SCB STATISTISKA CENTRALBYRÅN

1985-08-01

PROMEMORIOR FRÅN P/STM

NR 18

RECONCILING TABLES AND MARGINS USING LEAST-SQUARES

AV HARRY LÜTJOHANN

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1985:18. Reconciling tables and margins using least-squares / Harry Lütjohann.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-18

SCB STATISTISKA CENTRALBYRÅN

1985-08-01

PROMEMORIOR FRÅN P/STM

NR 18

RECONCILING TABLES AND MARGINS USING LEAST-SQUARES

AV HARRY LÜTJOHANN

SCB STATISTICS SWEDEN

Statistical Research Unit

1985-08-01

RECONCILING TABLES AND MARGINS USING LEAST-SQUARES

by Harry Lütjohann

Contents

1 The problem.
2 The solution in outline.
3 Some matrix notation.
4 Regression with exact linear restrictions.
5 The case of exact margins.
6 The case of estimated margins.
7 An historical note.
8 An artificial example.
References.
Appendix. Two computer programmes.

SCB C 124 Arkiv. 9 1982 2 000 ex

ABSTRACT

Data are given from one source for the cells of a two-dimensional table, and from another source for the two margins of the table. The two sets of data do no agree mutually. One wants to reconcile them.

There are two cases. In one case, the margins are given exactly. In the other case, the margins are estimates just like the cells.

The problem can be formalized using the concepts of parameter and estimator. Two solutions emerge, one for either of the two cases. Both solutions apply least-squares.

The approach proposed is intended to replace the traditional approach by means of iterative proportional fitting.

KEY WORDS

tables, margins, least-squares, iterative proportional fitting

1 The problem

Consider a two-dimensional table. For each cell of the table, there is a true number. These true numbers are not known.

Assume that there are two statistical investigations giving information on the table. The two investigations will be called the cells investigation and the margins investigation.

The cells investigations provides one estimate for each cell of the table. These estimates imply estimates of the row sums, the column sums, and the total, of the table.

The margins investigation provides either of two kinds of information on the margins of the table. Either it provides exact information on each row sum and/or each column sum and/or the total of the table. Or else it provides one estimate of each row sum and/or each column sum and/or the total of the table.

Usually, the information on the margins provided by the margins investigation is more precise than that implied by the cells investigation.

There are thus two sets of data relating to the table. Except by accident, these two sets of data are not consistent with each other. The row and column sums, and the total, implied by the cells investigation, do not agree with those given by the margins investigation.

We should use all the information from both investigations. We should use it to provide better estimates of the cells. If the margins are known exactly, the new estimates of the cells should be consistent with them. If the margins are not known exactly, there should be new estimates of the margins consistent with the new estimates of the cells.

The classical formulation of this problem was given by Deming and Stephan [1940]. The solution usually applied is to use an algorithm called Iterative Proportional Fitting (IPF). A recent development in this direction is the Structure Preserving Estimates (SPREE) of Purcell [1979].

Here, another solution will be proposed. It uses least-squares. It does not call for iterative computations. It gives minimum variance linear unbiased estimators for any sample size.

The solu tion proposed is not new, but it seems to be little used in practice. A paper giving an application of it, which is at once more general and much more specific than the present paper, is van der Ploeg [1982].

2 The solution in outline

Let the rows and the columns of the table be indexed $h = 1,\ldots,r$ and $k = 1,\ldots,c$, respectively.

Associated with each cell of the table there is a parameter β_{hk}, the unknown true value of the cell.

The information on the (hk)'th cell given by the cells investigation is denoted y_{hk}. It is assumed that y_{hk} can be regarded as an outcome of an unbiased estimator of the corresponding parameter β_{hk}.

Let us assume that the margins investigation does provide information on the row sums of the table. What will be said applies equally to column sums and total, if information is provided on those.

The information on the h'th row sum provided by the margins investigation will be denoted z_{Rh}. If this information is exact, it gives the true value of the corresponding row sum $\beta_{h1} + \ldots + \beta_{hc}$ of parameters. If it is not exact, it is assumed that z_{Rh} can be regarded as an outcome of an unbiased estimator of this row sum of parameters.

In the case of exact information on the margins, the situation is as follows. There are unbiased estimates for each of rc known (trivial) linear combinations of the rc unknown parameters. There is exact information on each of a number of known linear combinations of the parameters. This is precisely the situation where least-squares with exact linear restrictions is applicable. The solution proposed is, therefore, to apply least-squares with exact linear restrictions. The suggestion to do so was given already by Deming and Stephan [1940] but not pursued by them due to mathematical and computational difficulties. Matrix algebra and the electronic computer have now eliminated these difficulties.

In the case where the information given by the margins investigation is not exact, the situation is as follows. From the cells investigation, there are unbiased estimates for each of rc known (trivial) linear combinations of the rc unknown parameters. From the margins investigation, there are unbiased estimates for each of a number of known linear combinations of the parameters. (In the case where the margins investigation provides information on the row sums only, there are r such estimates.) There are thus two sets of unbiased estimates for known linear combinations of the same set of parameters. This is precisely the situation where least-squares is applicable. Preferably, the least-squares analysis should be weighted so as to take into account any difference in precision within and between the cells investigation and the margins investigation. The solution proposed is, therefore, to apply weighted least-squares. If the weights are not known exactly, they should be estimated or even guessed.

