A Regression Formulation of the Matrix Estimation Problem

SUE McNEIL
Princeton University, Princeton, New Jersey

CHRIS HENDRICKSON
Carnegie-Mellon University, Pittsburgh, Pennsylvania

Matrices are widely used in transportation planning to represent the distribution of characteristics or as origin-destination matrices. Developing such matrices by means of surveys is expensive and time consuming, and once the survey data are collected and compiled the matrices are rapidly outdated. Other methods which are commonly used are unable to include all available data or to provide a measure of the uncertainty of the estimates. This paper formulates a quadratic programming method to estimate matrix entry estimates as an equivalent constrained generalized least squares estimation problem. As well as being able to include any available information in the form of constraints, the variance-covariance matrix of the entry estimates may be found and confidence intervals calculated for matrix entry estimates with some added distributional assumptions. The problem of updating the proportions of nationwide automobile trips by purpose and trip length from 1970 to 1977 is included as a simple example to illustrate the method.

INTRODUCTION

Matrices are widely used in transportation planning in many different forms. Origin/destination matrices with each entry representing the travel from a particular zone and to some other zone are an important intermediate output in the application of the Urban Transportation Planning System (UTPS). Route or corridor specific origin/destination matrices are used in transit route planning and for forecasting changes in travel. Matrices representing the distribution of traffic or trips by such characteristics as vehicle type and roadway category are used in highway cost allocation studies. Supplemental matrices indi-
cating the impact of travel on energy consumption or other attributes are also employed.\textsuperscript{[8]}

Entries in these matrices are commonly estimated using surveys, but such surveys are time consuming and expensive. To overcome these problems, three general techniques have been used. First, existing matrices may be updated to current values to reflect changes over time. Second, incomplete data or data from a narrow application might be expanded to form matrix entry estimates. Finally, matrix estimates might be obtained ab initio using a distribution function such as a gravity model. Each of these three estimation procedures relies upon some current information such as row and column totals or the attributes within a distribution function.

One of six methods is usually employed to accomplish matrix estimation, updating and expansion. The six methods are:

1. Survey or resurvey can be used to estimate a matrix for any existing situation. As already mentioned, surveys are time consuming and expensive. Matrices compiled from surveys may even require an update as soon as they become available. Alternatively, small sample surveys may be used to reduce costs, but they will be less accurate.

2. Ad hoc expansions are applied in practice as quick and inexpensive means of updating and expanding matrices. For example, each matrix entry might be multiplied by a constant factor indicating population growth from the base year to the study year.

3. Distribution models calculate each matrix entry as a function of zonal characteristics. In particular, gravity models are widely used to estimate origin/destination matrices.

4. The RAS, biproportional\textsuperscript{[2]} or doubly constrained distribution technique is used to update or expand an existing matrix. The new entries are proportional to the base entries and consistent with current row and column totals.

5. Optimization methods such as linear programming\textsuperscript{[12]} and entropy maximization or information minimization\textsuperscript{[23]} can be used. The matrix entry estimates vary with the actual objective function used but in all cases they are consistent with the available information.

6. A distribution model (described in 3 above) may provide the values for the base matrix and then be combined with the biproportional method (described in 4 above).

Each of these methods has advantages but no one method is completely satisfactory. The biproportional and ad hoc methods cannot easily accommodate additional constraints on the variables. The estimation of parameters in the gravity model presents a problem. Furthermore, SIKDAR AND HUTCHINSON\textsuperscript{[16]} question the accuracy of the gravity model in transportation studies. Thus, the methods outlined above are widely used.
but may not produce reliable estimates or even estimates which are consistent with all available data. Finally, it is difficult to develop estimates of the uncertainty of the entry estimates obtained from the application of these methods. A further comparison of the alternative methods appears in McNeil and Hendrickson.\cite{13}

Carey et al.\cite{6} suggested a quadratic programming optimization method to estimate origin/destination trip matrices. Their method minimizes the sum of squared deviations from a distribution model or base year matrix entries, subject to restrictions on row totals, column totals or general observations of volumes. As part of the estimation problem, any unknown parameters in the distribution function are also estimated. This paper is an extension and generalization of the work by Carey et al. Their quadratic programming formulation is equivalently reformulated as a constrained generalized least squares estimation problem, permitting the uncertainty of the estimates to be forecast as well as explicitly including all available information in the form of constraints.

