WORKING
PAPERS
FROM
DEPARTMENT
FOR
STATISTICS
ON
INDIVIDUALS
AND
HOUSEHOLDS

METHODS
FOR
COLLECTIONS
AND
ANALYSIS

ARBEIDSNOTAT
FRA
AVDELING
FOR
PERSONSTATISTIKK

METODER
FOR INNSAMLING
OG ANALYSE

3/1992

Johan Heldal
Division for Methods and Standards

A Method for Calibration of Weights in Sample Surveys

CENTRAL
BUREAU
OF STATISTICS
OF NORWAY

Statistisk
sentralbyrå

FORORD

I denne serien samles notater innen feltet metoder for innsamling og analyse som har krav på en viss allmenn interesse, men som ikke presenterer avsluttede arbeider. Det som presenteres vil ofte være mellomprodukter på vei fram mot en endelig artikkel eller publikasjon, eller andre arbeider som forfatteren eller avdelingen er interessert i en viss spredning av og å få kommentert. Når de er ferdig bearbeidet, vil noen av arbeidene bli publisert i andre sammenhenger.

Synspunktene som presenteres er forfatternes egne, og er ikke nødvendigvis uttrykk for for SSBs oppfatning.

PREFACE

This series contains papers within the field of methodology. The papers are expected to be of some general interest, and presents work in progress, or other notes worth a limited distribution.

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of the Central Bureau of Statistics of Norway.

Statistisk
sentralbyrå

CENTRAL
BUREAU
OF STATISTICS
OF NORWAY

OSLO
PB 8131 DEP
N-0033 OSLO
TELEFON (02) 86 45 00
FAX (02) 86 49 88
TELEX 11 202 SSB-N
BESØKSADRESSE
SKIPPERGATEN 15

KONGSVINGER
POSTUTTAK
N-2201 KONGSVINGER
TELEFON (066) 85 000
FAX (066) 85 289
BESØKSADRESSE
OTERVEIEN 23

A Method for Calibration of Weights in Sample Surveys.

Johan Heldal
Central Bureau of Statistics of Norway
P.O. Box 8131, Dep., N-0033 Oslo

October 29, 1992

Abstract

Sometimes statistics based on sample surveys are published for population totals for which the true values are known in advance from other sources, such as registers. This paper describes a method to calibrate the weights of persons and households in such a way that the estimates from the sample are forced to fit the true values exactly. The external information which is thereby incorporated in the weights may also help improving the estimation of other quantities. Applications are given.

Keywords: Regression estimation, weighting procedures.

1 Introduction

Central statistical offices perform series of surveys on samples of persons, households, establishments and other kinds of units. These surveys make up much of the foundation for the official statistics being published by the offices, such as estimates for population totals and averages.

But there are other sources as well, among them registers comprising the entire populations. In Norway, the Central Population Register, the file of Incomes and Taxes are two of now several registers covering various populations.

Sometimes, statistics for the same quantity are published in different publications based on different sources of data. As an example, in Norway, statistics on income is published based on the file of Incomes and Taxes. Estimates for the same quantities are being published based on the Survey of Income, which is a sample survey where income tax returns for persons in a sample of households are collected from the municipal tax offices. Unneccessary to say, the two statistics differ. The estimates from the Survey of Income have sampling errors while the statistics from the file of Incomes and Taxes have not. The latter source can be considered to give the "true" numbers for the incomes. (This is however not always the case for register files.)

From a publication point of view, a situation with two different statistics for the same quantity bearing the same official authorization is rather awkward. In many situations it is therefore desireable to force the estimates from the survey to comply with those of the register. But this has to be done in such a way that it does not destroy the mutual consistency among the variables in the survey. For instance, the quantities on the income tax return define an account which must agree also when estimating population totals. Thus, the statistics from the register cannot just replace the estimates from the survey without any further reference.

There are however methods that can be used to obtain what we desire. The method to be described here is based on regression estimation. It adjusts the weights used to multiply the individual observations when totals are being estimated. These adjustments can also improve estimates of totals for other variables for which totals are not known in advance or incorporated in the calibration procedure. The detailed description of this procedure and application of it is the topic for this paper.

2 The ratio estimator

Consider a finite population with N units numbered by the index i=1, . . . ,N. Let s denote a probability sample drawn from the set $\mathcal{S}$ of all possible sampes from the given population. The probability that unit i will be drawn to the sample will be denoted p_i.

