@ssb.no
1 Introduction

Broadly speaking, post-stratification refers to any method of data analysis which involves forming units into homogeneous groups after the sample has been taken (Holt and Smith, 1979, Smith 1991). Typically, however, the term is restricted to those cases where auxiliary information external to the sample is available in addition. As such post-stratification is a central concept in survey sampling. It induces a structure to the population according to the auxiliary information, on which many of the standard methods are based including post-stratified estimation, generalized regression estimation and calibration estimation.

We explain all these methods from such a synthetic point of view. All of them are more or less a special case of calibration, and all of them are based on post-stratification. Indeed, post-stratification is the finest calibration and calibration the relaxed post-stratification. Throughout, we assume that the estimation aims at some population total, and that the estimator is of the linear class.

In addition, the appendix describes a program package CALWGTT for calibration written in S-Plus for Unix.

2 Post-stratification and post-stratified estimation

We shall distinguish between post-stratification and post-stratified estimation. While the former defines a structure of the population according to the auxiliary information, the latter refers to a special way in which this structure is utilized for estimation purposes.

2.1 Post-stratification

Denote by $y$ the object variable of the survey and by $x$ the auxiliary variable, both may possibly be vector-valued. Denote by $U$ the population of the size $N$, i.e. $U = 1, \ldots, N$, and by $i$ the unit index. Post-stratification is carried out w.r.t. $x$ after the sample has been collected, which divides the population into, say, $H$ disjoint (population) post-strata, i.e. $U = \bigcup_{h=1}^{H} U_h$ where $U_h \cap U_g = \emptyset$ for $h \neq g$. Meanwhile, applying post-stratification to the sample, denoted by $s$, gives rise to sample post-strata $(s_1, \ldots, s_H)$.

The post-stratification introduces the structural transition from $(s, U)$ to $\{(s_1, U_1), \ldots, (s_H, U_H)\}$, which allows us to think of $s_h$ as a sample taken from the homogeneous sub-population $U_h$. 
2.2 Post-stratified estimation

Post-stratification gives us \( Y = \sum_{i \in U} y_i = \sum_{h=1}^{H} Y_h = \sum_h (\sum_{i \in U_h} y_i) \). Given the knowledge of the distribution of the population post-strata, denoted by \( p_h = N_h/N \) where \( N_h \) is the size of the \( h \)th population post-strata, and that none of the sample post-strata is empty, the post-stratified estimator for \( Y \) is of the form \( \hat{Y}_{pst} = \sum_h \hat{Y}_h \), i.e. estimating \( Y_h \) based on \( s_h \) and taking summation over \( s_1, ..., s_H \). Notice that \( p_h \), though implicit, is necessary for constructing \( \hat{Y}_h \).

Estimator \( \hat{Y}_h \) differs according to whether the inclusion probability, denoted by \( \pi_i \), is constant or not within each \( U_h \). In case \( \pi_i = \pi_h \) for \( i \in s_h \), \( Y_h \) is estimated by means simple expansion, i.e.

\[
\hat{Y}_{pst} = \sum_h \hat{Y}_h = \sum_h N_h \hat{y}_h = \sum_h (N_h/n_h) \sum_{i \in s_h} y_i = \sum_h w_h \sum_{i \in s_h} y_i, 
\]

where \( n_h \) is the size of the \( h \)th sample post-stratum. We call this the simple post-stratified estimator.

Under some complex design where \( \pi_i \) differs within each post-stratum, an unbiased estimator of \( Y_h \) is given by the Horvitz-Thompson estimator within the post-stratum, i.e. \( \hat{Y}_h = \sum_{i \in s_h} y_i/\pi_i \). However, the suggested estimator in such cases (Smith, 1991), the so-called Hajek estimator, applies a ratio estimator within each \( U_h \) instead, i.e.

\[
\hat{Y}_h = N_h (\hat{Y}_h/\tilde{N}_h) = (N_h \hat{Y}_h)/\left(\sum_{i \in s_h} 1/\pi_i\right) = N_h (\sum_{i \in s_h} y_i/\pi_i)/\left(\sum_{i \in s_h} 1/\pi_i\right). 
\]

The weight for \( i \in s_h \) is now \( N_h (1/\pi_i)/(\sum_{i \in s_h} 1/\pi_i) \). The reason is that \( \hat{Y}_h/\tilde{N}_h \) is often more efficient for the post-stratum mean than \( \hat{Y}_h/N_h \) even when \( N_h \) is known (Särndal, Swensson and Wretman, 1992, Section 5.7).

