Survey Methodology, December 1987
Vol. 13, No. 2, pp. 183–198
Statistics Canada

A Class of Methods for Using Person Controls in Household Weighting

CHARLES H. ALEXANDER[1]

ABSTRACT

A class of "constrained minimum distance" methods is considered for constraining household weights to be consistent with auxiliary information on the number of persons in various age × race × sex cells. The constrained weights are as close as possible to the initial weights based on the inverse probability of selection. This class of methods includes raking and generalized least square methods, as well as multinomial maximum likelihood, (where the cells of the distribution are household types.) The properties of the methods in the presence of systematic undercoverage of the household types are studied through some simple models for coverage. Comparisons with the principal person method are made and the paper concludes with the observation that it is necessary to know more about the nature of survey undercoverage before deciding on which of the constrained minimum distance or principal person methods is to be preferred in applications.

KEY WORDS: Weighting; Auxiliary information; Raking ratio estimation; Principal person method; Survey coverage.

1. INTRODUCTION

Post-stratification is commonly used to adjust survey weights to take into account independent information about the number of units of certain kinds in the population. For example, independent estimates of the population in various age × race × sex post-stratification cells may be available from adjusting census counts for known changes in the number of persons since the census. These independent estimates are often referred to as "control counts". Prior to post-stratification, each sample person (or household) has an initial weight, typically corresponding to the inverse of the selection probability. A post-stratification ratio adjustment factor is applied to the weights of all sample persons in each cell, so that the sum of the adjusted person weights equals the independent control count for the cell. This adjustment is especially important when there is systematic undercoverage of households or persons within households.

For most U.S. Census Bureau demographic surveys, post-stratification is used in assigning weights to sample persons, but is not used directly in assigning weights to sample households. This is due to the greater difficulty of obtaining independent estimates for households. Instead, household weights for these surveys are assigned using some version of the "principal person" method. In the basic principal person method, the household weight is set equal to the final post-stratified person weight of the "principal" person in the household. The rule for identifying this person will be described in Section 2. By using the post-stratified person weight, the principal person method does incorporate the independent estimates of persons into the weights assigned to households.

The most obvious problem with the principal person method is that when the resulting household weights are used to calculate weighted estimates of the number of persons in each post-stratification cell, with each person being given his or her household's weight, these

[1] Charles H. Alexander, Statistical Methods Division, U.S. Census Bureau, Washington, D.C. 20233 U.S.A.

estimates do not agree with the control counts used in the post-stratification. Consequently, there has been interest in methods of assigning weights to households which are constrained to produce person estimates which agree with the independent control counts.

This paper considers a class of methods for assigning survey weights to households, constrained to be consistent with the "known" control counts in various person cells. The general idea is to find household weights which satisfy the constraints and are as close as possible to the initial vector of weights assigned to the households. The different methods within the class correspond to different ways of measuring the distance between the initial vector of weights and the adjusted vector of weights.

Section 2 describes six "constrained minimum distance" weighting methods of this type plus a version of the principal person method. Three of the six methods have been investigated previously, and the others are added in this paper to round out the picture. Section 3 describes the computation of the weights. Section 4 discusses how the adjusted weight depends on the composition of the household. Section 5 discusses results and examples which may help in understanding what these methods do. Section 6 describes areas for further research.

This work has numerous antecedents. The general class of constrained minimum distance methods is suggested for household weighting by Luery (1986). Extending Luery's work, Zieschang (1986a) proposes using one of these methods, generalized least squares, for weighting the U.S. Consumer Expenditure Surveys. Another member of the class is the "minimum discriminant information method", otherwise known as raking ratio estimation or, simply, raking. Oh and Scheuren (1978a) specifically discuss the raking approach to the household weighting problem, and give additional references to a rich literature on raking and related methods. The idea of viewing raking as a constrained minimum distance problem dates back at least to Deming and Stephan (1940). The fundamental principles of this approach are explored in Ireland and Kullback (1968). Applications to survey weight adjustment are well covered in Brackstone and Rao (1979). The class of methods also includes two criterion functions related to multinomial maximum likelihood. The relationship of this to raking has been extensively studied; see, for example, Bishop, Fienberg, and Holland (1976). Fienberg (1986) points out that the distance criteria considered in this paper may be viewed as special cases of a parametric family of functions considered in Cressie and Read (1984).