Let y_i be the value attached to unit i of some variable of interest in a sample survey. We want to estimate the population total

$$Y = \sum_{i=1}^{N} y_i$$

The traditional estimator for this quantity is the Horwitz-Thompson estimator

$$Y_{HT} = \sum_{i \in s} w_i y_i$$

where $w_i=1/p_i$ (Horwitz and Thompson 1952). The principal feature of the Horwitz-Thompson estimator is that it is design-unbiased. It has no optimality properties what concerns *precision*, which depends completely on the relation between y_i and p_i.

Suppose that we have access to an auxiliary variable x_i which is known in the sample and for which we can compute the total

$$X = \sum_{i=1}^{N} x_i$$

from some other source of data. In the following such a variable will be called a *key* variable. Let

$$X_w = \sum_{i \in s} w_i x_i$$

where the weights $\{w_i, i = \ldots, N\}$ are arbitrary. Let Y_w be similarly defined. Then the *ratio estimator* for the total Y has the form

$$Y_r = \frac{Y_w}{X_w} X \qquad (1)$$

The ratio estimator works best if there is an approximately linear relation between the ys and the $x$es of the form $y_i \approx b x_i$. Particularly, if the y-variate is the x-variate itself, (1) becomes

$$X_r = \frac{X_w}{X_w} X = X. \qquad (2)$$

The ratio estimator can be considered as a method to change the weights. Substituting for Y_w in (1) yields

$$Y_r = \sum_{i \in s} y_i \frac{X}{X_w} w_i.$$

In other words, the ratio estimator replaces the weights $\{w_i\}$ with weights of the form

$$v_i = \frac{X}{X_w} w_i.$$

The choice of weights $\{w_i\}$ will depend on the kind of statistical philosophy lying behind the use of the estimator. This is a subject of its own and will not be discussed here. The important point is that (2) holds whatever weight system is chosen. If the new weights shall be used as general weights for all kinds of variables, the ws must just not depend on what kind of variable y is, but they can in principle depend on the key variable x.

3 Regression estimator with one keyvariabel

The ratio estimator is an intuitive and simple estimator to use, and its properties are well studied (See Cochran 1977). But there are other methods which can be used to create new weights having the property that they estimate a given key-variable correctly. Consider the approximate linear relationship

$$y_i = \beta x_i + e_i \tag{3}$$

where e_i is an error term. (3) can either be interpreted as an in some sense "true" statistical model where e_i is a stochastic variable having expectation $E(e_i)$=0, or it can be interpreted as a purely descriptive relation in a finite population. Whatever interpretation, a best empirical fit for β in a least squares sense can be found by minimizing the expression

$$\sum_{i \in s} w_i e_i^2 = \sum_{i \in s} w_i [y_i - b x_i]^2.$$

The solution to this minimization problem is

$$\hat{\beta} = \frac{\sum_{i \in s} w_i x_i y_i}{\sum_{j \in s} w_j x_j^2}.$$

The regression estimator for Y can now be written

$$Y_R = \hat{\beta} X = \sum_{i \in s} [\frac{w_i x_i X}{\sum_{j \in s} w_j x_j^2}] y_i. \tag{4}$$

The new weights estimating are given by the content of the brackets in (4), that is

$$v_i = \frac{x_i X}{\sum_{j \in \mathrm{s}} w_j x_j^2} w_i.$$

If one wishes to estimate X by X_R, one can do so by substituting x_i for y_i above. Then $\hat{\beta}$=1 and X_R=X. Thus X is estimated correctly.

Define $\hat{y}_i$=$\hat{\beta} x_i$. Then

$$Y_R = \sum_{i=1}^{N} \hat{y}_i. \tag{5}$$

For $i \in \mathrm{s}$, y_i is known. In model based inference, and especially for small populations, it is recognized that it is better to substitute y_i for $\hat{y}_i$ for $i \in \mathrm{s}$ in (5). Doing so, we obtain the estimator Y_P (P for *prediction*)

$$Y_P = \sum_{i \in \mathrm{s}} y_i + \sum_{i \notin \mathrm{s}} \hat{\beta} x_i. \tag{6}$$

Also this estimator can be written as a weighted sum of the observed ys in such a way that the estimator applied to the $x$es yield the true value of X. Substituting for $\hat{\beta}$ in (6), we get

$$Y_P = \sum_{i \in \mathrm{s}} [1 + \frac{x_i(X - X_{\mathrm{s}})}{\sum_{j \in \mathrm{s}} w_j x_j^2} w_i] y_i \tag{7}$$

where

$$X_{\mathrm{s}} = \sum_{i \in \mathrm{s}} x_i.$$

The new weight for unit i, say ν_i, is the expression in the brackets in (7).

In the next section, the regression approach to weighting will be extended to the case with several key-variables.