In the case where the information given by the margins investigation is not exact, the solution proposed produces new estimates not only of the cells, but of the margins too.

For simplicity, we assume that any two estimates, whether from the cells investigation or from the margins investigation, are mutually uncorrelated. Very likely, this is not exactly true. But at least it may often be true that the covariances are of a smaller order of magnitude than the variances, so that the assumption is acceptable as an approximation.

The solution proposed is based on one fundamental and critical assumption. This assumption is as follows. Any information given by the two investigations is, if not exact, an estimate without bias. If this assumption is not reasonable, the new solution proposed here is not applicable.

The solution proposed has one practical drawback. It calls for inversion of a matrix the order of which increases with r and/or c. If the number of cells is large, there is risk of numerical inaccuracy.

3 Some matrix notation

3.1 General matrix notation

Matrices will be denoted by capital letters. Vectors will be denoted by lower-case letters. Throughout, any vector symbol written without a transpose sign denotes a column vector.

The orders of vectors and matrices are indicated as follows. An n-vector is a (column) vector of n elements. A pxq matrix is a matrix of p rows and q columns.

The matrix I_n is the nxn unit matrix, i.e. a diagonal matrix, each diagonal element of which is 1.

The vector j_n is an n-vector, each element of which is 1.

Let A be an mxn matrix and L a pxq matrix. Than $A \otimes L$ is the direct product of A and L, the mp x nq matrix defined as follows, where a_{hk} is the element in the h'th row and k'th column of A.

$$A \otimes L = \begin{pmatrix} a_{11}L & a_{12}L & \cdots & a_{1n}L \\ a_{21}L & a_{22}L & \cdots & a_{2n}L \\ \vdots & \vdots & & \vdots \\ a_{m1}L & a_{m2}L & \cdots & a_{mn}L \end{pmatrix}$$

The relation

$$(A \otimes L)' = A' \otimes L'$$

is an immediate consequence of the definition of the direct product.

3.2 Matrix notation relating to least-squares

There are n units of analysis. The order of the design matrix is nxp and it is denoted X. The n-vector of observations is denoted y. Each row in x and y is associated with one unit of analysis.

Any p-vector of regression coefficients is denoted b. Special regression coefficient vectors are indicated by subscripts.

Let a diagonal matrix be called positive if and only if each of its diagonal elements is strictly positive. Any positive diagonal nxn matrix will be denoted Q with diagonal elements q_i, i=1,...,n.

Let Z be an sxp matrix and z an s-vector such that the rank of Z is s. Then

$$Zb = z$$

is a set of s linearly independent linear restrictions on the regression coefficients.

3.3 Matrix notation relating to tables and margins

The subscripts G, R, C and T indicate the table without margins (the Guts of the table), the Row sums, the Column sums, and the Total of the table, respectively.

Data from the cells investigation are denoted by the letters Y and y. Data from the margins investigation are denoted by the letter z.

The table and its margins are regarded and treated as a matrix and two vectors.

The table without margins given by the cells investigation is the r x c matrix Y. Its elements are y_{hk}, where h=1,...,r is the row index and k=1,...,c is the column index.

Take all the elements of Y in the order in which you read the letters of a printed page. Put m = rc. The m-vector consisting of the elements of Y in the order mentioned is denoted y_G. The elements of y_G keep their double-indexing from Y.

The r x c matrix of parameters corresponding to the table without margins is denoted B (read capital beta). Its elements are β_{hk}, indexed like y_{hk}.

Take all the elements of B in the order in which you read a printed page. The m-vector obtained is denoted β. Its elements keep their double-indexing from B.

The row sums implied by the cells investigation form the column r-vector $y_R = Yj_c$ with elements y_{Rh}, h=1,...,r.

The column sums implied by the cells investigation form the row c-vector $y_C' = j_r'Y$ with elements y_{Ck}, k=1,...,c.

The total implied by the cells investigation is the scalar $y_T = j_r'Yj_c$.

The row sums given by the margins investigation (if any) form the column r-vector z_R with elements z_{Rh}, h=1,...,r.

The column sums given by the margins investigation (if any) form the row c-vector z_C' with elements z_{Ck}, k=1,...,c.

If a unique total is given or implied by the margins investigation, it is denoted z_T.

4 Regression with exact linear restrictions

Throughout this section, weighted least-squares (WLS) is considered. Everything said is conditional upon a fixed set of weights given by the positive diagonal matrix Q.

4.1 Unrestricted descriptive least-squares

Consider a given n x p design matrix X of rank p and a given n-vector of observations y. Any p-vector b defines an approximation vector $\hat{y} = Xb$ and a vector of approximation errors $e = y - Xb$.

The weighted sum of squares of the approximation errors is the following function of b.

(4.1) $f(b) = (y-Xb)'Q(y-Xb).$

It is minimized by the p-vector

(4.2) $b_0 = (X'QX)^{-1}X'Qy,$

whose elements are called the WLS regression coefficients.

The subscript 0 on b_0 is a zero intended to indicate that there are no restrictions of the kind to be introduced below.

That b_0 minimizes the approximation error sum of squares is well known. The minimum is called the residual sum of squares.

4.2 Unrestricted least-squares estimation

The classical full rank Linear Model is as follows. The design matrix X is non-stochastic and known. Its rank is equal to its number of columns. The vector of observations y actually observed is regarded as an outcome of a corresponding

stochastic vector, which is usually also denoted y. There is a non-stochastic and unknown p-vector of parameters β. There is a non-stochastic and unknown average disturbance variance σ^2. There is a known non-stochastic positive-definite n x n matrix W.