2. CONSTRAINED MINIMUM DISTANCE METHODS

2.1 Methods Based on Household Weights

Consider a sample of K households, whose initial weights are given by the vector $\underline{S} = (S_1 \ldots, S_K)'$. In this paper, S_k will be the inverse of the probability of selection of the k-th household; in some applications other adjustments such as nonresponse factors may be included in the initial weight.

Suppose that there are J post-stratification cells, and that the number of persons in the population (N_j) is known for each cell. For example, for the U.S. Consumer Expenditure Survey, there are $J = 48$ cells corresponding to combinations of the two sexes, two races (black, nonblack), and twelve age categories. In that survey, persons younger than 14 are not included. The control counts for these cells will be treated as a vector $\underline{N} = (N_1, \ldots, N_J)'$.

The composition of the sample households will be described by a matrix $A = (a_{kj})$, where a_{kj} is equal to the number of persons in the k-th sample household who are in the j-th post-stratification cell. Summing over the post-stratification cells for the k-th household gives $a_{k.}$, the total number of persons in the k-th household. For household k, the vector

$(a_{k1}, \ldots, a_{kJ})$ describes the composition of the household. For example, if the vector is $(2,1,0,0, \ldots, 0)$, then the household contains exactly two persons in the first cell and one in the second.

Using the initial weights $\underline{S}$, the weighted sample estimate of the number of persons in cell j would be $\hat{N}_j = \Sigma_k a_{kj} S_k$ or in general $\underline{\hat{N}} = A'\underline{S}$.

Typically $\underline{\hat{N}} \neq \underline{N}$, i.e., the initial weighted estimate of persons in the post-stratification cells may not equal the known population of the cell.

The goal is to define a new vector of weights $\underline{W} = (W_1, \ldots, W_K)'$ for the sample households, so that $\underline{N} = A'\underline{W}$ or

$$\sum_k a_{kj} W_k = N_j \text{ for } j = 1, \ldots, J. \tag{1}$$

The solution to (1) is not necessarily unique. The idea of the constrained minimum distance methods is to chose $\underline{W}$ so as to minimize some measure $D(\underline{W},\underline{S})$ of the distance between the vectors $\underline{W}$ and $\underline{S}$, subject to (1). In this way, the initial weights $\underline{S}$ are changed as little as possible in meeting the constraint that the adjusted weights should agree with the known control totals. Note that, for certain possible values $N_1, \ldots, N_J$, it may be impossible for any vector of weights $\underline{W}$ to satisfy the constraints (1). Practically speaking, this possible infeasibility does not seem to be a problem, provided the sample is large enough to include a good representation of different types of households, since the controls $\underline{N}$ are generated from the actual population and therefore can be expected to be "feasible".

There are numerous ways of measuring the difference between two vectors. Three distance criteria $D(\underline{W},\underline{S})$ will be considered, corresponding to a household-level generalized least squares (GLS-H) objective function, a minimum discriminant information (MDI-H) function, and a maximum likelihood estimation (MLE-H) criterion. The criteria are:

$$\text{GLS - H:} \quad \sum_k (W_k - S_k)^2/S_k, \tag{2a}$$

$$\text{MDI - H:} \quad (S_. - W_.) + \sum_k W_k \ln(W_k/S_k), \tag{2b}$$

$$\text{MLE - H:} \quad (W_. - S_.) - \sum_k S_k \ln(W_k/S_k). \tag{2c}$$

Throughout the paper, the dot notation is used to denote summation over a subscript.

