AN ALGORITHM FOR ADJUSTING n-DIMENSIONAL TABULAR DATA TO CONFORM TO GENERAL LINEAR CONSTRAINTS

Kevin Hollenbeck, Mathematica Policy Research, Inc.*

1. Introduction

The purpose of this paper is to report progress on a computational algorithm being developed which adjusts tabular data to conform to linear constraints. A common application for which the algorithm is suited is the statistical adjustment of data to known marginal constraints. The particular problem which motivated this research, however, was the adjustment of data subject to the constraints that the ratios of some cell entries were to be kept constant. The paper provides, in turn, an example of a problem for which the algorithm was developed, the analytical formulation and solution of the general class of problems, and the computational experience gained to date in implementing the solution technique.

Briefly, the advantages of the proposed solution algorithm besides the fact that it allows general linear constraints on the cell entries and not just marginal constraints, are that it:

a) allows and adjusts data elements which are zero,

b) identifies redundant or inconsistent constraints, thus decreasing computer time and avoiding the possibility of nonconvergence,

c) weights the adjustment for each element,

d) gives an exact solution in a non-iterative process,

e) yields estimates which satisfy the constraints to the problem and may be useful as initial estimates in iterative procedures to derive non-least squares estimates.

2. An Example Problem

A cross-tabulation of data from the May, 1976 Current Population Survey (CPS) by the attributes of employment status and age, where employment status is classified into the mutually exclusive and exhaustive classes of employment, unemployment, not in the labor force, and noncivilian status is given in table 1.

It will be noted that the civilian labor force in May was comprised of 94.216 million persons and the (non-seasonally adjusted) unemployment rate was 6.69 percent.

Suppose that a projection of the same table to May, 1977 was desired. One of the major econometric models has forecast that in quarter 2 of 1977, the total noninstitutional population 16 and over will be 158.3 million, the civilian labor force will have 96.7 million members, the overall civilian unemployment rate will be 6.4 percent, the civilian unemployment rate of persons 16-19 will be 16.8 percent, and the civilian labor force participation rate of persons 16-19 will be 56.0 percent.[1] The problem is to adjust the table entries so that they conform to the five constraints imposed by the macroeconomic forecasts. If the entries in table 1 are labelled y_{ij}, $i=1,\ldots,4$; $j=1,\ldots,4$, then the five constraints can be written as

$$(1)\quad \sum_{i=1}^{4}\sum_{j=1}^{4} y_{ij} = 158{,}300$$

$$(2)\quad \sum_{i=1}^{4}\sum_{j=1}^{4} y_{ij} = 96{,}700$$

$$(3)\quad -.064 \sum_{i=1}^{4} y_{i1} + .936 \sum_{i=1}^{4} y_{i2} = 0$$

$$(4)\quad -.168\, y_{11} + .832\, y_{12} = 0$$

$$(5)\quad .44\, y_{11} + .44\, y_{12} - .56\, y_{13} = 0.$$

The solution then is to minimize a distance criterion function subject to (1) through (5).

3. General Problem Formulation and Solution

The general problem is set up and analytically solved in this section of the paper.

Notation -

D - an n-dimensional matrix of observed data in which each dimension represents an attribute of the data entities (such as age, race, sex, employment status, etc.) and in which each dimension is divided into mutually exclusive and exhaustive classes of the attribute values,

c_i; $i=1,\ldots,n$ - the number of classifications in dimension i; note that c_i need not be same for all i,

$q = \prod_{i=1}^{n} c_i$ - number of elements in D,

x - a (q x 1) vector of the elements of D reordered to form a vector,

y - a (q x 1) vector of the estimates which satisfy the constraints,

E - the n-dimensional matrix of estimates,

M,m - an (a x q) matrix and (a x 1) vector which express the a marginal constraints on the problem; note a may equal 0,

N,n - a (b x q) matrix and (b x 1) vector which express the b nonmarginal constraints; note b may equal 0,

TABLE 1

EMPLOYMENT STATUS IN MAY, 1976 OF THE NONINSTITUTIONAL POPULATION 16 YEARS OF AGE AND OVER, BY AGE
(numbers in thousands)

Age	Employment Status				
	Employed	Unemployed	Not in Labor Force	Noncivilian	Total
16 - 19	7,732	1,434	7,886	368	17,420
20 - 24	12,208	1,501	4,905	808	19,422
25 - 64	65,241	3,236	28,338	964	97,779
65 +	2,731	133	18,857	0	21,721
TOTAL	87,912	6,304	59,986	2,140	156,342

S - an n-dimensional matrix of weights for each element of D; these weights are "indicators" of the errors associated with the cell entries, and in most applications, are directly proportional to the cell frequency counts, and

$\hat{W}$ - a (q x q) diagonal matrix of weights associated with the elements of x (i.e., the diagonalization of the vector that results when the elements of S are reordered in the same fashion as the elements of D to form x).