4 Estimation with more than one keyvariable

The ratio estimator in section 2 and the regression estimator in section 3 presented methods that made it possible to change the weight system in such a way that the total of a specific variabel, the key-variable, would be estimated correctly by the new weights. Some y-variables, having a close to linear relationship to the key-variable, could also be estimated better by the new weights than by the Horwitz-Thompson estimator, while estimators not showing such a relationship not neccesarily will. We shall now see how it is possible to incorporate several key-variables jointly and adapt a new system of weights so that the totals of all the key-variables are estimated correctly by the new set of weights. The effect of using these weights to estimate total Y for a variable which is not a key-variable will also be discussed. The method is considered earlier in Bethlehem & Keller (1987).

Suppose that we replace the key-variable x_i with a vector of key-variables

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, N.$$

For these variables we know the total

$$X_j = \sum_{i=1}^{N} x_{ij}, \quad j = 1, \ldots, p.$$

The vector of all totals for the key variables will be denoted by $\boldsymbol{X} = (X_1, \ldots, X_p)$.

As in section 3 we consider an approximate linear relationship

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i = \boldsymbol{\beta}\boldsymbol{x}_i' + e_i, \quad i = 1, \ldots, N, \tag{8}$$

where $\boldsymbol{\beta}=(\beta_1, \ldots, \beta_p)$. The best least squares fit for $\boldsymbol{\beta}$ is found by minimizing the expression

$$\sum_{i \in s} w_i e_i^2 = \sum_{i \in s} w_i [y_i - \boldsymbol{\beta}\boldsymbol{x}_i']^2. \tag{9}$$

Let $\mathbf{X}$ be the $N \times p$ matrix with the key-variables $\boldsymbol{x}_i$ as rows,

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$$

and let $\mathbf{Y}$ be the matrix of variables of interest, that is

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

(Notice the difference between the **bold X** and the ***italic bold X***.) The regression model (8) can the be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}' + \mathbf{E}$$

where $\mathbf{E}$ is the $N \times 1$ vector of error terms e_i. Let $\mathbf{X}_\mathrm{s}$ and $\mathbf{Y}_\mathrm{s}$ be $n \times p$ and $n \times 1$ versions of $\mathbf{X}$ and $\mathbf{Y}$ for the units $i \in \mathrm{s}$ where n is the sample size. Let $\mathbf{W}_\mathrm{s}$ be the $n \times n$ diagonal matrix with the weights $\{w_i\}$ on the diagonal. The vector $\hat{\boldsymbol{\beta}}$ minimizing (9) is then given by

$$\hat{\boldsymbol{\beta}} = \mathbf{Y}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s}(\mathbf{X}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s})^{-1}$$

Let $\hat{y}_i = \hat{\boldsymbol{\beta}}\boldsymbol{x}_i'$ and $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}'$. The regression estimator can now be written

$$Y_R = \sum_{i=1}^{N} \hat{y}_i = \hat{\boldsymbol{\beta}}\boldsymbol{X}' = \mathbf{Y}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s}(\mathbf{X}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s})^{-1}\boldsymbol{X}'.$$

Again, as in section 3, the predicted values $\hat{y}_i$ can be substituted by y_i for $i \in \mathrm{s}$, giving the prediction estimator

$$Y_P = \sum_{i \in \mathrm{s}} y_i + \sum_{i \notin \mathrm{s}} \hat{y}_i = \sum_{i \in \mathrm{s}} y_i[1 + w_i\boldsymbol{x}_i(\mathbf{X}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s})^{-1}(\boldsymbol{X} - \boldsymbol{X}_\mathrm{s})']$$

Let

$$v_i = w_i\boldsymbol{x}_i(\mathbf{X}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s})^{-1}\boldsymbol{X}', \quad i \in \mathrm{s} \tag{10}$$

and $\mathbf{v}=[v_{i_1}, \ldots, v_{i_n}]$. $i_1, \ldots, i_n$ are the indexes i that are contained in s. Then the regression estimator can be written

$$Y_R = \mathbf{v}\mathbf{Y}_\mathrm{s} = \sum_{i \in \mathrm{s}} v_i y_i.$$

Similarly, let

$$\nu_i = 1 + w_i\boldsymbol{x}_i(\mathbf{X}_\mathrm{s}'\mathbf{W}_\mathrm{s}\mathbf{X}_\mathrm{s})^{-1}(\boldsymbol{X} - \boldsymbol{X}_\mathrm{s})', \quad i \in \mathrm{s}$$

and $\boldsymbol{\nu}=[\nu_{i_1}, \ldots, \nu_{i_n}]$. Then

$$Y_P = \boldsymbol{\nu}\mathbf{Y}_\mathrm{s} = \sum_{i \in \mathrm{s}} \nu_i y_i.$$