The plan of the paper is as follows. The model is formulated with a distribution function in Section 1 and the solution proposed in Section 2. An example is presented in Section 3. The least squares method is compared with the biproportional method in Section 4. An appendix summarizing the notation used throughout the paper is also included.

1. PROBLEM FORMULATION

The objective of matrix entry estimation, either for obtaining estimates ab initio or for expanding or updating existing estimates, is to form a consistent estimate, \( \hat{Q} \), of the unknown matrix \( Q \) representing the actual entries. The estimate, \( \hat{Q} \) is required to be consistent with all available and relevant information and generally has all nonnegative entries.

To formulate the estimation problem let the matrices \( Q \) and \( \hat{Q} \) be rearranged to form vectors, \( q \) and \( \hat{q} \). Associated with these vectors, \( q \) and \( \hat{q} \) is an estimated value of \( q \), denoted \( y \). The vector \( y \) can be referred to as a vector of base values and is obtained from a distribution function or direct demand function representing a functional relationship between attributes such as socioeconomic variables associated with the matrix entries. The function may be characterized as

\[
y = f(X, \alpha)
\]

where \( f \) is a distribution function,

\( \alpha \) = a vector of parameters, and 
\( X \) = a matrix of characteristics with one row for each entry of the vector \( q \).
The gravity model is one possible form for the function \( f(X, \alpha) \). Other forms of the distribution function for transportation applications are discussed in Carey et al.\(^6\). We will assume that the parameters \( \alpha \) of \( f(X, \alpha) \) are known, although distribution functions with unknown parameters are a direct extension of the results presented here using the approach of Carey et al.\(^6\) and are presented in MCNEIL.\(^{14}\) The special case where \( y \) is some former value of \( q \) such as an outdated origin/destination matrix is a linear function of the former value with zero intercept and unit slope.

If \( Q \) is an \( nm \) matrix then \( q, \hat{q} \) and \( y \) are \( nm \) vectors. For example, if the columns of \( Q \) are stacked end to end then:

\[
q' = [q_{11}, q_{21}, q_{31}, \ldots, q_{n1}, q_{12}, \ldots, q_{y}, \ldots, q_{nm}]
\]

where \( q_y \) is an element of \( Q \).

We assume \( p \) linear constraints exist, given by

\[
Rq = r
\]

where \( R \) is a \( p \times (nm) \) matrix representing the incidence of elements in each constraint, and \( r \) is a \( p \)-vector of constraint values.

Any information that may be expressed as a linear combination of the entries of \( q \) may be included in the constraints. Examples include:

1. Observations of some particular matrix entries.
2. Observations of the sums of a mutually exclusive set of matrix entries. For origin-destination matrices this is known as a cut volume and may be found by observing flows on individual links defined by the cutset.\(^7\)
3. Observations of row and/or column totals. For origin-destination matrices, these totals represent the total amount of traffic originating in or destined for particular zones. These observations are actually a special case of 2 but are so common they are classed as a separate type.

Only linearly independent constraints should be included. For example, if there are \( n \) row totals and \( m \) column totals available then there are \( n + m - 1 \) linearly independent constraints. Assuming \( q \) represents the columns of \( Q \) stacked end to end and \( Q \) is a three by three matrix, then if we choose 3 row constraints and 2 column constraints as the linearly independent constraints, the matrix \( R \) is:

\[
R = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0
\end{bmatrix}
\]
The least squares or quadratic programming formulation minimizes the sum of the weighted squared differences between the values to be estimated and base values subject to a set of constraints. Algebraically, the problem can be stated as a quadratic programming problem:

\[ P1: \text{Minimize} (y - q)' V^{-1} (y - q) \]  
\[ \text{subject to } Rq = r \]

where \( V \) is a matrix of weights and the other notation has been defined above and also appears in Appendix I.