2.3 Discussion

The main theoretical problem of the post-stratified estimation is conditioning. Post-stratification, according to Holt and Smith (1979), implies that the properties of an estimator for \( Y \) should be evaluated conditional to the realized sample configuration of the post-strata, i.e. \((n_1, ..., n_H)\). This is particularly convincing in case of the simple post-stratified estimator, which serves as the primary example of post-stratified estimation. Difficulties arise, however, when dealing with complex designs, because \( \{\pi_i; i \in s_h\} \) is not fixed when conditioning on \( n_h \) alone, and its distribution easily becomes untraceable (Rao, 1985).
Consider, for instance, stratified simple random sampling where post-stratification cuts across the stratification. Given categorical auxiliary variable, this is a common situation where such difficulties arise. However, whenever \( p_h \) indeed is based on some population register, it is in principle possible to combine this register with that from which the sample was drawn. In other words, post-stratification can be extended to include stratum index as an additional auxiliary variable, since the combined register would provide the necessary \( N_h \). For the general case, thus, the solution would be to include \( \pi_i \) as an additional auxiliary variable, followed by post-stratification in the usual way.

The practical problem of this approach, as well as for the post-stratified estimator at large, is the resulting empty sample post-strata. Another side of this problem is that the totals of the population post-strata may not always be available/reliable. Post-stratified estimation which ignores the empty sample post-strata is downward biased for non-negative \( y_i \) as noted by e.g. Jagers (1986). A few exceptions apart (Fuller, 1966), calibration estimation (Deville and Särndal, 1992; Deville, Särndal, and Sautory, 1993) provides an alternative general methodology.

3 Post-stratification and calibration (I)

3.1 Calibrating post-stratification

The weights for the given sample, i.e. \( \{w_i; i \in s\} \), are said to be calibrated w.r.t. a set of known totals in the population, if the estimates based on \( \{w_i; i \in s\} \) reproduce these totals. Given categorical auxiliary variable, such totals are typically the sizes of the various domains of the population. Indeed, from the calibration point of view, the post-stratified estimator should first of all be calibrated w.r.t. the sizes of the population post-strata, i.e. \( N_h = \sum_{i \in s_h} w_i \) for \( 1 \leq h \leq H \), which is true for the simple post-stratified estimator and the Hajek estimator, but not the Horvitz-Thompson estimator.

In particular, whenever the post-stratification has used up all the auxiliary information available, it must also define the finest division of domains w.r.t. whose totals calibration can be carried out. In other words, the set of calibration totals, denoted by \( T \), can only be taken from

\[ \zeta(1, ..., H) = \{t; t = \sum_{h \in R} N_h \} \text{ if } R \subseteq \{1, ..., H\} \}

Thus, if an estimator is calibrated w.r.t. \( (N_1, ..., N_H) \), it is necessarily so for any \( T \subseteq \zeta \).

Technically speaking, in case of empty sample post-strata, calibration avoids collapsing post-strata provided each population total of the empty sample post-strata is built into
more than one calibration totals. As a simplest case, assume non-empty sample post-
strata except from \( s_1 \). Since none of the sample units comes from \( U_1 \), calibration w.r.t. \( N_1 \) is impossible, i.e. \( N_1 \notin T \). To collapse \( U_1 \) into some other post-strata means, (a) a bipartition of \( T \) as \((T_1, T_2)\), (b) a choice of some \( g \in \{2, \ldots, H\} \) and let \( T_1 = N_1 + N_g \), and (c) letting \( T_2 = \{N_h; h \in \{1, g\}^c\} \). On the other hand, one could also let \( N_1 \) contribute to more than one of the components of \( T \subseteq \zeta(1, \ldots, H) \). For instance, let \( T = (T_1, T_2) \) where \( T_1 = (N_1 + N_2, N_1 + N_3) \) and \( T_2 = \zeta(4, \ldots, H) \). Since the calibrated weights satisfy \( N_1 + N_2 = \sum_{i \in s_2} w_i \) as well as \( N_1 + N_3 = \sum_{i \in s_3} w_i \), both units from \( s_2 \) and \( s_3 \) will now account for \( s_1 \), and no collapsing post-strata is needed. Moreover, in case \((N_2, N_3)\) are built into \( T_2 \) themselves, i.e. \( T_2 \subseteq \zeta(2, \ldots, H) \), more post-strata will be involved — the effect is sent down in a domino-motion.

**Remark 1** Calibration is sometimes known as the generalized raking. It resembles the method of raking in that both satisfy the known population marginal totals. Both avoid collapsing post-strata in case of empty sample post-strata, though the raking may become unstable or even fail to converge in such cases (OH and Scheuren, 1987). The difference occurs at the domain level, i.e. while raking is able to produce estimate for a post-stratum even if it is empty in the sample, this is never possible with calibration, or any linear estimator of the form \( \sum_{i \in s} w_i y_i \).