The regression estimate $\boldsymbol{X}_R$ for $\boldsymbol{X}$ then is

$$\boldsymbol{X}_R = \boldsymbol{X}(\mathbf{X}_s'\mathbf{W}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{W}_s\mathbf{X}_s = \boldsymbol{X}$$

and the prediction estimate $\boldsymbol{X}_P$ for $\boldsymbol{X}$ is

$$\boldsymbol{X}_P = \mathbf{X}_s(\mathbf{1}_n' + \mathbf{W}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{W}_s\mathbf{X}_s)^{-1}(\boldsymbol{X} - \boldsymbol{X}_s)') = \boldsymbol{X}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones. In other words, both the new weight systems satisfy the requirement that they jointly estimate the totals for all key-variables correctly.

When choosing key-variables, care should be taken so that co-linearity problems do not arise. Furthermore, if there are many candidates for key-variables not having significant impact on many possible y-variables, one may get many insignificant βs, decreasing the precision of Y_R and Y_P. Thus, if there is a large number of y-variables whose totals will be estimated by the new weights, it may pay to sacrifice the exact fit for some candidate key-variables that do not "explain" much variability for many y-variables.

In design based estimation where the original weights w_i are the Horwitz-Thompson weights $1/p_i$, the estimators Y_R and Y_P are not unbiased. However, if the $N \times 1$ vector $\mathbf{1}_N$ consisting of N ones is in the columnspan of $\mathbf{X}$, both estimators are consistent and asymptotically design-unbiased. If $\mathbf{1}_N$ is one of the columns of $\mathbf{X}$ it means that we have an intercept term in the regression.

Variables that are linear transformations of the key-variables are also estimated correctly by the method. This is a sometimes useful feature. Assume that the vectors $\boldsymbol{z}_i$ of q variabels can be written

$$\boldsymbol{z}_i = \boldsymbol{x}_i\mathbf{C}, \quad i = 1, \ldots, N$$

with a total $\boldsymbol{Z}=\sum_{i=1}^{N} \boldsymbol{z}_i$, where $\mathbf{C}$ is a $p\times q$ matrix. Then $\boldsymbol{Z}_R=\boldsymbol{X}_R\mathbf{C}=\boldsymbol{X}\mathbf{C}=\boldsymbol{Z}$ which implies that $\boldsymbol{Z}$ is also estimated correctly by the new weights.

A problem which may arise and make the estimation of the weights unstable, is near colinearity in the $\mathbf{X}$ matrix. Since variables that result from linear transformations of keyvariables will also be estimated correctly, they can themselves be used as key-variables and should yield the same set of weights as the original key-variables. By choosing the transforming matrix $\mathbf{C}$ carefully, one can obtain new variables $\mathbf{Z}$=$\mathbf{XC}$ with smaller condition number than the original $\mathbf{X}$. Scaling of the columns of $\mathbf{X}$ to the same order of size could be a first step in construction of such new variables. Forsuch a purpose $\mathbf{C}$ could also be chosen from the data, for instance by letting $\mathbf{C}$ be a transformation to the principal components of $\mathbf{X}$.

Example. What will the weights look like when the matrix $\mathbf{X}$ consists of two columns, one of which is the vector $\mathbf{1}_N$ and the other is an $N \times 1$ vector $\mathbf{x}$? Let

$$\mathbf{X} = (\mathbf{1}_N, \mathbf{x}) \quad \text{and} \quad \mathbf{X}_\mathrm{s} = (\mathbf{1}_n, \mathbf{x}_\mathrm{s}).$$

Define

$$N_w = \sum_{i \in \mathrm{s}} w_i.$$

Then

$$\mathbf{X}_\mathrm{s}' \mathbf{W}_\mathrm{s} \mathbf{X}_\mathrm{s} = \begin{bmatrix} N_w & X_w \\ X_w & \sum_{i \in \mathrm{s}} w_i x_i^2 \end{bmatrix}.$$

Let $\overline{X}=X/N$, $\overline{X}_w=X_w/N_w$ and

$$S_w^2 = \frac{1}{N_w} \sum_{i \in \mathrm{s}} w_i (x_i - \overline{X}_w)^2.$$

$N_w^2 S_w^2$ is the determinant of $\mathbf{X}_\mathrm{s}' \mathbf{W}_\mathrm{s} \mathbf{X}_\mathrm{s}$ and

$$(\mathbf{X}_\mathrm{s}' \mathbf{W}_\mathrm{s} \mathbf{X}_\mathrm{s})^{-1} = \frac{1}{N_w^2 S_w^2} \begin{bmatrix} \sum_{i \in \mathrm{s}} w_i x_i^2 & -X_w \\ -X_w & N_w \end{bmatrix}.$$