If \( V \) is an identity matrix \( (V = I) \) then it can be shown that the objective function of \( P1 \) is equivalent to a constrained ordinary least squares estimation problem. If \( V \) is a matrix with diagonal elements corresponding to the entries in \( y \) and off-diagonal elements equal to zero, then \( P1 \) has a chi-squared objective function to minimize the sum of the squared deviations divided by the values of \( y_i \).

Equivalently, \( P1 \) may be written as a Lagrangian problem:

\[ \text{Minimize } L = (y - q)' V^{-1} (y - q) - (Rq - r)' \lambda \]

where \( \lambda \) is a vector of Langrange multipliers. The objective of the quadratic programming estimation problem is to find values of \( q \) and \( \lambda \) which minimize the Lagrangian function, \( L \).\[^{1,24} \]

The quadratic programming problem \( P1 \) has an interesting and useful interpretation as a constrained generalized least squares (CGLS) regression. Least squares regression attempts to minimize the sum of the squared differences between observed and estimated variable values. Constraints may be applied on coefficients appearing in the regression equation. In order to emphasize the equivalence of the quadratic problem and CGLS regression, we formulate a general constrained regression problem and then formulate our estimation problem as a CGLS regression.

The CGLS regression model may be formulated as

\[ y = X\beta + \epsilon \]

Equation 4 is subject to \( p \) constraints on the coefficients \( \beta \) given by

\[ R\beta = r \]

where \( R \) and \( r \) are as above,

\( y \) = an \( n \) vector of observed endogenous or dependent variables,  
\( X \) = an \( n \times k \) matrix of observations of exogenous or independent variables,  
\( \beta \) = a \( k \) vector of parameters or coefficients to be estimated, and  
\( \epsilon \) = an \( n \) vector of errors which account for errors in observations of
y, the inherent variability of y, specification errors in the function and excluded variables.

Given the matrix X, the expected value of ε is assumed to be zero and the structure of the covariance matrix of ε is assumed to be known except for an unknown factor of proportionality, σ². Algebraically this can be expressed as

\[ E(\epsilon | X) = 0, \text{ and} \]
\[ \text{Var}(\epsilon | X) = \sigma^2 V \]

where σ² is an unknown parameter, and V is a known \( n \times n \) matrix with the trace of V equal to n.

The generalized least squares problem is to find estimates of β which minimize the weighted squared difference between y and Xβ. The assumptions associated with this formulation and some general solution results may be found in Theil.\[18\]

If the unknown values in the vector q are thought of as parameters or coefficients (represented by β in Equation 4) of variables which can only take the value zero or one, then minimizing the problem P1 is the same as finding the constrained least squares estimates for the “coefficients” q (or β), based on observations of the zero-one valued independent variables. These binary variables are one if the entry in vector q corresponds to an entry in the vector y and zero otherwise.

Thus, the matrix entry estimation problem’s regression equation is:

\[ y = Iq + \epsilon \]

where I is the identity matrix, and ε is a vector of errors.

In effect, I is the set of observations of the binary, independent variables described above, but the problem is nontrivial because of the constraints on q in Equation 2.

If we assume

\[ E(\epsilon) = 0, \text{ and} \]
\[ \text{Var}(\epsilon) = \sigma^2 V \]

then the problem is a constrained generalized least squares estimation of the type described above and in Theil.\[18\]

2. SOLUTION OF THE ESTIMATION PROBLEM

The solution to problem P1 may be found by solving the Lagrangian equation (3). The solution is a linear function of q and r:

\[ q = y + Vr'QV^{-1}(r - R_y) \]

The Kuhn-Tucker sufficient conditions\[4\] ensure that Equation 8 is a
global optimal solution to $P1$ if the matrix of weights, $V$, is nonsingular and positive definite, which is likely to be true for typical specifications of $V$.

Using results from constrained generalized least squares regression$^{[18]}$, it can be shown that Equation $8$ is the best linear unbiased estimate of $q$ for the problem

$$y = Iq + \epsilon$$
subject to $Rq = r$

where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2V$.