In each case $D(\underline{W},\underline{S})$ is nonnegative and is equal to zero if and only if $\underline{W} = \underline{S}$. This can be shown, in the usual way, by examining the first and second partial derivatives of each expression with respect to the W_k.

Algorithms for calculating $\underline{W}$ to minimize these three criteria, while meeting the constraint (1) to the degree of approximation desired, will be discussed in Section 3.

2.2 Methods Derived from Person Weights

An alternative approach to this problem leads to a slight but important modification of the three distance criteria. These modified criteria are given by (5a), (5b), and (5c) below. Although these criteria lead to weights for households, they are generated by an approach which starts out by trying to define weights for persons. Accordingly, first consider the problem as one of defining person weights as close as possible to their original household weights, subject to the constraint that the weighted estimate of persons in each post-stratification cell equals the known control. Let the persons in the k-th household be numbered $i = 1, \ldots, a_{k.}$ and let S_{ki} be the initial weight of the i-th person in the k-th household; note that $S_{ki} = S_k$.

Let b_{kij} be a zero-one indicator variable showing whether the i-th person in the k-th household is in the j-th post-stratification cell. Then the condition for consistency with the controls is

$$\sum_k \sum_i b_{kij} W_{ki} = N_j. \tag{3}$$

The three criteria for the person weighting problem would be

$$\sum_k \sum_i (W_{ki} - S_{ki})^2 / S_{ki}, \tag{4a}$$

$$S_{..} - W_{..} + \sum_k \sum_i W_{ki} \ln(W_{ki}/S_{ki}), \tag{4b}$$

$$W_{..} - S_{..} - \sum_k \sum_i S_{ki} \ln(W_{ki}/S_{ki}). \tag{4c}$$

These criteria could be used for defining person weights. In fact the criterion (4c) would lead to the post-stratification weights which are used in person weighting for the Consumer Expenditure Survey, as described in Alexander (1986). However, our problem is to define weights for households. Household weights may be obtained from these criterion functions by imposing upon the person problem the additional constraint that all persons in the same household must have the same weight. Therefore, let $W_{ki} = W_k$ for $i = 1, \ldots, a_k$. Under this constraint, (3) becomes

$$N_j = \sum_k \left(\sum_i b_{kij} \right) W_{ki} = \sum_k a_{kj} W_k,$$

which is the same as the constraint (1) in Section 2.1. The distance criteria (4a), (4b), and (4c) now become:

$$\text{GLS-P:} \quad \sum_k a_{k.} (W_k - S_k)^2 / S_k, \tag{5a}$$

$$\text{MDI-P:} \quad \sum_k a_{k.} S_k - \sum_k a_{k.} W_k + \sum_k a_{k.} W_k \ln (W_k / S_k), \tag{5b}$$

$$\text{MLE-P:} \quad \sum_k a_{k.} W_k - \sum_k a_{k.} S_k - \sum_k a_{k.} S_k \ln (W_k / S_k). \tag{5c}$$

The criteria are now summations at the household level, but the household size $a_{k.}$ has been brought into the criterion for measuring the distance between the initial and adjusted vector of weights. These criteria will be seen to have advantages over the more direct approach which led to (2a), (2b), and (2c).

2.3 The Principal Person Method

In the basic principal person method, the post-stratified person weight of the household's "principal person" is used as the household's weight. To determine the principal person, it is first necessary to determine the household's "reference person". The reference person is identified by the interviewer as the first person mentioned in response to the instruction "start by giving me the name of someone who owns or rents this house." Household relationships are defined in terms of the other members' relationship to this reference person. "Reference person" has replaced the "head of household" concept for this purpose.

The principal person is the wife of the reference person if the reference person is a married male with spouse present. Otherwise, the principal person is the reference person himself or herself. The rationale for this choice is that the principal person should be a person who is not likely to be missed due to within-household undercoverage. In general, women have better coverage than men. Further, the principal owners or renters of the house or apartment seem unlikely to be overlooked.