For the ith unit the weight v_i is

$$\begin{aligned} v_i &= \frac{1}{N_w^2 S_w^2} [1,\ x_i] \begin{bmatrix} \sum_{i \in \mathrm{s}} w_i x_i^2 & -X_w \\ -X_w & N_w \end{bmatrix} \begin{bmatrix} N \\ X \end{bmatrix} w_i \\ &= \frac{N}{N_w} (1 - \frac{(\overline{X} - \overline{X}_w)(\overline{X}_w - x_i)}{S_w^2}) w_i \end{aligned}$$

As a special case, take $w_i = c, i = 1, \ldots, N$, for instance $c = N/n$. Take $\overline{X}_\mathrm{s}=X_\mathrm{s}/n$. Then we get $N_w=cn$, $\overline{X}_w=\overline{X}_\mathrm{s}$ and $S_w^2=\frac{1}{n}\sum_{i \in \mathrm{s}}(x_i - X_\mathrm{s}/n)^2$ which is usually denoted s^2., Then v_i simplifies to

$$v_i = \frac{N}{n}(1 - \frac{(\overline{X} - \overline{X}_\mathrm{s})(\overline{X}_\mathrm{s} - x_i)}{s^2}])$$

If the original constant weights are N/n, v_i will give a smaller weight to the unit if $\overline{X}> \overline{X}_{\mathrm{s}} >x_i$ and if $\overline{X} < \overline{X}_{\mathrm{s}}< x_i$. Otherwise v_i will give a greater weight.

The corresponding formula for the weight system ν_i is obtained by substituting $N-n$ for N and $\overline{X}_{\bar{\mathrm{s}}}$ for $\overline{X}$ in the above formula, where $\overline{X}_{\bar{\mathrm{s}}}=(X - X_{\mathrm{s}})/(N-n)$, the average of the $x$es that are not in s, giving

$$\nu_i = \frac{N-n}{n}(1 - \frac{(\overline{X}_{\bar{\mathrm{s}}} - \overline{X}_{\mathrm{s}})(\overline{X}_{\mathrm{s}} - x_i)}{s^2}).$$

ν_i is less than N/n if $(\overline{X}_{\bar{\mathrm{s}}} - \overline{X}_{\mathrm{s}})(\overline{X}_{\mathrm{s}} - x_i)>-ns^2/(N-n)$.

5 Consistency between samples for households and persons.

The Survey of Income and many other surveys cover both persons and households. Often a household is sampled by drawing a person and take that persons household as a sample household. A sample of persons is then constructed from all persons in the sample households. The probability that a given household shall be sampled is therefore equal to the probability that at least one of its persons shall be drawn in the first instance. Finally then, all persons in a household get the same probability of being included in the sample, and this probability is equal to the inclusion probability of their household. Thus, if the Horwitz-Thompson estimator is used, the household itself and all persons in it will have the same weight.

Consider a situation were we have a population consisting of M households containing a total of N persons. Let H_h be household no. h, h=1,...,M and let still i=1,...,N index the persons. Let π_h be the probability that household h is sampled and let as before p_i be the probability that person no. i is included. The situation described above can then be formulated as

$$\pi_h = p_i \text{ if } i \in H_h \tag{11}$$

and thus

$$p_i = p_j \text{ if both } i \text{ and } j \in H_h.$$

This is a very useful property if we wish to make statistics for households based on the sample. A large number of household variables are constructed by aggregating variables attached to the persons up to household level.

In order to be able to discuss problems concerning this rather trivial situation in the context of new weight systems, more notation is needed. Let as before y_i be the value of a variable of interest for *person* no. i and let ψ_h be the value of the same quantity aggregated for household no. h. That is

$$\psi_h = \sum_{i \in H_h} y_i.$$

(As a convention greek letters will be used for household quantities.) For the respective totals we have of course

$$\Psi = \sum_{h=1}^{M} \psi_h = \sum_{i=1}^{N} y_i = Y.$$

Let the weights for the persons be $\{w_i\}$ and denote the household weights by $\{\omega_h, h{=}1, \ldots, M\}$. Furthermore, let ς be the sample of huseholds. When making household samples the way described above, one should require that the total $Y_w {=} \sum_{i \in s} w_i y_i$ estimated from the sample of persons and the total $\Psi_\omega {=} \sum_{h \in \varsigma} \omega_h \psi_h$ estimated from the sample of households should give the same number. Something else would be awkward. If

$$\omega_h = w_i \text{ for all } i \in H_h \tag{12}$$

we have

$$Y_w = \sum_{i \in s} w_i y_i = \sum_{h \in \varsigma} \sum_{i \in H_h} \omega_h y_i = \sum_{h \in \varsigma} \omega_h \psi_h = \Psi_\omega.$$

It follows from (11) that this required property holds for the Horwitz-Thompson estimator.