### 3.2 Dummy index: an example

Let post-stratification be based on auxiliary variable (a) Sex — (Men, Women) and denoted by \( x_1 = 0 \) or 1, (b) Civil Status I — (Married, Not-Married) and denoted by \( x_2 = 0 \) or 1, and (c) Civil Status II — (With Children, Without Children) and denoted by \( x_3 = 0 \) or 1. This gives rise to 8 post-strata, i.e. \((x_1, x_2, x_3) = (i, j, k)\) for \( i, j, k = 0, 1 \), where e.g. \((0, 0, 1)\) stands for "married men without children".

Dummy indexing of the post-strata for each sample unit consists of a vector of the same number of components as the number of post-strata, i.e. 8 in this case. Each component corresponds to a post-stratum, and takes value 1 if the unit belongs to this post-stratum and 0 otherwise. In this way, the dummy index of the auxiliary variable is \( z_i = (1, 0, 0, 0, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), \ldots, (0, 0, 0, 0, 0, 0, 0, 1) \), depending on which post-stratum the unit belongs to. Notice that the sum of the components of any vector is constant unity. In particular, using dummy indexing, calibration w.r.t. the post-strata totals can now be expressed as the *calibration equation*, i.e.

\[
T = \sum_{i \in s} w_i z_i \quad \Leftrightarrow \quad (N_1, \ldots, N_H) = \sum_h z_h (\sum_{i \in s_h} w_i) \quad \Leftrightarrow \quad N_h = \sum_{i \in s_h} w_i.
\]
Since the dummy indexing arises from crossing all the three auxiliary variables, it is sometimes shorthanded as “Sex × Civil Status I × Civil Status II” (Bethlehem and Wouter, 1987).

In general, dummy indexing for calibration w.r.t. $T$ refers to the arrangement of the binary vector for the sample units such that the calibration equation retains the form $T = \sum_{i \in s} w_i z_i$. It follows that such a dummy index would have the same number of components as that of $T$. Consider the next two illustrations.

Let first $T$ be the population marginal totals of $(x_1, x_2, x_3)$, i.e. the total of (a) Men, (b) Women, (c) Married, (d) Not-Married, (e) With Children and (f) Without Children — six of them in all. Dummy indexing each $x_j$, for $j = 1, 2, 3$, in the usual way gives us sub-vectors, say, $(0, 1)$ if $x_1 = 0$ and $(1, 0)$ if $x_1 = 1$, $(0, 1)$ if $x_2 = 0$ and $(1, 0)$ if $x_2 = 1$, and $(0, 1)$ if $x_3 = 0$ and $(1, 0)$ if $x_3 = 1$. Juxtapose the three sub-vectors leads to

$$(0, 1, 0, 0, 1, 1) \text{ if } (x_1, x_2, x_3) = (0, 0, 0), \quad (0, 1, 0, 1, 1, 0) \text{ if } (x_1, x_2, x_3) = (0, 0, 1),$$

$$(0, 1, 1, 0, 0, 1) \text{ if } (x_1, x_2, x_3) = (0, 1, 0), \quad (0, 1, 1, 1, 0, 1) \text{ if } (x_1, x_2, x_3) = (0, 1, 1),$$

$$(1, 0, 0, 1, 0, 1) \text{ if } (x_1, x_2, x_3) = (1, 0, 0), \quad (1, 0, 0, 1, 1, 0) \text{ if } (x_1, x_2, x_3) = (1, 0, 1),$$

$$(1, 0, 1, 0, 0, 1) \text{ if } (x_1, x_2, x_3) = (1, 1, 0), \quad (1, 0, 1, 1, 0, 1) \text{ if } (x_1, x_2, x_3) = (1, 1, 1).$$

Notice that the sum of the components of any vector no longer remains constant unity. In addition, the way in which the calibration totals here arise from the auxiliary variable will be referred to as natural, shorthanded as “Sex + Civil Status I + Civil Status II”.

Let now the calibration be defined w.r.t. the following marginal population totals: (a) Married Men, (b) Not-Married Men, (c) Married Women, (d) Not-Married Women, (e) Men With Children, (f) Men Without Children, (g) Women With Children, and (h) Women Without Children — eight of them in all. These can be shorthanded as “(Sex × Civil Status I) + (Sex × Civil Status II)”. Post-stratification according to (Sex, Civil Status I) leads to sub-vector $(1, 0, 0, 0)$ for $(x_1, x_2) = (0, 0)$, $(0, 1, 0, 0)$ for $(x_1, x_2) = (0, 1)$, $(0, 0, 1, 0)$ for $(x_1, x_2) = (1, 0)$, $(0, 0, 0, 1)$ for $(x_1, x_2) = (1, 1)$. Similarly, post-stratification according to (Sex, Civil Status II) leads to sub-vector $(1, 0, 0, 0)$ for $(x_1, x_3) = (0, 0)$, $(0, 1, 0, 0)$ for $(x_1, x_3) = (0, 1)$, $(0, 0, 1, 0)$ for $(x_1, x_3) = (1, 0)$, $(0, 0, 0, 1)$ for $(x_1, x_3) = (1, 1)$. Care needs to be taken so that the juxtaposition of the two sub-vectors is carried out consistently, i.e.