The basic idea of the principal person method is that there is exactly one principal person in each household. Consequently, the number of households may be estimated by estimating the number of principal persons. This basic method is used for the U.S. National Crime Survey. Other surveys such as the U.S. Consumer Expenditure Surveys or Current Population Survey, make additional adjustments based on assumptions about within-household undercoverage of principal persons, as compared to other persons in the same post-stratification cell (Alexander 1986.)

The principal person method is difficult to model theoretically because the designation of the reference person is somewhat arbitrary. In the hypothetical examples of Section 5, a simplified version of the principal person method will be used, in which the principal person is the household member whose post-stratification cell has the best coverage, i.e., whose post-stratification factor is closest to one. A similar idea is used in Scheuren (1981).

This simplified principal person method will be represented symbolically as follows. For the k-th sample household, let $j(k)$ be the post-stratification cell of the household's principal person. Then the household's principal person weight is

$$W_k = S_k(N_{j(k)}/\hat{N}_{j(k)}).$$

3. COMPUTATION OF THE WEIGHTS

The two least squares methods, GLS-H and GLS-P, have closed-form expressions for $\underset{\sim}{W}$, providing that there exists some solution to the constraints (1). For the GLS-H weights, the adjusted weights are given by

$$\underset{\sim}{W} = \underset{\sim}{S} + MA(A'MA)^{-1}\ (\underset{\sim}{N} - A'\underset{\sim}{S}) \tag{6}$$

where $\underset{\sim}{S} = (S_1, \ldots, S_K)$, $\underset{\sim}{N} = (N_1, \ldots, N_J)$, A is the matrix (a_{kj}) and M is the $K \times K$ diagonal matrix with the elements of $\underset{\sim}{S}$ on the main diagonal. The weights $\underset{\sim}{W}$ for the GLS-P method are also given by (6), except that M is the $K \times K$ diagonal matrix with the values $S_1/a_{1.}, \ldots, S_K/a_{K.}$ on the main diagonal.

A disadvantage of (6) for either method GLS-H or GLS-P is that the solution $\underset{\sim}{W}$ may include negative weights. Conceptually this is unsettling, and for practical users negative weights are unacceptable. It is usually possible to incorporate additional constraints that the weights must be positive. Ways of doing this are given by Zieschang (1986a) and Huang and Fuller (1978). However, the advantage of a simple closed-form solution is lost with these additional constraints.

The raking method (MDI-P) has been used before for household weighting, e.g., by Oh and Scheuren (1978a). A related method which has been extensively tested is described in Pugh, Tyler, and George (1976), based on the approach of Stephan (1942). Luery (1986) gives an iterative algorithm based on Darroch and Ratcliff (1972), which is proved to converge whenever there is a solution to (1). This method is presented here, since the iterative step has a simple interpretation. The iteration starts with "step 0" weights

$$W_k(0) = S_k(N_. / \hat{N}_.)$$

In other words, the initial weight S_k is adjusted by an overall inflation factor equal to the known population $N_.$ divided by the initial weighted total population. At subsequent iterative steps, the adjustment is

$$W_k(i) = W_k(i-1) \prod_j \left(N_j / \sum_s a_{sj} W_s(i-1) \right)^{a_{kj}/a_{k.}}$$

Note that $W_k(i-1)$ is multiplied by the geometric mean of the post-stratification factors for the persons in the k-th household, where the post-stratification factors are calculated using the weights after iteration $i - 1$.

The other three methods, MDI-H, MLE-H, and MLE-P, have not been extensively studied. The following iterative algorithms have worked successfully in small hypothetical examples such as those given in Section 5. In each case, a system of equations, which the weights must satisfy in order to minimize the distance criterion subject to the constraints, can be found by the use of Lagrange multipliers. The equations cannot be solved directly, but if an iterative method produces solutions of the proper form, then the solution minimizes the criterion. If the algorithms converge, the solutions will satisfy the equations. However, the author has no general proof of convergence. A possible alternative approach for the "maximum likelihood" criteria would be to apply the approach of Haber and Brown (1986). Other related work is Fagan and Greenberg (1985).