When changing the person weights by the methods described in the sections 3 and 4, we will soon see that the new weights for different persons in the same household are different. It is no longer possible to aggregate the persons belonging to the same household and deduce a sensible weight for the household. However, if the original weights for the persons and the households satify the requirement (12), we shall see that it is still possible to construct new weights which also do and at the same time make all the estimates for totals of key variabes fit their true values exactly as in the sections 3 and 4. Two methods for doing this will be described.

The first method stems from an article by Lemaître and Dufour (1987). With their method, the average in each household of each key variable $\{x_{ji}, i \in H_h\}$ is computed. Let n_h be the size of household h. Define

$$\xi_{hj} = \sum_{i \in H_h} x_{ij} \text{ and } \bar{\xi}_{hj} = \xi_{jh}/n_h, \ \ j = 1, \dots, p.$$

For each person we make a new variable $\{u_{ij}, i \in H_h\}$ which is exactly this quantity, that is

$$u_{ij} = \bar{\xi}_{hj}, \ i \in H_h; \ \ j = 1, \dots, p.$$

Then, instead of using the original $x$es as key variables, we use the us. Since all persons in the same household will have the same values for the u-variables, the new weights for the persons, say v_{1i}, will be the same for all persons in the same household. This weight can be taken as the new household weight ω_{1h}. The totals for the us will be estimated correctly, and for these totals we have

$$\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{u}_i = \sum_{h=1}^{M} \boldsymbol{\xi}_h = \sum_{i=1}^{N} \boldsymbol{x}_i = \boldsymbol{X},$$

where $\boldsymbol{u}_i = (u_{1i}, \dots, u_{pi})$ and $\boldsymbol{\xi}_h = (\xi_{1i}, \dots, \xi_{pi})$. Thus the vector of totals, $\boldsymbol{X}$, will be estimated correctly. Let

$$\mathbf{U} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_N \end{bmatrix}$$

and let $\mathbf{U}_s$ be the corresponding matrix for the sample. The new weight system for the households and persons can by (10) be written

$$\omega_{1h} = v_{1i} = w_i \mathbf{u}_i (\mathbf{U}_s' \mathbf{W}_s \mathbf{U}_s)^{-1} \boldsymbol{X}', \ \ h \in \varsigma. \qquad (13)$$

For the discussion here, regression estimator weights will be used. However, the discussion and the conclusions to come apply equivalently to the prediction estimator weights.

In the second method one first makes a household sample by aggregating over the persons in each household, then using the method described in section 4 with the household variables $\{\boldsymbol{\xi}_h\}$ as key variables and the variables

$\{\psi_h\}$ as "target" variables. The new household weights generated this way can be used directly as weights for the persons in the household. Let

$$\Xi = \begin{bmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_M \end{bmatrix}$$

and let Ξ_ς be the corresponding sample version, having m rows where m is the number of households in the sample. Let $\{\omega_h\}$ be the original weights for the households and let $\boldsymbol{\Omega}_\varsigma$ be the $m \times m$ diagonal matrix with the elements $\{\omega_h,\ h \in \varsigma\}$ on the diagonal. The new household weights can be written

$$\omega_{2h} = \omega_h \boldsymbol{\xi}_h (\Xi_\varsigma' \boldsymbol{\Omega}_\varsigma \Xi_\varsigma)^{-1} \boldsymbol{X}', \ \ h \in \varsigma \tag{14}$$

and

$$v_{2i} = \omega_{2h}, \ \ i \in H_h$$

Consider equation (13). The elements of the matrix $(\mathbf{U}_s'\mathbf{W}_s\mathbf{U}_s)$ can be written

$$(\mathbf{U}_s'\mathbf{W}_s\mathbf{U}_s)_{jk} = \sum_{i \in s} w_i u_{ij} u_{ik} = \sum_{h \in \varsigma} \sum_{i \in H_h} \omega_h \xi_{hj} \xi_{hk} / n_h^2 = \sum_{h \in \varsigma} \xi_{hj} \xi_{hk} \omega_h / n_h. \tag{15}$$