The Linear Model states that the vector of observations has the expected values.

$$(4.3) \qquad E(y) = X\beta$$

and the covariance matrix

$$(4.4) \qquad V(y) = \sigma^2 W.$$

Throughout this paper, the matrix W will be assumed to be positive diagonal. This gives a subclass of the general Linear Model, the heteroscedastic linear model.

Consider the WLS regression coefficients

$$(4.5) \qquad b_W = (X'W^{-1}X)^{-1}X'W^{-1}y$$

obtained by using the weights $Q = W^{-1}$. These weights are inversely proportional to the variances of the units of analysis.

The following theorem is given without proof: The elements of b_W are the minimum variance linear unbiased estimators of the corresponding elements of the parameter vector β.

This is a variant of the well-known Gauss-Markov theorem of least-squares theory. It is proved in many text-books. It can be derived as a simplified special case of the theorem stated and proved later in this section.

4.3 Restricted descriptive least-squares

Consider a given n x p design matrix X of rank p and a given n-vector of observations y. Consider a given s x p matrix Z of rank s and a given s-vector z jointly defining the linear restrictions

(4.6) $Zb = z$

on the regression coefficients. Assume a given positive diagonal weight matrix Q.

The weighted sum of squares of the approximation errors is (4.1) as above. It is minimized by the p-vector.

(4.7) $b_Z = b_0 - (X'QX)^{-1}Z'[Z(X'QX)^{-1}Z']^{-1}(Zb_0-z),$

where b_0 is as in (4.2). The elements of b_Z are called the restricted WLS regression coefficients.

The restricted least squares regression coefficients are equal to their unrestricted counterparts, minus a correction "proportional to" the amounts by which the unrestricted coefficients fail to satisfy the restrictions.

For the special case $Q = I_n$, the restricted regression coefficients were derived by Theil [1961].

A simple demonstration that b_Z minimizes the weighted sum of squares of the approximation errors will now be sketched.

Consider any p-vector $b_Z + d$ of regression coefficients satisfying the restrictions (4.6). Then $Zd = 0$.

The approximation error sum of squares is

$$f(b_Z+d) = [y-X(b_Z+d)]'Q[y-X(b_Z+d)] =$$

$$= (y-Xb_Z)'Q(y-Xb_Z) + d'X'QXd$$

$$- d'X'Q(y-Xb_Z) - (y-Xb_Z)'QXd.$$

Because $Zd = 0$ it can be shown that

$$d'X'Q(y-Xb_Z) = 0$$

(just substitute (4.7) and develop!), so that

$$f(b_Z+d) = f(b_Z) + d'X'QXd$$

Now $d'X'QXd \geq 0$ with $d'X'QXd = 0$ if and only if $d = 0$.

4.4 Restricted least-squares estimation

Assume the heteroscedastic Linear Model consisting of (4.3) and (4.4).

Assume that the parameter vector is known to satisfy the s linear restrictions

(4.8) $$Z\beta = z$$

analogous to (4.6).

Choose $Q = W^{-1}$.

Consider the restricted WLS regression coefficients b_Z defined in (4.7).

Theorem

The elements of b_Z are the minimum variance linear unbiased estimators of the corresponding elements of the parameter vector β.

Proof.

For convenience, define $M = (X'QX)^{-1}$.

The restricted least-squares regression coefficients are

$$b_Z = Ay + Hz,$$

where A is the p x n matrix

$$A = [M - MZ'(ZMZ')^{-1}ZM]X'Q,$$

while H is the p x s matrix

$$H = MZ'(ZMZ')^{-1}.$$

The restricted least-squares regression coefficients are linear in y and z. Consider any other estimator which is linear in the same sense.

Let D be any p x n matrix and K any p x s matrix. Consider the p-vector

$$b = (A+D)y + (H+K)z$$

of alternative linear estimators of β.

A necessary condition for b to be unbiased estimators of β is that $DX\beta + Kz = 0$, which by (4.8) implies that $(DX + KZ)\beta = 0$. This equation must hold identically in β, so

$$DX + KZ = 0$$

is a necessary condition for unbiasedness.

Consider any vector of linear unbiased estimators of β. Its covariance matrix is

$$V[(A+D)y + (H+K)z] = V[(A+D)y] =$$

$$= \sigma^2 AWA' + \sigma^2 AWD' + \sigma^2 DWA' + \sigma^2 DWD'.$$

Now, since $QW = I$ and $DX = -KZ$, $AWD' = 0$.

Consequently,

$$V(b) = V(b_Z) + \sigma^2 DWD'.$$

Since W is positive-definite, any diagonal element of $\sigma^2 DWD'$ is non-negative. All the diagonal elements are 0 if and only if $D = 0$. Thus

$$D = 0$$

is a necessary condition for minimum variances.

Now, if $D = 0$, then $KZ = 0$ and then, because the rank of Z is s, $K = 0$.

We have shown that the equations

$$D = 0 \text{ and } K = 0$$

are, together, necessary conditions for the general restricted linear estimators b to be unbiased and have minimum variances. That they are also, together, sufficient, is evident.

End of proof

A theorem much more general than ours, which includes ours as a special case, is stated and proved by Theil [1971]. We include a proof here, because it is not quite trivial to derive our theorem from that of Theil.