3.1 Method for MDI-H

The equation for the weights is

$$W_k = S_k \prod_j \gamma_j a_{kj} \tag{7}$$

subject to (1). If values $\gamma_1, \ldots, \gamma_J$ can be found so that the weights calculated according to (7) satisfy (1), then those weights minimize (2b) subject to (1). An iterative algorithm for generating such a vector $\underline{W}$ is as follows.

Initialize $W_k(0) = S_k$ and $\gamma_j(0) = 1$. Then at the i-th iteration let

$$\gamma_j(i) = \gamma_j(i-1) \left[1 - (\hat{N}_j(i-1) - N_j) / \sum_s a_{sj}^2 W_s(i-1) \right],$$

where $\hat{N}_j(i-1) = \sum_s a_{kj} W_s(i-1)$. Then let $W_k(i) = S_k \prod_i (\gamma_j(i))^{a_{kj}}$.

3.2 Method for MLE-H

The solution is of the form:

$$W_k = S_k / \left(1 + \sum_j \gamma_j \, a_{kj}\right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \qquad \text{and} \qquad \gamma_j(0) = 0,$$

$$\gamma_j(i) = \gamma_j(i-1) + (\hat{N}_j(i-1) - N_j) / \left(\sum_s (a_{sj} \, W_s(i-1))^2 / S_k\right),$$

$$W_k(i) = S_k / \left(1 + \sum_j \gamma_j(i) \, a_{kj}\right).$$

3.3 Method for MLE-P

The solution is of the form:

$$W_k = S_k / \left(\sum_j \gamma_{kj} \, a_{kj} / a_{k.}\right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \qquad \text{and} \qquad \gamma_j(0) = 1,$$

$$\gamma_j(i) = \gamma_j(i-1) \, \hat{N}_j(i-1) / N_j,$$

$$W_k(i) = S_k / \left(\sum_j \gamma_j(i) \, a_{kj} / a_{k.}\right).$$

4. THE ROLE OF A HOUSEHOLD'S "COMPOSITION TYPE"

For the six constrained minimum distance methods, the ratio of a household's initial weight to its adjusted weight depends on the number of people in the household in the different post-stratification cells. To discuss this further, the notion of a household's "composition type" will be introduced. Two sample households, say k and m will be said to "have the same type" if they have exactly the same number of people in each of the post-stratification cells, i.e., if

$$a_{kj} = a_{mj} \text{ for } j = 1, \ldots, J. \tag{8}$$

As an example, one household type would be a "household consisting of a white male 35-39 and a white female 30-34." Note that the composition type does not depend on family relationships.

The ratio of the adjusted weight to the initial weight, W_k / S_k, is the same for all households with the same type. In other words, if k and m satisfy (8), then $W_k / S_k = W_m / S_m$. This fact was used in Ireland and Scheuren (1975). A formal proof is given in Alexander and Roebuck (1986).

A useful consequence of this fact is that, in calculating the weights for the constrained minimum distance methods, the calculations may be done using the household type as the unit of analysis rather than the individual household. A simple example may make the implications of these results clearer. Suppose that there are two post-stratification cells, $j = 1$ for females and $j = 2$ for males. The sample consists of K households. For household k, the vector (a_{k1}, a_{k2}) describes how many females and males are in the household; a household with vector (2,1) has two females and one male.

Practically speaking, there is some upper limit on the size of a household, and there are only finitely many household types. For the example, assume that no household has more than three people. Then there are $T = 9$ household types corresponding to the vectors: (1,0), (0,1), (2,0), (1,1), (0,2), (2,1), (1,2), (3,0), (0,3). These types will be numbered consecutively $t = 1, \ldots, 9$. The types will also be labelled mnemonically, F, M, FF, FM, MM, FFM, FMM, FFF, MMM. Hypothetical sample data and control totals are given in Table 1. Note that S_t is the total initial weight given to households of type t.