The estimate $q$ is best in the sense that the covariance matrix of any other linear unbiased estimator exceeds that of $q$ by a positive definite matrix. Furthermore, the variance-covariance matrix of $q$ is given by

$$\text{Var}[q] = \sigma^2[V - VR'(RVR')^{-1}RV] \quad (9)$$

and an unbiased estimator of $\sigma^2$ is

$$\hat{\sigma}^2 = (y - Iq)'V^{-1}(y - Iq)/p \quad (10)$$

where $p$ is the number of constraints. The degrees of freedom of $\hat{\sigma}^2$ are equal to the number of observations in $y$ plus the number of constraints less the number of parameters to estimate in $q$.$^{[18]}$

Thus, the solution to the quadratic estimation problem, $P1$, may be obtained through a series of matrix operations on $y$, $V$, $R$, and $r$ as given by Equation 8. Commonly available computer software programs such as the International Mathematical and Statistical Library, IMSL$^{[10]}$ can easily handle matrices corresponding to $50 \times 50$ $Q$ matrix or 2500 entries of $q$ to be estimated. Fortunately, many of the matrix manipulations required to calculate the variance covariance matrix of $\bar{q}$ may be obtained while calculating $\bar{q}$. Further, if $V$ is a diagonal matrix, then the inversion of $V$ is trivial, and only matrix multiplications and additions are required for the estimation of the variance covariance matrix for $\bar{q}$.

Alternatively, the estimation problem $P1$ may be solved using any general quadratic programming technique. Algorithms which solve quadratic programming problems have been widely used and numerous alternative methods exist. For example, Carey et al.$^{[6]}$ propose a solution using the Frank-Wolfe algorithm. Another alternative is to formulate and solve the equivalent Lagrangian problem as a set of simultaneous linear equations.

Carey et al.$^{[6]}$ note that computational efficiency can be improved by taking into account the special structure of the constraints. For example, if the only constraints in problem $P1$ are row and column totals, then the problem $P1$ reduces to a transportation problem for which several
specialized algorithms have been developed. The result is reduced computation time and storage requirements. More general problems other than those with only row and column constraints can often be transformed into transportation problems.\textsuperscript{[6]} Similarly, partitioning of matrices may be used to reduce the computational burden of matrix inversion in Equation 8.

3. EXAMPLE APPLICATION TO A TRIP DISTRIBUTION ESTIMATION PROBLEM

To illustrate the quadratic estimation technique, we shall apply it to the problem of updating a matrix (or table) of the U.S. distribution of trip purpose by trip length for 1970 to 1977, given the percentages of trips made for each trip purpose and the percentages of trips in each class of trip lengths as the row and column totals, respectively. Data for this example are drawn from the 1970 and 1977 Nationwide Personal Transportation Studies.\textsuperscript{[21, 22]}

Consider a matrix $Y$, which represents the distribution of trip purpose by trip length in 1970 for automobile trips as shown in Table I. We wish to estimate a similar distribution for 1977 using 1970 as the base matrix and some preliminary information which is available for 1977 and given

\begin{table}[h]
\centering
\begin{tabular}{lccc}
\hline
\textbf{Purpose and Data} & \textbf{Trip Length} \\
 & \textless 5 miles & 6-20 miles & \textgreater 20 miles \\
\hline
1970 Measured Proportions (Base Year)\textsuperscript{a} & & & \\
Work & 18.8 & 13.6 & 3.9 \\
Family related & 23.0 & 6.9 & 1.1 \\
Other & 20.6 & 8.8 & 3.3 \\
1977 Estimated Proportions (Unweighted Objective Function, $V = I$) & & & \\
Work & 17.0 & 13.1 & 3.1 \\
Family related & 22.5 & 7.6 & 1.6 \\
Other & 20.7 & 10.1 & 4.3 \\
1977 Estimated Proportions (Chi-square Weighted Objective Function) & & & \\
Work & 16.1 & 13.2 & 3.9 \\
Family related & 22.9 & 7.6 & 1.2 \\
Other & 21.2 & 10.0 & 3.9 \\
1977 Measured Proportions\textsuperscript{b} & & & \\
Work & 16.4 & 13.3 & 3.5 \\
Family related & 22.9 & 7.3 & 1.5 \\
Other & 20.9 & 10.2 & 4.0 \\
\hline
\end{tabular}
\caption{Measured and Estimated Proportions of Automobile Trips by Purpose and Length for 1970 and 1977}
\end{table}