Finally, a word of explanation. The unrestricted least-squares estimators b_0 have minimum variances in the class of unbiased estimators linear in y. The restricted least-squares estimators b_z have minimum variances in the class of unbiased estimators linear in y and z. The latter class contains the former class as a subset. Thus, the latter minimum may be lower than the earlier one.

In fact, it is easy to show that

$$V(b_0) - V(b_z) =$$

$$= \sigma^2 MZ'(ZMZ')^{-1}ZM,$$

where M is as defined above. The difference matrix is positive-semidefinite, so its diagonal elements are non-negative.

5 The case of exact margins

The cells investigation provides an m-vector of estimates y_G of the cell parameters β of the table.

The margins investigation provides exact knowledge of the r-vector of row sums Z_R and/or the c-vector of column sums Z_C and/or the total Z_T of the r x c parameter matrix B.

Corresponding to the cells investigation there is the linear model obtained by putting $X = I_m$ and $y = y_G$ in (4.3). The model is

$$E(y_G) = \beta,$$

$$V(y_G) = \sigma^2 W.$$

In the estimator formulas, choose $Q = W^{-1}$.

The unrestricted least-squares estimators (4.5) are

$$b_0 = y_G ,$$

i.e. each cell parameter is estimated by the corresponding cell value from the cells investigation.

The estimators to be used are the restricted least-squares estimators (4.7), It remains to determine the restrictions (4.6) to be applied.

5.1 Only row sums given

The margins investigation gives the restrictions obtained by putting

$$Z = I_r \otimes j_c' ,$$

$$z = z_R$$

in (4.6).

Some additional notation is now needed. Associated with each cell of the table there is a variance $\sigma^2 w_{hk}$, $h = 1, \ldots, r$ and $k = 1, \ldots, c$. These variances form an analogous $r \times c$ variance matrix, say $\sigma^2 V$. These variances, taken in the order in which you read the letters of a printed page, are the main diagonal elements of the positive diagonal $m \times m$ matrix $\sigma^2 W$.

Let the h'th row vector of the $r \times c$ matrix of variances V be denoted w_h', $h = 1, \ldots, r$. Define the $m \times r$ block-diagonal matrix

$$W_R^* = \begin{pmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & \ldots\ 0 \\ 0 & 0 & w_3 & \ldots 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots\ w_r \end{pmatrix}$$

Let the h'th row sum of the matrix of variances V be denoted w_{Rh}, $h = 1, \ldots, r$. Define the $r \times r$ positive diagonal matrix

$$W_R = \begin{bmatrix} w_{R1} & 0 & 0 & \ldots 0 \\ 0 & w_{R2} & 0 & \ldots 0 \\ 0 & 0 & w_{R3} & \ldots 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots w_{Rr} \end{bmatrix}$$

So far for additional notation.

Now, apply the formula (4.7) for the restricted least-squares estimators b_Z. We get

$$(5.1) \qquad b_Z = y_G - W(I_r \otimes j_c) \cdot$$

$$\cdot [(I_r \otimes j_c')W(I_r \otimes j_c)]^{-1}[(I_r \otimes j'_c)y_G - z_R]$$

It can be seen that

$$W(I_r \otimes j_c) = W_R^*$$

and that

$$(I_r \otimes j_c')\, W_R^* = W_R.$$

Further,

$$(I_r \otimes j_c')y_G = y_R.$$

Thus, (5.1) simplifies into

$$(5.2) \qquad b_Z = y_G - W_R^* W_R^{-1} (y_R - z_R).$$

This is the matrix expression for the restricted least-squares estimators.

Re-written in scalar notation, our result (5.2) is as follows. For any h and k,

$$(5.3) \qquad b_{hk} = y_{hk} - w_{hk}\left(\sum_{k=1}^{c} w_{hk}\right)^{-1} (y_{Rh} - z_{Rh}).$$

This is the restricted least-squares estimator for the (hk)'th cell.

The estimate given by the estimator can be described verbally as follows. For any cell in the h'th row of the table, consider the difference $y_{Rh} - z_{Rh}$ between the h'th row sum of the cells investigation and the h'th row sum of the margins investigation. Subtract a suitable fraction of this difference from the cell value y_{hk} given by the cells investigation. What fraction? The difference is shared out among the cells of the h'th row in proportion to the cell variances $\sigma^2 w_{hk}$.
By the theorem of section 4, this very simple estimate is the minimum variance linear unbiased estimate when the margins investigation gives only the row sums, but gives them exactly.

5.2 Only column sums given

The margins investigation gives the restrictions obtained by putting

$$Z = j_r' \otimes I_c,$$

$$z = z_C$$

in (4.6)

This case is entirely analogous to the case of given row sums. It will not be developed here.

5.3 Row and column sums given

The margins investigation gives the r row sum restrictions above plus the c column sum restrictions above. But consistency requires that the sum of the row sums is equal to the sum of the column sums. There are thus only $s = r + c - 1$ linearly independent linear restrictions on β.

In order to get a formula to compute from, the simplest way is to drop the last column sum restriction. As a preliminary step, put

$$Z^* = \begin{bmatrix} I_r \otimes j_c' \\ j_r' \otimes I_c \end{bmatrix},$$

$$z^* = \begin{bmatrix} z_R \\ z_C \end{bmatrix}$$

Then drop the last row of Z^* and z^*, and substitute the remaining s x m matrix Z and s-vector z into (4.7).