There are several criteria of "closeness" which may be used to adjust the data. Ireland and Kullback [6] demonstrate that the method of iterative proportions suggested by Deming and Stephan [2] minimizes discrimination information. Stephan [8] developed an algorithm which provides a least squares solution. More recently, Feinberg and Holland [3] consider a Bayesian approach, Grizzle, Starmer, and Koch [4] assume a linear model on functions of the cell probabilities and employ least squares, and Brown and Muenz [1] propose a reduced mean square error estimation technique for two dimensional contingency table.

The present study has been limited to using a weighted least squares criterion to exploit the linearity of the gradient of the objective function. (Exploration of the question of whether the analytical and computational solution techniques are applicable to other criteria is intended). The weighted least squares problem is:

$$(6)\quad \min_{e} \sum_{i=1}^{c_1} \sum_{j=1}^{c_2} \cdots \sum_{m=1}^{c_n} s_{ij\ldots m}(e_{ij\ldots m} - d_{ij\ldots m})^2,$$

where $s_{ij\ldots m}$, $e_{ij\ldots m}$, $d_{ij\ldots m}$ are typical elements of S, E, and D, and subject to a marginal constraints and b nonmarginal constraints.

Reordering the elements of D and E into vectors x and y and reordering the elements of S into the diagonal matrix $\hat{W}$, the problem can be stated as:

$$(7)\quad \min_{y} (y - x)^T \hat{W} (y - x)$$

$$\text{s. t. } M y = m$$
$$N y = n.$$

Analytically the problem is solved using straigntforward constrained maximization techniques. Form the Lagrangean function L:

$$(8)\quad L = (y - x)^T \hat{W} (y - x) - \Lambda_1(m - My) - \Lambda_2(n - Ny).$$

The first order conditions for a solution lead to the following set of simultaneous equations:

$$\begin{pmatrix} \hat{W} & M^T & N^T \\ M & 0 & 0 \\ N & 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ \Lambda_1 \\ \Lambda_2 \end{pmatrix} = \begin{pmatrix} \hat{W}x \\ m \\ n \end{pmatrix} . \quad (9)\ (10)\ (11)$$

The second order conditions for a minimum are satisfied as long as the elements of $\hat{W}$ are positive. Premultiplying (9) by $M\hat{W}^{-1}$ and subtracting the result from (10) and premultiplying (9) by $N\hat{W}^{-1}$ and subtracting the result from (11) and then multiplying the Λ_1 and Λ_2 vectors by -1 yields the system:

$$
\begin{matrix}(9')\\(10')\\(11')\end{matrix}
\begin{pmatrix} I & -\hat{W}^{-1}M^T & -\hat{W}^{-1}N^T \\ 0 & M\hat{W}^{-1}M^T & M\hat{W}^{-1}N^T \\ 0 & N\hat{W}^{-1}M^T & N\hat{W}^{-1}N^T \end{pmatrix}
\begin{pmatrix} y \\ \Lambda_1 \\ \Lambda_2 \end{pmatrix}
=
\begin{pmatrix} x \\ m-mX \\ n-nX \end{pmatrix}.
$$

To solve the problem, it suffices to solve the subsystem of equations (10') and (11'). This is done by Gaussian elimination on the (a+b) x (a+b) symmetric matrix

$$
\begin{pmatrix} M\hat{W}^{-1}M^T & M\hat{W}^{-1}N^T \\ N\hat{W}^{-1}M^T & N\hat{W}^{-1}N^T \end{pmatrix}.
$$

By back substitution, we derive estimates of y:

$$
(12)\quad y = \hat{W}^{-1}M^T\Lambda_1 + \hat{W}^{-1}N^T\Lambda_2 + x.
$$

A very nice feature of this algorithm is that if the Gaussian elimination procedure to solve the subsystem of equations (10') and (11') cannot eliminate a row while it is pivotting because all of the elements are zero, then this implies that the associated constraint is redundant or inconsistent and that constraint is eliminated from consideration. This decreases computer time and may point out inconsistencies to the analyst.

4. Computational Experience

A program to solve the data adjustment problem using the above technique has been written for an IBM 370 system.[2] The program, LFNJUST,[3] was used to solve the above example problem. The data for the problem are given in table 2. The solution is given in table 3. It can be seen in that table that, except for rounding error, the projected population is 158.3 million; the civilian labor force is 96.7 million; the civilian unemployment rate is 6,187/96,699.8 = 6.40 percent; the civilian unemployment rate of persons 16-19 is 1,610.7/9,597.4 - 16.80 percent; and the civilian labor force participation rate of persons 16-19 is 9,587.4/17,120.4 = 56.0 percent. Thus all the constraints are satisfied.