$$(1, 0, 0, 0, 1, 0, 0, 0) \text{ if } (x_1, x_2, x_3) = (0, 0, 0), \quad (0, 1, 0, 0, 1, 0, 0, 0) \text{ if } (x_1, x_2, x_3) = (0, 0, 1),$$

$$(0, 0, 1, 0, 0, 0, 1, 0) \text{ if } (x_1, x_2, x_3) = (0, 1, 0), \quad (0, 0, 1, 0, 0, 0, 1, 0) \text{ if } (x_1, x_2, x_3) = (0, 1, 1),$$

$$(1, 0, 0, 0, 0, 0, 1, 0) \text{ if } (x_1, x_2, x_3) = (1, 0, 0), \quad (0, 1, 0, 0, 0, 0, 0, 1) \text{ if } (x_1, x_2, x_3) = (1, 1, 0),$$

$$(1, 0, 0, 0, 0, 0, 1, 0) \text{ if } (x_1, x_2, x_3) = (1, 1, 0).$$
Finally, since the dummy indexing amounts to some one-to-one transformation of the auxiliary variable, we shall not make an effort to distinguish the two forms from now on. That is, we simply write \( x_i \) as the auxiliary vector of the \( i \)th unit, and \( X \) the corresponding totals in the population, in which way the calibration equation becomes now \( X = \sum_{i \in s} w_i x_i \). It also becomes clear that the calibration breaks down only if there are all zero-element columns in the sample auxiliary matrix, whose \( i \)th row is given by \( x_i \).

4 Calibration and generalized regression estimation

4.1 Linear calibration and generalized regression

The calibration equation alone, i.e. the choice of the calibration totals, is insufficient in determining the weights. Two more things are used: (a) a set of initial weights, denoted by \( \{ a_i; i \in s \} \), e.g. weights from the simple post-stratified estimator or the Horvitz-Thompson estimator, and (b) a metric function, denoted by \( G \), which measures the distance between \( \{ a_i; i \in s \} \) and the calibrated weights \( \{ w_i; i \in s \} \). Deville, Särndal, and Sautory (1993) chose \( r_i = w_i/a_i \) as argument of \( G \), and the measure of distance for the whole sample as \( \sum_{i \in s} a_i G(r_i) \). The idea is now to find \( \{ w_i \} \) which differs least from \( \{ a_i \} \) while subject to the calibration equation.

Let \( g = \partial G/\partial r \) be its first partial derivative. Let \( \lambda = (\lambda_1, ..., \lambda_J)^T \) be the Lagrange multiplier, we solve for \( \{ w_i; i \in s \} \),

\[
\frac{\partial}{\partial w_i} \{ \sum_{i \in s} a_i G(r_i) - (\sum_{i \in s} w_i x_i - X) \lambda \} = g(r_i) - x_i \lambda = 0.
\]

Denote by \( h(u) = g^{-1}(u) \), i.e. the inverse function of \( g \). The calibrated weights are then formally \( w_i = a_i h(x_i \lambda) \) where \( \lambda \) satisfies the calibration equation, i.e. \( X = \sum_{i \in s} a_i h(x_i \lambda) x_i \). Special attention has been paid to the so-called linear method where \( G = (r - 1)^2/2 \), which gives \( g = r - 1 \), and \( h(u) = 1 + u \), and the calibrated weights

\[
w_i = a_i (1 + x_i \lambda) = a_i \{ 1 + (X - \sum_{i \in s} a_i x_i)(\sum_{i \in s} a_i x_i^T x_i)^{-1} x_i^T \}.
\]

This is identical to generalized regression (GREG) estimation with \( \{ a_i; i \in s \} \) as weights (Bethlehem and Wouter, 1987; Lemaitre and Dufour, 1987). Though the GREG estimation was historically strongly motivated by empty post-strata, it does offer an alternative interpretation to the resulting estimator. For any finite population vector \( y = (y_1, ..., y_N)^T \) with auxiliary vector \( x_i \) for the \( i \)th unit, we make the transformation from \( y \) to \( \epsilon = (\epsilon_1, ..., \epsilon_N)^T \), i.e. \( \epsilon_i = y_i - x_i \beta \), through the vector \( \beta \) of the same dimension as
the auxiliary vector. In particular, the ordinary least-square fit based on the population
is defined as \( \beta = (x^T x)^{-1} x^T y \) where \( x \) is the auxiliary matrix whose \( i \)th row is set to \( x_i \).