The actual computation of a solution must be left to an electronic computer. The programming needed is easy, because of the simple structure of Z.

The result is again, by the theorem of section 4, minimum variance linear unbiased estimates.

5.4 A note

In the case of exact margins, the general linearly restricted least-squares estimator (4.7) is simplified as follows. First, $b_0 = y_G$. Second, $(X'QX)^{-1} = W$. Hence,

$$b_Z = y_G - WZ'[ZWZ']^{-1}(Zy_G - z).$$

It may be noted that this is analogous to the linearly restricted minimum modified chi square estimator given, in the context of contingency tables, be Grizzle and Williams [1972].

6 The case of estimated margins

The cells investigation provides an m-vector of estimates y_G of the cell parameters β of the table.

The margins investigation provides an r-vector of estimates z_R of the row sums, and/or a c-vector of estimates z_C of the column sums, and/or an estimate z_T of the total, of the r x c parameter matrix B.

In the linear models below, we use a common average variance σ^2. This is not a limitation.

Corresponding to the cells investigation there is the linear model.

$$E(y_G) = \beta \ ,$$

$$V(y_G) = \sigma^2 W_G \ ,$$

where W_G is a known positive diagonal m x m matrix.

Corresponding to the margins investigation, there are one, two or three additional analogous linear models.

If the margins investigation provides estimates of the row sums, there is the model

$$E(z_R) = (I_r \otimes j_c') \beta \ ,$$

$$V(z_R) = \sigma^2 W_R .$$

If the margins investigation provides estimates of the column sums, there is the model

$$E(z_C) = (j_r' \otimes I_c)\,\beta,$$

$$V(z_C) = \sigma^2 W_C.$$

If the margins investigation provides an independent estimate of the total, there is also the model

$$E(z_T) = j_m'\,\beta,$$

$$V(z_T) = \sigma^2 W_T\,,$$

where W_T is a 1 x 1 matrix.

The models that there are, are (one may say) concatenated into a single heteroscedastic Linear Model involving the parameter vector β. For example, if the margins investigation provides estimates of the row and column sums, the joint model of the cells and margins data is as follows.

$$E\begin{bmatrix} y_G \\ z_R \\ z_C \end{bmatrix} = \begin{bmatrix} I_m \\ I_r \otimes j_c' \\ j_r' \otimes I_c \end{bmatrix} \beta\,,$$

$$V\begin{bmatrix} y_G \\ z_R \\ z_C \end{bmatrix} = \sigma^2 \begin{bmatrix} W_G & 0 & 0 \\ 0 & W_R & 0 \\ 0 & 0 & W_C \end{bmatrix}$$

If the margins investigation is such that the sum of the row sums is always equal to the sum of the column sums, one should, for theoretical correctness, drop the last column sum observation and the last row of the model. But there is probably little harm in keeping them, so as to preserve symmetry.

The estimators to use are the weighted least-squares estimators (4.5) corresponding to the joint model. Here, cells data and margins data are given weight in inverse proportion to their variances.

The cells estimates b_{hk} obtained are linear in the cells and margins observations. By the Gauss-Markov theorem, they are minimum variance linear unbiased estimates.

The row and column sum estimates obtained are the row and column sums of the cells estimates. By a well-known generalisation of the Gauss-Markov theorem, they too are minimum variance linear unbiased estimates.

7 An historical note

In 1940, Deming and Stephan formulated what has been called here the case of exact margins of the problem of reconciling tables and margins; Deming and Stephan [1940].

Deming and Stephan applied least-squares with exact linear restrictions, without the convenient matrix algebra nowadays customary. They considered a two-dimensional table, and proceeded first to treat the situation where the margins investigation gives the row sums only.

Deming and Stephan assumed a heteroscedastic linear model, but a very particular one. They assumed that the variance of the (hk)'th cell observation y_{hk} was proportional to the value observed. In our notation, their assumption was that $w_{hk} = y_{hk}$ for $h = 1, \ldots, r$ and $k = 1, \ldots, c$.

Under this special assumption only, Deming and Stephan derived the restricted least-squares estimators of the cell parameters β_{hk}.

Let us substitute Deming's and Stephan's special assumption into our more general expression (5.3). The outcome is

$$b_{hk} = y_{hk}(y_{Rh})^{-1} z_{Rh},$$

which is Deming's and Stephan's result too.

The expression found says the following. Take the (hk)'th cell observation. Adjust it by a multiplicative factor. The factor to be used is the ratio between the h'th row sum given by the margins investigation and that given by the cells investigation.

This looks very much like a multiplicative correction of the cells observation. But as our more general expression shows, it is in fact a very particular case of an additive (or "subtractive") correction, where two terms cancel.

The Interative Proportional Fitting algorithm is much used in attempts to solve the problem of reconciling tables and margins. The IPF algorithm may be characterized as essentially multiplicative. From Deming and Stephan [1940], it seems that the IPF algorithm was originally inspired by the simple estimator formula derived above for a very special case.

The theoretical basis for the IPF algorithm is perhaps somewhat slender. The theoretical basis for least-squares with linear restrictions seems more firm and adequate. The IPF algorithm may however, be fitted into the new theoretical framework as a numerical device for computing the restricted regression coefficients in the case where the cell variances are assumed to be proportional to the cell values.

8 An artificial example

The table and the margins given are as follows.

102	51	191	350
205	68	86	350
250	112	53	450
297	302	413	1000
900	500	750	

The variances of the cells are assumed to be all equal.