\textsuperscript{a} Source: Tables A18 and A19.\textsuperscript{[21]}
\textsuperscript{b} Source: Tables A5 and A7.\textsuperscript{[22]}

Copyright © 2001 All Rights Reserved
in Table II. The preliminary information is the row and column totals which are the percentage of trips made for each purpose and the percentage of trips of each length. If we arrange \( Y \), given in Table I, to form \( y \) then

\[
y' = [18.8, 23.0, 20.6, 13.6, 6.9, 8.8, 3.9, 1.1, 3.3].
\]

In the constrained least squares regression problem, the vector \( y \) is the vector of observations of the dependent variables. The vector \( \mathbf{q} \) is the corresponding 9 element vector of the entries to be estimated. In the constrained least squares framework, these are the coefficients to be estimated. The constraint set is derived from Table II and represents row and column totals (less one since one total is linearly dependent on the other constraints):

\[
R = \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\quad \text{and} \quad r = \begin{bmatrix}
60.2 \\
30.8 \\
9.0 \\
33.2 \\
31.7 \\
\end{bmatrix}.
\]

The first row of \( R \) is the column constrained associated with the column for trips less than 5 miles and the first entry in \( r \) is the column total given in Table II. If we assume \( V \) is a 9 \( \times \) 9 identity matrix, then the unweighted quadratic programming problem is formulated as

\[
\text{Minimize } (y - \mathbf{q})'(y - \mathbf{q})
\]

subject to \( R\mathbf{q} = \mathbf{r} \)

and the equivalent constrained ordinary least squares problem is to find the least squares estimate of \( \mathbf{q} \) such that

\[
y = I\mathbf{q} + \epsilon
\]

subject to \( R\mathbf{q} = \mathbf{r} \)

\[\]

**TABLE II**

*Distribution of Trip Purpose and Trip Length as a Percentage of All Trips for 1977*

<table>
<thead>
<tr>
<th>Trip Length</th>
<th>6-20 miles</th>
<th>&gt;21 miles</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5 miles</td>
<td>60.2</td>
<td>30.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trip Purpose</th>
<th>Family Related</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>33.2</td>
<td>31.7</td>
</tr>
<tr>
<td>Family Related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>35.1</td>
<td></td>
</tr>
</tbody>
</table>

\[\]

*a Source: Tables A5 and A7.[22]\]
where $I$ is a $9 \times 9$ identify matrix analogous to the matrix of observations of independent variables in the constrained generalized least squares formulation,

$$E[\epsilon] = 0, \text{ and}$$
$$\text{Var}(\epsilon) = \sigma^2 I.$$

Alternatively, if $V$ is a matrix with diagonal elements corresponding to $y$ and off-diagonal elements equal to zero then the formulation is chi-square formulation as described in Section 1. The chi-square formulation is algebraically expressed as follows:

$$y = Iq + \epsilon$$
subject to $Rq = r$

where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 V$. This formulation is somewhat more appealing since we expect that the variations will be larger for entries with large absolute values.

Table I shows the resulting estimates for the unweighted formulation. When compared with the measured values also given in Table I, the average absolute error and mean square error between the estimates and the actual values are 0.3 and 0.11, respectively. If the base matrix for 1970 is compared with the measured values for 1977, the average absolute error and mean square error are 0.6 and 0.99 respectively.

If we assume that the problem has a chi-squared objective function then the matrix entry estimates are as shown in Table I. In this case, the average absolute error and mean square error are 0.2 and 0.05, respectively, representing a significant reduction of approximately 30% and 50% in the average absolute and mean square errors respectively, when compared with the errors in the estimates obtained using unweighted quadratic programming.