Notice that the GREG estimator can thus be regarded as a linear adjustment of the
initial estimator based on \( \{a_i; i \in s\} \) (Särndal, Svensson, and Wretman, 1992, Chapter
6-7), after which the weights necessarily satisfy the calibration equation \( \sum_{i \in s} w_i x_i = X \).

The GREG estimation provides thus an alternative mathematical formulation of the
calibration estimation. That is, in case the transformation \( y_i - x_i \beta \) is made w.r.t. the
calibration totals \( X \), the resulting weights will be calibrated. This is managable via
suitable dummy indexing. On the other hand, the final weights depends now on how
the parameter \( \beta \) is defined, instead of the distance function \( G \) — though the two can
be made identical in “the linear case”. As an extreme case, post-stratified estimation
can be obtained by setting the dummy index to be the post-stratum indicator (Särndal,
Swensson, and Wretman, 1992, Section 7.6). Post-stratified estimation can therefore be
regarded as the “full regression model” which has included all the interaction among the
auxiliary variables.

4.2 Variations of calibration estimation

Deville and Särndal (1992) considered in fact a class of distance functions. In an even more
general form, individual coefficients \( 1/q_i \) can be attached to \( G \) to form a weighted overall
distance of the sample, i.e. the weighted calibration, though applications are dominated
by the standard case of \( q_i = 1 \). In any case, it was shown (Deville and Särndal, 1992)
that the linear method provides asymptotically the common linear approximation to all
the calibration estimators in this class. It is at the same time the fastest since it does
not require iterative fitting. It has also been noted that the calibrated estimate \( \hat{Y}_{cal} \) often
differs rather little from one method to another.

When the sample is small, the linear method might produce negative weights from
time to time. Should this be found undesirable, iterative algorithms can be developed
to restrict the range of the weights. See e.g. Jayasuriya and Valliant (1996) for an
application of this type of restricted regression estimation. Basically, one decides on the
lower and upper limits of the calibrated weights — weight ratio \( w_i/a_i \) exceeding 3 or 4
are considered large. After each iteration, the weights which fall outside of these limits
will be truncated, and the fitting algorithm are re-runned for the remaining sample, with
corresponding adjustment of the calibration equation. It is to be noticed that too strong
restrictions may cause the algorithm not to converge. We also note that the extent and
consequences of adjusting negative weights through weighted calibration has not been
much studied.

Inspection of the GREG estimator shows that the sign of the linearly calibrated weights depends largely on the inverse of the matrix $\sum_{i \in s} a_i x_i^T x_i$. The so-called ridge regression (Chambers, 1996) adds to this a user-specified positive diagonal matrix $D$ of the same dimensions, i.e. substituting $(D^{-1} + \sum_{i \in s} a_i x_i^T x_i)^{-1}$ for $(\sum_{i \in s} a_i x_i^T x_i)^{-1}$ in the formula for the linearly calibrated weights. It turns out the ridged weights can be obtained from minimizing the ridged loss function

$$\frac{1}{2} \sum_{i \in s} a_i (r_i - 1)^2 + \frac{1}{2} (X - \sum_{i \in s} w_i x_i) D (X - \sum_{i \in s} w_i x_i)^T,$$

whose second term can properly be regarded as a penalty to be paid for deviation from the population totals contained in $X$. For this reason the method can be classified as **penalized calibration**, which does not satisfy the calibration equation unless $D = \text{diag}(\infty)$. In particular, negative weights can almost always be eliminated if one is willing “to pay a large enough penalty”.

5 **Post-stratification and calibration (II)**

5.1 **A synthesis: Post-stratification is the finest calibration, and calibration the relaxed post-stratification**

By gradually relaxing the calibration equation from post-stratified estimation to GREG estimation and finally to the weighted and penalized calibration, calibration estimation increases the applicability of the population structure defined by the post-stratification. The question which remains is whether, or to which degree, this gain is accompanied by the preservation of a number of properties derived from the primary case of the simple post-stratified estimator. We shall concentrate here on the linear calibration estimator. In the light of the synthesis here, our approach is different from the standard one with a Horvitz-Thompson-start. The results in such cases can e.g. be found in Särndal, Swensson, and Wretman (1992). Throughout, we assume that the calibration totals are selected from $((1, H))$ where $h = 1, \ldots, H$ is the post-stratum index.