In the case of exact margins, the outcome of the reconciliation is as follows.

114	43	193	350
212	55	83	350
271	114	65	450
303	288	409	1000
900	500	750	2150

In the case of estimated margins, the variance of any row sum is assumed to be 50% of the variance of a cell, and the variance of any column sum is assumed to be 10% of that of a cell.

In this case, the reconciliation gives the following result.

114	43	193	350
212	56	84	352
270	113	63	446
303	289	410	1002
899	501	750	2150

The outcome of the reconciliation is not a set of integers. In the tables above, the numbers obtained are rounded off to the nearest integer. This may cause minor problems, but does not do so in the present example.

References

Deming, W.E. and Stephan, F.F. [1940]: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known.
Annals of Mathematical Statistics, XI, 1940, 427-444.

Grizzle, J.E. and Williams, O.D. [1972]: Log linear models and tests of independence for contingency tables.
Biometrics, 28, 137-156.

van der Ploeg, F. [1982]: Reliability and the adjustment of sequences of large economic accounting matrices.
Journal of the Royal Statistical Society, ser. A, 145, 169-194.

Purcell, N.J. [1979]: Efficient estimation for small domains: A categorical data analysis approach.
Unpublished Ph.D. dissertation, University of Michigan.

Theil, H. [1961]: Economic Forecasts and Policy, 2nd rev.ed.
North-Holland, Amsterdam.

Theil, H. [1971]: Principles of Econometrics.
North-Holland, Amsterdam.

APPENDIX

TWO COMPUTER PROGRAMMES

A.1 An outline

Two computer programmes have been written, one for the case of exact margins and one for the case of estimated margins. The input and output parts of the two programmes are very similar.

Both programmes are limited to the situation where the margins investigation gives information on the row sums and the column sums, no less, no more.

Both programmes assume that the Y and z data read in are integers, snd give results rounded off to integers.

The programming was made inside the SAS procedure MATRIX. When the programmes are to be run, there must be a JCL card telling the computer that it is a SAS programme that is coming.

A.2 The indata.

For the time being, the programmes accept at most 6 rows and at most 6 columns. This restriction is imposed in order to get a convenient input procedure via a telescreen terminal.

The indata are arranged into a 14 x 7 matrix. Any element of this matrix not used for a number must be filled in with a dot. SAS reads the indata as if it were 14 observations on 7 variables, and accepts the dots as missing values.

The r x c matrix Y given by the cells investigation must be placed in rows 1 to r and columns 1 to c.

The row sum vector z_R given by the margins investigation must be placed in rows 1 to r and column c+1.

The column sum c-vector z_C' given by the margins investigation must be placed in row r+1 and columns 1 to c.

The r x c matrix V of variances for the cells investigation must be given and placed in rows 8 to 7+r and columns 1 to c.

In the case of estimated margins, there must also be given variances for the row and column sums z_R and z_C.

The r-vector of row sum variances must be placed in rows 8 to 7+r and column c+1.

The c-vector of column sum variances must be placed in row 8+r and columns 1 to c.

In short, the matrix Y begins in position (1,1) and is bordered by the row and column sums. Analogously, the matrix of cell variances V begins in position (8,1) and is bordered by the row and columns um variances if any.

Finally, the bottom right corner (14,7) of the indata matrix is used to call in some optional printing. If this element is a dot, there is the standard output. If it is 1 or 2, there is additional output, the result matrix with 1 or 2 decimals, respectively.

The 14 x 7 indata are referred to in the programmes as "infile rtmdat". There must be a JCL card telling the computer where to find the data set "rtmdat".

A.3 The outdata.

The output consists of four tables, the last of which is optional.

The first table is the matrix Y given by the cells investigation, bordered by its row and column sum vectors y_R and y_C', the whole again bordered by the row and column sum vectors z_R and z_C' given by the margins investigation.

In the case of exact margins, the first table also gives the total of the table given by the cells investigation, and the total of the row sums given by the margins investigation. It is assumed that the latter total is equal to the total of the column sums given by the margins investigation.

In the case of estimated margins, the first table also gives, instead, the total of the table given by the cells investigation, the total of the row sums given by the margins investigation, and the total of the column sums given by the margins investigation. It is not necessary that the two latter sums are equal.

The second table is the matrix of variances V for the cells investigation, in the case of estimated margins bordered by the vectors of variances for the row and column sums given by the margins investigation.

The third table gives the results of the reconciliation of table and margins. The numbers printed are rounded off to integers.

The third table is the matrix of adjusted Y values, bordered by their row and column sums. In the case of exact margins, these sums should agree with the row and columns sums given by the margins investigation. In the case of estimated margins, in general, they do not so agree.

The fourth table is optional. It is the third table rounded off to one or two decimals.