The components of the first terms in (13), $w_i \mathbf{u}_i$ can similarly be written out as

$$w_i u_{ij} = \xi_{hj} \omega_h / n_h. \tag{16}$$

Consider the elements of the matrix $\Xi_\varsigma' \boldsymbol{\Omega}_\varsigma \Xi_\varsigma$ in equation (14). They can be written as

$$(\Xi_\varsigma' \boldsymbol{\Omega}_{sh} \Xi_\varsigma)_{jk} = \sum_{h \in \varsigma} \xi_{hj} \xi_{hk} \omega_h \tag{17}$$

The components of the term $\boldsymbol{\xi}_h \omega_h$ of (14) is

$$\xi_{hj} \omega_h \tag{18}$$

Compare (17) by (15) and (18) by (16). Then it becomes clear that the second method uses the method described in 4 directly on the households with the weigths that follow naturally from $\omega_h = w_i$ for $i \in H_h$. The metod of Lemaître and Dufour first divides these weights by n_h and then applies the method in section 4. If the original weights were the Horwitz-Thompson weights,

$1/\pi_h=1/p_i$, the second method applies these weights directly as the original household weights, while the first method uses the weights $1/n_h\pi_h=1/n_hp_i$. With the method of household sampling described in the beginning of this section, the original Horwitz-Thompson weights w_i and ω_i are approximately proportional to $1/n_h$. Lemaître and Dufours method replaces ω_h by weights approximately proportional to $1/n_h^2$. Which method is the best is studied empirically in the next section.

The discussion in this section has been carried through for the weights of the regression estimator. However, the discussion and the same results hold true also for the weights of the prediction estimator.

6 An application

The application presented here is based on the Norwegian Survey of Income from 1990 which has already been mentioned. True totals of the key-variables used, and of other variables of interest not used as key-variables, are identified from the file of Incomes and Taxes belonging to the taxation and revenue authorities. The sample was drawn essentially as described in section 5 and consisted of 13677 persons, 13 years or more, in 6046 households.

The key variables being used are:

- Exemption group with two groups, group 1 and group 2.
- Net receipts (income after deductions) falling in each of five graduation intervals.

The graduation steps in 1990 for the two exemption groups (in Norwegian kroner) were

Group1	61000	122000	158000	201000
Group2	76000	153000	182000	207000

The reason for dividing the net receipts into graduation intervals is twofold. First, the distribution of income is of interest in itself. Secondly, since the tax rate is constant within each interval, correct estimation of total net receipt within each interval will cause the total tax revenues will be estimated correctly as well.

Table 1 below shows true values and estimates for the key-variables and for some other variables of interest by the weight systems discussed in 5. Metod I is Lemaître and Dufours metod. The original weights w_i used for estimation were the Horwitz-Thompson weights which in this case were approximately inversely proportional to household size.

Table 1. Estimates of totals for selected variables by different weight systems. Amounts in 100 kroner.[1]

Variable	Answers	H-T weights	Method I	Method II
No. of persons in				
*Exempt. gr. 1	3016322	2904915	3016322	3016322
*Exempt. gr. 2	414437	411721	414437	414437
Net receipts in intervals of exemption group I:				
*0 - 610	1491944248	1391742252	1491944248	1491944248
*611 - 1220	907948473	917136290	907948473	907948473
*1221 - 1580	286850181	317258637	286850181	286850181
*1581 - 2010	168368914	198891698	168368914	168368914
*≥2011	194834249	222825943	194834249	194834249
Total n.r.	3049946065	3047854820	3049946065	3049946065
Estim. tax	90429556	103417862	90429556	90429556
Net receipts in intervals of exemption group II:				
*0 - 760	224230603	215787580	224230603	224230603
*761 - 1530	134138751	131359499	134138751	134138751
*1531 - 1820	26496023	30841370	26496023	26496023
*1821 - 2070	15646301	17566630	15646301	15646301
*≥2071	58380040	61558707	58380040	58380040
Total n.r.	458891718	457113786	458891718	458891718
Estim. tax	15234080	16535444	15234080	15234080

[1] * marked variables are key-variables

Table 1 continued.