The estimated covariance matrix for the chi-square objective function is shown in Table III. If we assume that the error term $\epsilon$ is multivariate

<table>
<thead>
<tr>
<th>TABLE III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance Matrix for the Estimates Using the Chi-squared Objective Function</td>
</tr>
<tr>
<td>0.87 0.38 -0.49 -0.68 0.33 0.36 -0.19 0.05 0.13</td>
</tr>
<tr>
<td>0.65 -0.31 0.33 -0.59 0.26 0.04 -0.11 0.05 0.05</td>
</tr>
<tr>
<td>0.80 0.35 0.26 -0.61 0.14 0.05 -0.20</td>
</tr>
<tr>
<td>0.79 -0.36 -0.44 -0.11 0.03 0.08</td>
</tr>
<tr>
<td>0.63 -0.28 0.03 -0.04 0.01</td>
</tr>
<tr>
<td>0.70 0.08 0.02 -0.09</td>
</tr>
<tr>
<td>0.29 -0.08 0.22</td>
</tr>
<tr>
<td>0.14 0.07</td>
</tr>
<tr>
<td>0.28</td>
</tr>
</tbody>
</table>
normally distributed,

\[ \epsilon \sim MVN(0, \sigma^2 V) \]

then t-statistics and confidence intervals may be calculated for each matrix entry estimate using the estimated variance covariance matrix given in Table III.

For example, the variance of the estimate of \( q_1 \), which is the proportion of work trips of less than 5 miles, is the diagonal element \((1,1)\) in the covariance matrix, or 0.87 in Table III. Based on the assumption of normality of the error terms the \( t \)-statistic for \( q_1 \) is \( q_1 / \{ \text{var}(q_1) \}^{1/2} \) or \( 16.1 / \{ 0.87 \}^{1/2} = 17.26 \), and the 95% confidence interval for this estimate is \([13.3, 18.8]\). The measured value of \( q_1 \) equals 16.4, which is within this confidence interval.

4. COMPARISON OF QUADRATIC ESTIMATION RESULTS WITH OTHER ESTIMATES

The results obtained in Section 3 may be compared with estimates from the RAS or biproportional method as well as entropy or information optimization techniques. Using the notation introduced earlier, the biproportional estimation problem is formulated as\(^{[2]}\):

\[
P2: \quad q_{ij} = b_i y_{ij} c_j \\
\text{subject to } \sum_j q_{ij} = u_i, \\
\sum_i q_{ij} = v_j
\]

where \( u_i \) and \( v_j \) are the row and column totals, respectively and the factors are \( b \) and \( c \) are found relatively easily by an iterative solution algorithm. The matrix entry estimates \( q_{ij} \) may be shown to be equivalents to the following optimization problem\(^{[2]}\):

\[
P3: \quad \text{Minimize } \sum_i \sum_j q_{ij} \ln(q_{ij}/y_{ij}) \\
\text{subject to } \sum_j q_{ij} = u_j \\
\text{and } \quad \sum_i q_{ij} = v_j.
\]

The solution of the example problem using the RAS formulation is shown in Table IV along with the estimates using the chi-square quadratic function described in the previous section. The estimates obtained from these two methods are quite similar. Indeed, they result in approximately the same average absolute and mean square errors.

These similar empirical results are not surprising since there is a close connection between the two estimation problems when they are written as optimization problems such as \( P1 \) and \( P3 \). Bacharach shows that the objective function of the chi-square estimation problem is a first order approximation to the objective function of the biproportional estimation.
problem. This entropy maximization problem is also similar to the biproportional problem. The entropy maximization problem or information minimization estimation can be formulated as:

Minimize \( \sum_i \sum_j y_{ij} \ln(y_{ij}/q_{ij}) \)

subject to the same constraints as previously. This formulation is equivalent to the RAS form \((P3)\) with the exchange of \(q_{ij}\) for \(y_{ij}\) and vice versa. Since we assume \(y_{ij}\) and \(q_{ij}\) are similar, estimates obtained using this entropy formulation should be similar to those obtained using both the chi-squared and the RAS formulation. Unfortunately, the entropy formulation leads to awkward equations for solution; Uribe et al. [20] suggest a two-stage solution approach to estimate \(q\).