A.4 The case of exact margins.

```
 1  * RECONCILING TABLES AND MARGINS USING LEAST-SQUARES;
 2  * HARRY LUETJOHANN STATISTICS SWEDEN;
 3  * THE CASE OF EXACT MARGINS;
 4  DATA INDATA;
 5  INFILE RTMDAT;
 6  INPUT V1 V2 V3 V4 V5 V6 V7;
 7  PROC MATRIX;
 8  * READING IN DATA;
 9  FETCH INMATR DATA=INDATA;
10  * DETERMINING ORDERS;
11  R=6;
12  IF INMATR(7,1)=. THEN R=5;
13  IF INMATR(6,1)=. THEN R=4;
14  IF INMATR(5,1)=. THEN R=3;
15  IF INMATR(4,1)=. THEN R=2;
16  C=6;
17  IF INMATR(1,7)=. THEN C=5;
18  IF INMATR(1,6)=. THEN C=4;
19  IF INMATR(1,5)=. THEN C=3;
20  IF INMATR(1,4)=. THEN C=2;
21  M=R*C;
```

```
22  * DEFINING DATA MATRICES AND VECTORS;
23  YM=INMATR(1:R,1:C);
24  YR=YM*J(C,1,1);
25  YCT=J(1,R,1)*YM;
26  YC=YCT';
27  YT=J(1,R,1)*YR;
28  YG=SHAPE(YM,1);
29  ZR=INMATR(1:R,C+1);
30  ZCT=INMATR(R+1,1:C);
31  ZC=ZCT';
32  ZT=J(1,R,1)*ZR;
33  * DEFINING VARIANCES;
34  WM=INMATR(8:7+R,1:C);
35  WV=SHAPE(WM,1);
36  W=DIAG(WV);
37  * ARRANGING INDATA FOR PRINTING;
38  A=.;
39  INDA1=YM||YR;
40  INDA2=INDA1||ZR;
41  INDA3=YCT||YT;
42  INDA4=INDA3||A;
43  INDA5=ZCT||A;
44  INDA6=INDA5||ZT;
45  INDA7=INDA2//INDA4;
46  INDA8=INDA7//INDA6;
47  INDAW=WM;
48  * DEFINING RESTRICTIONS (HB=Z);
49  HR=I(R)@J(1,C,1);
50  HC=J(1,R,1)@I(C);
51  H1=HR//HC;
52  H=H1(1:R+C-1,1:M);
53  Z1=ZR//ZC;
54  Z=Z1(1:R+C-1,);
55  * COMPUTING REGRESSION COEFFICIENTS;
56  WH=W*H';
57  HWH=H*WH;
58  HWHI=INV(HWH);
59  DIF=H*YG-Z;
60  BZ=YG-WH*HWHI*DIF;
61  * ARRANGING OUTDATA FOR PRINTING;
62  YGH=SHAPE(BZ,C);
63  YRH=YGH*J(C,1,1);
64  YCHT=J(1,R,1)*YGH;
65  YTH=J(1,R,1)*YRH;
66  OUDA1=YGH||YRH;
67  OUDA2=YCHT||YTH;
68  OUDA3=OUDA1//OUDA2;
69  * PRINTING;                                    (SQUARES;
70  TITLE1 RECONCILING TABLES AND MARGINS USING LEAST-)
71  TITLE2 HARRY LUETJOHANN STATISTICS SWEDEN;
72  TITLE3 THE CASE OF EXACT MARGINS;
73  PRINT INDA8 INDAW OUDA3 FORMAT=6.0;
74  IF INMATR(14,7)=1 THEN PRINT OUDA3 FORMAT=8.1;
75  IF INMATR(14,7)=2 THEN PRINT OUDA3 FORMAT=9.2;
```

A.5 The case of estimated margins.

```
 1  * RECONCILING TABLES AND MARGINS USING LEAST-SQUARES;
 2  * HARRY LUETJOHANN STATISTICS SWEDEN;
 3  * THE CASE OF ESTIMATED MARGINS;
 4  DATA INDATA;
 5  INFILE RTMDAT;
 6  INPUT V1 V2 V3 V4 V5 V6 V7;
 7  PROC MATRIX;
 8  * READING IN DATA;
 9  FETCH INMATR DATA=INDATA;
10  * DETERMINING ORDERS;
11  R=6;
12  IF INMATR(7,1)=. THEN R=5;
13  IF INMATR(6,1)=. THEN R=4;
14  IF INMATR(5,1)=. THEN R=3;
15  IF INMATR(4,1)=. THEN R=2;
16  C=6;
17  IF INMATR(1,7)=. THEN C=5;
18  IF INMATR(1,6)=. THEN C=4;
19  IF INMATR(1,5)=. THEN C=3;
20  IF INMATR(1,4)=. THEN C=2;
21  M=R*C;
22  * DEFINING DATA MATRICES AND VECTORS;
23  YM=INMATR(1:R,1:C);
24  YR=YM*J(C,1,1);
25  YCT=J(1,R,1)*YM;
26  YC=YCT';
27  YT=J(1,R,1)*YR;
28  YG=SHAPE(YM,1);
29  ZR=INMATR(1:R,C+1);
30  ZCT=INMATR(R+1,1:C);
31  ZC=ZCT';
32  ZTR=J(1,R,1)*ZR;
33  ZTC=ZCT*J(C,1,1);
34  * DEFINING WEIGHT (INVERSE VARIANCE) MATRICES;
35  WM=INMATR(8:7+R,1:C);
36  WV=SHAPE(WM,1);
37  WGD=DIAG(WV);
38  QG=INV(WGD);
39  WR=INMATR(8:7+R,C+1);
40  WRD=DIAG(WR);
41  QR=INV(WRD);
42  WCT=INMATR(8+R,1:C);
43  WC=WCT';
44  WCD=DIAG(WC);
45  QC=INV(WCD);
46  * ARRANGING INDATA FOR PRINTING;
47  A=.;
48  INDA1=YM||YR;
49  INDA2=INDA1||ZR;
50  INDA3=YCT||YT;
51  INDA4=INDA3||ZTR;
52  INDA5=ZCT||ZTC;
53  INDA6=INDA5||A;
54  INDA7=INDA2//INDA4;
55  INDA8=INDA7//INDA6;
56  INDAW=INMATR(8:8+R,1:C+1);
```

```
57  * DEFINING DESIGN MATRICES (H);
58  HR=I(R)@J(1,C,1);
59  HC=J(1,R,1)@I(C);
60  * COMPUTING REGRESSION COEFFICIENTS;
61  XQXG=QG;
62  XQXR=HR'*QR*HR;
63  XQXC=HC'*QC*HC;
64  XQX=XQXG+XQXR+XQXC;
65  XQYG=QG*YG;
66  XQYR=HR'*QR*ZR;
67  XQYC=HC'*QC*ZC;
68  XQY=XQYG+XQYR+XQYC;
69  B=SOLVE(XQX,XQY);
70  * ARRANGING OUTDATA FOR PRINTING;
71  YGH=SHAPE(B,C);
72  YRH=YGH*J(C,1,1);
73  YCHT=J(1,R,1)*YGH;
74  YTH=J(1,R,1)*YRH;
75  OUDA1=YGH||YRH;
76  OUDA2=YCHT||YTH;
77  OUDA3=OUDA1//OUDA2;
78  * PRINTING;                                    (SQUARES;
79  TITLE1 RECONCILING TABLES AND MARGINS USING LEAST-)
80  TITLE2 HARRY LUETJOHANN STATISTICS SWEDEN;
81  TITLE3 THE CASE OF ESTIMATED MARGINS;
82  PRINT INDA8 INDAW OUDA3 FORMAT=6.0;
83  IF INMATR(14,7)=1 THEN PRINT OUDA3 FORMAT=8.1;
84  IF INMATR(14,7)=2 THEN PRINT OUDA3 FORMAT=9.2;
```