5.2 **The properties of the calibration estimator without empty sample post-strata**

Suppose first that the sample post-strata are all non-empty, i.e. $n_h > 0$ for $1 \leq h \leq H$. The linear calibration estimator can, in virtue of the transformation $y_i = x_i \beta + \epsilon_i$, be
rewritten as an adjustment of the simple post-stratified estimator \( \hat{Y}_{\text{pst}} \), i.e.

\[
\hat{Y}_{\text{cal}} = \hat{Y}_{\text{pst}} + \sum_{i \in s} v_i(x_i \beta + \epsilon_i)
\]

\[
v_i = w_i - q_h = w_i - N_h/n_h \quad \text{for} \ i \in s_h
\]

\[
\sum_{i \in s} w_i \epsilon_i = \sum_{h} q_h n_h x_h = X.
\]

If (a) \( \pi_{ih} = \pi_h \) for \( i \in s_h \), where \( \pi_i \) is the inclusion probability of the \( i \)-th unit and \( \pi_{ih} \) its inclusion probability conditional to \( n = (n_1, \ldots, n_H) \), and (b) \( w_i = w_h \) for \( i \in U_h \), then the conditional bias of \( \hat{Y}_{\text{cal}} \) simplifies to

\[
E[\hat{Y}_{\text{cal}} - Y | n] = \sum_h E[v_h \sum_{i \in s_h} \epsilon_i | n] = \sum_h n_h v_h (\sum_{i=1}^{N_h} \epsilon_i / N_h) = \sum_h n_h v_h \bar{E}_h,
\]

such that it is conditionally, and therefore unconditionally as well, unbiased regardless of the initial weights apart from condition (b), provided that, \( \forall 1 \leq h \leq H \),

\[
\sum_{i=1}^{N_h} \epsilon_i = 0.
\]

Notice that condition (b) can be generalized to (b)' \( \{w_i; i \in s_h\} \) remains constant conditional to \( n \), which however makes little difference in practice. In the transformation which results into the calibration estimator, \( \beta \) is such that \( \sum_{i \in U} \epsilon_i^2 \) is minimized for the given population. It follows that \( \sum_{i \in U} x_i \epsilon_i = 0 \), i.e. the residuals sum up to zero for each marginal, which is necessary yet not sufficient for (1), since the latter requires that the residuals sum up to zero within each population post-stratum. If we have (i) stratified srswr conditional to \( n \), and (ii) \( w_i = w_h \) for \( i \in U_h \), then

\[
\text{Var}(\hat{Y}_{\text{cal}} | n) = \sum_h n_h (1 - f_h) w_h^2 \sigma_h^2 \quad f_h = \frac{n_h}{N_h} \quad \sigma_h^2 = \sum_{i \in U_h} (y_i - \bar{Y}_h)^2 / N_h - 1.
\]

A key condition above is that \( w_i = w_h \) for \( i \in U_h \), which is satisfied whenever \( a_i = a_h \) for \( i \in s_h \). This follows since \( \{w_i\} \) minimizes, subject to the calibration equation,

\[
\sum_h a_h \sum_{i \in s_h} (\frac{w_i}{a_h} - 1)^2 = \sum_h (a_h^{-1} \sum_{i \in s_h} w_i^2 - 2 \sum_{i \in s_h} w_i + a_h n_h).
\]

Since the calibration equation, i.e. \( \sum_h x_h (\sum_{i \in s_h} w_i) = X \), will not be disturbed by the particular choice of \( \{w_i; i \in s_h\} \) as long as \( W_h = \sum_{i \in s_h} w_i \) remains the same, for arbitrary fixed \( W_h \), the distance is minimized at \( w_i = W_h/n_h \). In other words, \( w_i = w_h \) for \( i \in s_h \).
5.3 The properties of the calibration estimator with empty sample post-strata

Let \( R_0 \cup R_0^c = \{1, \ldots, H\} \), where \( R_0 \cap R_0^c = \emptyset \) and \( n_h = 0 \) for \( h \in R_0 \) and \( n_h > 0 \) for \( h \in R_0^c \), i.e.

\[
\hat{Y}_{\text{cal}} = \sum_{h \in R_0^c} q_h \left( \sum_{i \in s_h} y_i \right) + \sum_{i \in s} v_i y_i \quad q_h = N_h/n_h \quad \text{ and } \quad v_i = w_i - q_h \text{ for } i \in s_h.
\]