Variable	Answers	H-T weights	Method I	Method II
Basis for high-income taxation in intervals of exemption group I:				
0-2050	3281115419	3259749485	3278397235	3277991617
≥2051	340435377	342106663	336474738	340936156
Basis for high-income taxation in intervals of exemption group II:				
0-2470	500788756	509935195	506720173	5058906517
≥2471	65144250	72834991	73478679	72600124
Totals for three kinds of pensionable incomes:				
From wages and salaries	3151247759	3151897530	3151433997	3236911606
Self empl. I	101320790	103252904	104261636	130755631
Self empl.II	190788034	194996765	189210175	202431753
Property tax	14842036	14593802	14310915	13517815

Table 1 shows that the key-variables have been estimated correctly by both new weight systems. The total net receipts and the net income tax have also been estimated correctly since they are both linear functions of the key variables with known coeffissients. However, the new weights do not always estimate the non-key variables better than the Horwitz-Thompson weights. The new estimates for the bases for the high-income taxation hit their target values approximately as well as the H-T estimates with method II possibly slightly better than method I. For the pensionable incomes (which are the bases for calculation of the National Pension Insurance premiums) and the property tax, method I hits approximately as well as th H-T estimates while method II hits significantly worse.

Experiments with the two metods, also using other sets of key-variables and datasets for the years 1986 to 1989, show consistently a result indicated in table 1: Method I of Lemaître and Dufour hits the target values better than method II for most variables most years. This is so in spite of the extreme initial downweighting of the large households produced by method I.

In a recent paper, Deville & Särndal (1992) showed that the calibrated weights presented in section 4 could be obtained by minimizing the distance between the old and the new weights under the restriction that the new weights should estimate the key-variables correctly. More precisely, they proposed to minimize

$$D_{wv}^2 = \frac{1}{N} E\{\sum_{i \in s} (w_i - v_i)^2 / w_i\} \tag{19}$$

where w_i is the Horwitz-Thompson weight. They thereby suggest that D_{wv}^2 should be as small as possible for the new estimator to be stable. Bearing the above results in mind, it is therefore of interest to compare this quantity for method I and II. This has been done in table 2 for the 1986 to 1990 Surveys of Income. The choices of key-variables have for all these years been kept as similar as possible to the choice above for 1990. n is the number of persons in the sample and m is the number of households.

Table 2 D_{wv}^2 estimated for 5 years

Year	n	m	Method I	Method II
1986	12087	4975	15.51	75.54
1987	8119	3393	17.11	97.17
1988	7872	3423	34.10	94.55
1989	7710	3475	14.46	99.94
1990	13677	6046	8.84	51.08

Table 2 shows that D_{wv}^2 is smaller for method I than for method II for all five years. This is in consistent with our experience that method I is the more stable. One interpretation of this result may be that method I exploits an information which method II does not take advantage of, namely the household size and that this information effectively increases the sample size from m to near n. This interpretation becomes more reasonable when it is considered that method I actually works on a sample of n household averages. Nevertheless, table 2 shows that the Horwitz-Thompson weight is not neccessarily the best initial weight in such a method, even form a design-based point of view.

The variations in the estimate of D_{wv}^2 over the five years may to some extent reflect variations in sample design for the five surveys.

References

[1] Bethlehem, J. G. & Keller, W. J. (1987): *Linear Weighting of Sample Survey data.* Journal of official statistics, Vol. 3, no. 2, p. 141-153.

[2] Cochran, W. G. (1977): *Sampling Techniques* 3rd ed., Wiley

[3] Horwitz, D. G. & Thompson, D. J. (1952): *A Generalization of Sampling Without Replacement from a Finite Universe.* Jour. of Am. Stat. Ass., Vol 47, p. 663-685.

[4] Lemaître, G. & Dufour, J. (1987): *An Integrated Method for Weighting Persons and Families.* Survey Methodology, Vol 13, no. 2, p. 199-207.

[5] Deville, Jean-Claude & Särndal, Carl-Erik (1992) *Calibration Estimators in Survey Sampling* Jour. of Am. Stat. Ass., Vol 87, p. 376-382.

References

[1] Bethlehem, J. G. & Keller, W. J. (1987): *Linear Weighting of Sample Survey data.* Journal of official statistics, Vol. 3, no. 2, p. 141-153.

[2] Cochran, W. G. (1977): *Sampling Techniques* 3rd ed., Wiley

[3] Horwitz, D. G. & Thompson, D. J. (1952): *A Generalization of Sampling Without Replacement from a Finite Universe.* Jour. of Am. Stat. Ass., Vol 47, p. 663-685.

[4] Lemaître, G. & Dufour, J. (1987): *An Integrated Method for Weighting Persons and Families.* Survey Methodology, Vol 13, no. 2, p. 199-207.

[5] Deville, Jean-Claude & Särndal, Carl-Erik (1992) *Calibration Estimators in Survey Sampling* Jour. of Am. Stat. Ass., Vol 87, p. 376-382.