5. Conclusions

The LFNJUST algorithm and program have demonstrated the utility of directly using constrained optimization techniques rather than iterative algorithms for the adjustment of tabular data to conform to linear constraints. Because it adjusts cell entries rather than frequencies, it is particularly useful when reweighting survey data. Additional study is warranted to consider whether the technique is efficient when minimum discrimination information estimates are desired and to consider multivariate table adjustment.

TABLE 2

DATA FOR EXAMPLE PROBLEM

$$
X^T = (\ 7732\quad 12208\quad 65241\quad 2731\quad 1434\quad 1501\quad 3236\quad 133\quad 7886\quad 4905\quad 28338\quad 18857\quad 368\quad 808\quad 964\quad 0)
$$

$$
\hat{W}^{-1} \cdot X = \begin{pmatrix} 7732 \\ 12208 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 964 \\ \cdot \\ 0 \end{pmatrix}
$$

$$
M = \begin{pmatrix} 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0 \end{pmatrix}
$$

$$
N = \begin{pmatrix} -.064 & -.064 & -.064 & -.064 & -.936 & -.936 & -.936 & -.936 & 0&0&0&0&0&0&0&0 \\ -.168 & 0 & 0 & 0 & .832 & 0 & 0 & 0 & 0&0&0&0&0&0&0&0 \\ .440 & 0 & 0 & 0 & .440 & 0 & 0 & 0 & -.560&0&0&0&0&0&0&0 \end{pmatrix}
$$

$$
m = \begin{pmatrix} 158,300 \\ 96,700 \end{pmatrix}
$$

$$
n = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
$$

TABLE 3

ESTIMATED EMPLOYMENT STATUS IN MAY, 1977 OF THE NONINSTITUTIONAL POPULATION OVER 15 YEARS OF AGE, BY AGE
(Numbers in thousands)

	Employment Status				
Age	Employed	Unemployed	Not in Labor Force	Noncivilian	Total
16 - 19	7,976.7	1,610.7	7,533.0	366.8	17,487.2
20 - 24	12,566.5	1,411.0	4,889.4	805.4	19,672.3
25 - 64	67,156.7	3,042.0	28,247.6	960.9	99,407.2
65+	2,811.2	125.0	18,796.9	0.0	21,733.1
TOTAL	90,511.1	6,188.7	59,466.9	2,133.1	158,299.8

FOOTNOTES

*The research on which this paper is based was supported by the Socioeconomic Impact Division, Office of Economic Impact, Federal Energy Administration. Computational assistance was provided by Marjorie Odle.

1. The estimates were computed from the Data Resources, Inc. (DRI) CONTROL0524 forecast published in The Data Resources Review, June 1976 (Lexington, Mass.: Data Resources, Inc.).

2. The program was written by Marjorie Odle of The Hendrickson Croporation and is in the testing stage. It will be fully documented and available from the author on request after December 1, 1976. The program utilizes GELS, an IBM-developed procedure to solve a system of linear equations, in which the coefficient matrix is symmetric. See IBM [5].

3. The name is derived from its predecessor, NJUST, a least squares iterative adjustment algorithm developed by the Office of Research and Statistics, Social Security Administration, (see Pugh, Tyler, and George [7]), and because it will be applied to labor force adjustments.

REFERENCES

Brown, C. C. and Muenz, L. R., "Reduced Mean Square Error Estimation in Contingency Tables," Journal of the American Statistical Association, 71 (March 1976), 176-182.

Deming, W. E. and Stephan, F. F., "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known," Annals of Mathematical Statistics, 11 (1940), 427-44.

Feinberg, S. E. and Holland, P. W., "Simultaneous Estimation of Multinomial Cell Probabilities," Journal of the American Statistical Association, 68 (September 1973), 683-91.

Grizzle, J. E., Starmer, C.F., and Koch, G. G., "Analysis of Categorical Data by Linear Models," Biometrics, 25 (September 1969), 489-504.

International Business Machines, Inc., System/360 Scientific Subroutine Package Version III (White Plains, New York: IBM, Inc., 1968).

Ireland, C. T. and Kullback, S., "Contingency Tables with Given Marginals," Biometrics, 55 (March 1968), 179-88.

Pugh, R. E., Tyler, B. S., George, S., "Computer-Based Procedure for the n-Dimensional Adjustment of Data--NJUST," Unpublished Paper, Office of Research and Statistics, Social Security Administration (May 1974).

Stephan, F. F., "Iterative Method of Adjusting Sample Frequency Tables When Expected Margins are Known," Annals of Mathematical Statistics, 13 (1942), 166-78.