Let \( X_0 = \sum_{h \in R_0} \sum_{i \in U_h} x_i \), and \( E_0 = \sum_{h \in R_0} \sum_{i \in U_h} \epsilon_i \), and \( E_h^c = \sum_{i \in U_h} \epsilon_i \) for \( h \in R_0^c \). Notice that \( \sum_{i \in s} w_i x_i = X \) and \( \sum_{h \in R_0} q_h n_h x_h = X - X_0 \). Under the same condition (a) and (b) as before,

\[
E[\hat{Y}_{\text{cal}} - Y | n] = \sum_{h \in R_0^c} \left( \sum_{i \in U_h} v_i \right) \left( \sum_{i \in U_h} y_i / N_h \right) - \sum_{h \in R_0} \sum_{i \in U_h} y_i = \sum_{h \in R_0^c} V_h^c \bar{Y}_h - Y_0
\]

\[
= \left( \sum_{h \in R_0^c} V_h^c x_h \right) \beta + \sum_{h \in R_0^c} V_h^c \bar{E}_h^c - Y_0 = \left( X_0 \beta + \sum_{h \in R_0^c} V_h^c \bar{E}_h^c \right) - (X_0 \beta + E_0).
\]

In other words, \( \hat{Y}_{\text{cal}} \) is unbiased regardless of the initial weights apart from (b), provided

\[
(2) \quad \sum_{h \in R_0} \sum_{i \in U_h} \epsilon_i = 0 \quad \text{ and } \quad \sum_{i \in U_h} \epsilon_i = 0 \text{ for } h \in R_0^c.
\]

It is worth noting here that, since (2) follows from (1), the unbiasedness of the calibrated estimator can, for such populations, be “immune” towards empty cells in the sample, just like the method itself. Moreover, given (i) and (ii) as before, we have

\[
\text{Var}(\hat{Y}_{\text{cal}} | n) = \sum_{h \in R_0^c} n_h (1 - f_h) w_h^2 \sigma_h^2 \quad f_h = \frac{n_h}{N_h} \quad \sigma_h^2 = \sum_{i \in U_h} \frac{(y_i - \bar{Y}_h)^2}{N_h - 1}.
\]

Since this conditional variance probably underestimates the uncertainty in the estimation an ad hoc remedy consists in collapsing the empty and singular (where \( n_h = 1 \)) post-strata into other non-empty post-strata in some reasonable fashion, and use the combined totals instead of \( N_h \) for \( n_h > 1 \) alone. This we call the poorman’s variance estimator.
A CALWGT: A program package for calibration

A.1 General information

The program package for calibration CALWGT is written in S-plus for Unix — “Version 3.2 Release 1 for Sun SPARC, SunOS 4.x: 1993”. The installation diskette for CALWGT is available on request to the author at

\[ E-mail: \text{icz@ssb.no} \quad Tel: +47\ 22\ 00\ 44\ 78 \quad Fax: +47\ 22\ 86\ 47\ 34. \]

CALWGT can be freely distributed. To ensure version-consistency, however, OTHER names ought to be used after any modifications by the users. It is kindly requested that the author at the above address be contacted in case of any ambiguities or errors which may arise for improvements and corrections.

A.2 Installation and on-line help

The CALWGT installation diskette comes with the following files: “calwgt.aux”, “calwgt.drv”, “calwgt.ini”, “calwgt.src”, “calwgt.txt”, “readme.txt”. A description of the installation procedure can be found in “readme.txt”.

CALWGT has its own on-line help which will automatically be invoked under the installation. It contains information on how to set up the data for CALWGT, its calling parameters, how to handle abnormal exit of CALWGT, as well as a few practical tips on how to extend the standard theory of calibraiton to deal with some special cases. Once installed, the on-line help can be invoked any time in S-plus environment by typing in the command

\[ > .\text{calwgt.hlp}() \]

A.3 Calibrating the weights

The main part of CALWGT which deals with calibration is invoked in S-plus environment by

\[ > .\text{calwgt}(\text{calling.parameters}) \]

Please refer to the on-line help for how to set up the “calling.parameters”. In particular, CALWGT handles both categorical and continuous auxiliary variables.
Once started, CALWGT proceeds interactively where each prompt will be coupled with a number of helpful notes/comments. The built-in error detective mechanism should prove adequate in most cases provided the instructions are being followed. Basically, the user is able to choose between the linear and the multiplicative methods, with all their unrestricted, truncated or restricted options having been made available.

As a special note, one should avoid the logit \((L,U)\) (Deville, Särndal, and Sautory, 1993) method whenever possible. On the other hand, the user is encouraged to run both the linear and the multiplicative methods, and compare the resulting calibration estimates — these should be fairly close to each other for “nice” samples.

On normal exit, the calibrated weights will be written into “wgt.cal”, and the Lagrange multipliers into “lambda.cal” — both under the same directory as CALWGT.