As these methods yield similar numerical results and the biproportional method is computationally easier to apply, why use the quadratic method? First, the biproportional method assumes that matrices behave in a biproportional manner. If they do not, then there is no way of estimating how closely the changes approximate a biproportional change. The constrained generalized least squares or quadratic methods assume that errors exist, and estimates of the uncertainty of the results may be determined using standard statistical techniques. The variances calculated in Section 3 along with some distributional assumptions may be used to determine confidence intervals for the estimates.

Second, the biproportional method can only include row and column totals and constraints on individual elements without adding other dimensions to the problem. The quadratic method can easily include any available information which can be represented by a linear relationship between the entries.

Third, the flexibility in defining the error structure (given by the matrix \(V\)) or equivalently the weights in the objective function permits
the inclusion of intuitive information or normative judgments concerning
the structure of changes or errors in the distribution function
\( f(X, \alpha) \).

Fourth, the biproportional method assumes that the base matrix and
constraint totals are known with certainty. This is rarely, the case, but
the real situation is ignored in the biproportional method. The errors can
be accounted for in the quadratic formulation provided the estimates of
the base matrix and constraint total are unbiased.\textsuperscript{[14]}

Finally, the initial elements required for the quadratic formulation
may be of many different kinds. Like the RAS method, a matrix from a
previous study or a distribution function may be used to obtain initial
estimates. The quadratic formulation can also use a distribution function
with unknown parameters to obtain the solution.\textsuperscript{[6]}

5. CONCLUSION

We have shown that the use of a quadratic programming formulation to
obtain matrix entry values yields estimates that are comparable to
estimates found using the biproportional method and consistent with all
available information. An equivalent formulation of the quadratic pro-
gramming problem as a constrained generalized least squares estimation
problem allows estimates of the variance-covariance matrix to be ob-
tained. Furthermore, by making some distributional assumptions about
the error terms in the model, confidence intervals for the matrix entry
estimates may also be estimated. The method is computationally feasible
using either matrix manipulation or quadratic programming computer
software. Compared with methods currently used, the quadratic formu-
lation has several advantages. In particular, the quadratic method is
flexible, as any available information may be included as constraints,
and it allows estimates of the variance of the matrix entry estimates to
be made.

APPENDIX: NOTATION

This appendix includes notation which is used more than once in this
paper. Capital letters are matrices, and bold face lower case letters are
vectors.

\[ \alpha = \text{vector of parameters in a distribution or regression function.} \]
\[ \beta = \text{vector of coefficients in a regression equation.} \]
\[ \epsilon = \text{vector of error terms in a regression function.} \]
\[ f(X, \alpha) = \text{distribution function with variables } X \text{ and parameters } \alpha. \]
\[ I = \text{identity matrix.} \]
\[ \lambda = \text{vector of Lagrange multipliers.} \]
\( Q = \) matrix to be estimated.
\( \hat{Q} = \) matrix estimate of \( Q \).
\( q = \) matrix to be estimated, \( Q \), rearranged as a vector.
\( \hat{q} = \) vector estimate of \( q \).
\( R = \) matrix representing the incidence of elements of \( q \) in each constraint. All matrix elements are equal to zero or one.
\( r = \) vector of constraint values.
\( V = \) matrix of weights in the quadratic estimation problem objective.
\( X = \) matrix of observed attributes for each element in \( q \).
\( Y = \) base matrix corresponding to \( Q \).
\( y = \) base matrix, \( Y \), rearranged as a vector with columns end to end and entries equal to \( f(X, \alpha) \).

ACKNOWLEDGMENTS

The authors are grateful for research support from the American Association of University Women and U.S. National Science Foundation (Grant CEE-8209755). The comments of an anonymous referee are also appreciated.

REFERENCES