A.6 A reference.

SAS USER'S GUIDE
1979 EDITION
SAS Institute Inc., Raleigh NC 27605.

A.7 On matrix inversion.

Both programmes call for inversion of matrices.

In the programme for the case of exact margins, the order of the matrix inverted is r+c-1. (See line 58 of the programme.)

In the programme for the case of estimated margins, the order of the matrix inverted is m=rc. (See line 69 of the programme.) In addition, this programme inverts three diagonal matrices of the same order m. (See lines 38, 41 and 45.)

The matrices to be inverted are in neither case functions of the really observed data Y and z.

A.8 An artificial example:
The indata "rtmdat".

102	51	191	350	.	.	.
205	68	86	350	.	.	.
250	112	53	450	.	.	.
297	302	413	1000	.	.	.
900	500	750	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
100	100	100	50	.	.	.
100	100	100	50	.	.	.
100	100	100	50	.	.	.
100	100	100	50	.	.	.
10	10	10	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

In the case of exact margins, the variances given for margins data are not used.

When the case of estimated margins was run, element (14,7) of the indata was put equal to 1 so as to get information on the rounding-off effects.

A.9 An artificial example:
Exact margins, the output.

RECONCILING TABLES AND MARGINS USING LEAST-SQUARES
HARRY LUETJOHANN STATISTICS SWEDEN
THE CASE OF EXACT MARGINS

INDA8	COL1	COL2	COL3	COL4	COL5
ROW1	102	51	191	344	350
ROW2	205	68	86	359	350
ROW3	250	112	53	415	450
ROW4	297	302	413	1012	1000
ROW5	854	533	743	2130	.
ROW6	900	500	750	.	2150

INDAW	COL1	COL2	COL3
ROW1	100	100	100
ROW2	100	100	100
ROW3	100	100	100
ROW4	100	100	100

OUDA3	COL1	COL2	COL3	COL4
ROW1	114	43	193	350
ROW2	212	55	83	350
ROW3	271	114	65	450
ROW4	303	288	409	1000
ROW5	900	500	750	2150

A.10 An artificial example: Estimated margins, the output.

RECONCILING TABLES AND MARGINS USING LEAST-SQUARES
HARRY LUETJOHANN STATISTICS SWEDEN
THE CASE OF ESTIMATED MARGINS

INDA8	COL1	COL2	COL3	COL4	COL5
ROW1	102	51	191	344	350
ROW2	205	68	86	359	350
ROW3	250	112	53	415	450
ROW4	297	302	413	1012	1000
ROW5	854	533	743	2130	2150
ROW6	900	500	750	2150	.

INDAW	COL1	COL2	COL3	COL4
ROW1	100	100	100	50
ROW2	100	100	100	50
ROW3	100	100	100	50
ROW4	100	100	100	50
ROW5	10	10	10	.

OUDA3	COL1	COL2	COL3	COL4
ROW1	114	43	193	350
ROW2	212	56	84	352
ROW3	270	113	63	446
ROW4	303	289	410	1002
ROW5	899	501	750	2150

OUDA3	COL1	COL2	COL3	COL4
ROW1	113.5	43.2	193.0	349.8
ROW2	212.2	56.0	83.7	351.9
ROW3	269.8	112.5	63.3	445.6
ROW4	303.4	289.1	409.9	1002.3
ROW5	898.9	500.8	749.9	2149.6

In this case, there are no rounding-off problems.

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

☆Garnisonstryckeriet Stockholm 1985☆