A.4 An example

Suppose calibration is to be carried out towards (Unit index, Employment Status, Sex). The first of them is a constant auxiliary variable for all the members of the population; while the last of them is a binary variable. Suppose the employment status is divided into the three categories, i.e. “Employed”, “Unemployed”, “Labour-InActive”. CALWGT considers this calibration as having 3 auxiliary variables, with configuration vector \((1,3,2)\).

The population is now cross-classified into 6 \((= 1 \times 3 \times 2)\) post-strata. Instead of simply naming them as \((1,1,1), (1,1,2), \ldots, (1,3,2)\), the dummy indexing for natural calibration leads to the following model design matrix, which contains all the possible dummy auxiliary vectors,

\[
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 
\end{pmatrix}
\]

To actually carry out the calibration, the user must supply the population marginal counts — 6 of them here in this case, the sample design matrix, and the initial weights. Suppose the population marginal counts are \((60, 25, 15, 20, 25, 35)\), and that we have a sample of size 4 with sample design matrix given as

\[
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 
\end{pmatrix}
\]
\[
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0
\end{pmatrix},
\]

and the initial weights are (15, 15, 15, 15). CALWGT returns (12.5, 20, 15, 12.5) as the calibrated weights — the transcript is given below:

```R
> Splus
S-PLUS : Copyright (c) 1988, 1993 Statistical Sciences, Inc.
S : Copyright AT&T.
Version 3.2 Release 1 for Sun SPARC, SunOS 4.x : 1993
Working data will be in /ssb/lynx/h1/lcz/.Data
> .calwgt(F,F,F)
  Starting CALWGT...

Model specification — a vector which identifies the model.
For instance, calibration towards (sex,age,area) with, say,
four age groups and ten area codings implies 3 auxiliary
variables, with configuration vector (2,4,10).
The number of auxiliary variables (<number> <return>):
1: 3
The configuration vector (<number> <space> ... <number> <return>):
1: 1 3 2

The defined model has 3 auxiliary variables, each
with 1 3 2 levels, giving in total 6 marginal
counts w.r.t. which the calibration is to be carried out.

The size of the sample (<number> <return>): 1: 4

Typing in the population marginal counts on-line (6 of them )...
1: 60 25 15 20 25 35
  Typing in the initial weights of the sample units on-line (4 of them )...
1: 15 15 15 15
```
Typing in the sample design matrix on-line (4 * 6)...
No. 1, 1: 1 1 0 0 1 0
No. 2, 1: 1 0 1 0 0 1
No. 3, 1: 1 0 0 1 0 1
No. 4, 1: 1 1 0 0 1 0

The method of calibration:
press <l> and <return> for the iterative linear method;
press <r> and <return> for the NON-iterative linear method;
press <m> and <return> for the multiplicative method
— using IPS and for dummy indexing only;
press <n> and <return> for its quicker, all-round version
— using Newton-Raphson method;
press <g> and <return> for the logit (L,U) method
— a restricted multiplicative method.
1: r

With bounded weights or not (<y>/<n> <return>)?
1: n

Calibrating the weights... (See ‘calwgt.log’ for more information.)
CALWGT has successfully converged.
The calibrated weights have been stored under the name ‘wgt.cal’,
and the parameters of the model under ‘lambda.cal’.

Exit CALWGT... Bye!
> scan(“wgt.cal”)
[1] 12.5 15.0 20.0 12.5
> scan(”lambda.cal”)
[1] 0.3333333 -0.5000000 -0.3333333 0.0000000 0.0000000 0.0000000
> q()
References


Recent publications in the series Discussion Papers

125 E. Bijn and T.J. Klette (1994): Errors in Variables and Panel Data: The Labour Demand Response to Permanent Changes in Output
126 I. Svendsen (1994): Do Norwegian Firms Form Extrapolative Expectations?
127 T.J. Klette and Z. Griliches (1994): The Inconsistency of Common Scale Estimators when Output Prices are Unobserved and Endogenous
131 L.A. Grünfeld (1994): Monetary Aspects of Business Cycles in Norway: An Exploratory Study Based on Historical Data
134 K. A. Brekke and H.A. Gravningsmyr (1994): Adjusting NNP for instrumental or defensive expenditures. An analytical approach
139 L. Andreassen (1995): Aggregation when Markets do not Clear
146 T. O. Thoresen (1995): The Distributional Impact of the Norwegian Tax Reform Measured by Disproportionality
163 S. Grepperud (1996): Soil Conservation as an Investment in Land
168 R. Aaberge, A. Bjørklund, M. Jantti, M. Palme, P. J. Pedersen, N. Smith and T. Wennemo (1996): Income Inequality and Income Mobility in the Scandinavian Countries Compared to the United States