The constrained minimum distance adjustments effectively may be calculated from the total weights for the household composition types, $S_1, \ldots, S_9$, without actually looking at the individual household weights. Adjusted weights $W_1, \ldots, W_9$ may be calculated using the algorithms from Section 3 replacing summation over k by summation over t. Then for any type t household, the adjusted weight given by the method is W_t / S_t times the initial weight for the household. (The potentially confusing notation of using S_k for the household weight and S_t for the total weight for a t household type is adopted to emphasize that the formulas of Sections 2 and 3 apply equally well to households or household types. In doing calculations, the meaning will be clear from the context.)

The reduction of the problem from individual households to household types is extremely convenient for presenting small examples. Even when applied to the full 48 post-stratification cells, the household-type approach may still be practical: despite the astronomical number of possible household types, the actual number of types in the sample can never be larger than the sample size and often is substantially smaller. This was found to be the case for related cells of households in Ireland and Scheuren (1975). Simply reducing the size of the computational task by combining the weights for single-person households of the same type may be useful; this has been done at the U.S. Bureau of Labor Statistics in applying the generalized least squares method to the Consumer Expenditure Surveys.

The simplified version of the principal person method also depends only on the household type. If two households have the same composition, then their principal persons will be in the same post-stratification cell, the one with the post-stratification factor closest to one. Consequently, the same ratio adjustment factor would be used for both households. In the actual principal person method, the principal person depends in part on who happens to be designated as reference person, so the adjustment factor is not completely determined by the household's composition type.

Note that the MLE-H method corresponds to calculating multinomial maximum likelihood estimates (subject to the constraint (1)) of p_t, $t = 1, \ldots, T$, where p_t is the population proportion of households with type t. The MLE-P method has a related interpretation. Neither of these models, which also pertain to the corresponding GLS and MDI methods, allows for systematic undercoverage.

5. DISCUSSION OF THE METHODS

This Section begins with some speculations about properties of the constrained minimum distance methods, based on the results of Section 4, and follows with some simple hypothetical examples, which generally appear to support the speculations.

The first conjecture is that MLE-H, GLS-H, and MDI-H will tend to give similar results, and also that MLE-P, GLS-P, and MDI-P will tend to be similar to one another, at least for large samples. This is based on the observation that these are all best asymptotic normal estimators under the relevant multinomial sampling model, where the cells are the household types. For small or moderate sample sizes, greater differences between the methods might be anticipated, especially if there are a large number of household composition types, so that the sample in individual "cells" of the multinomial may be small.

The examples given below tend to support this conjecture; the "household" methods all give very similar results, as do the "person" methods. This is true even in some cases when the hypothetical data do not fit the model very well. However, these examples involve only a small number of household types and post-stratification cells, and so are illustrative rather than conclusive.

The second conjecture is based on considering the nature of the sampling models under which the constrained minimum distance methods may be viewed as maximum likelihood estimates, or asymptotic approximations thereto. In these models, perfect coverage is assumed. The models assume a distribution corresponding to probabilities which are the actual proportions in the population, and these probabilities are consistent with the "true" control totals used in the constraints (1). According to these models, for sufficiently large samples, the initial sample estimates would approach agreement with the control totals. This would not be true when there is substantial undercoverage in the sampling frame. Such undercoverage is an important reason for using post-stratification. Coverage considerations may be especially important for telephone surveys where there is no supplemental frame to include households without telephones. If there is no special adjustment for noninterview "nonresponse", such as refusal or inability to provide the requested information, then nonresponse may be a further departure.

Based on these remarks, the second conjecture is that without adjustment the constrained minimum distance methods may not perform well in adjusting for systematic undercoverage, even for large samples. The methods are optimal under models which assume perfect coverage; one would expect that they might be less than optimal when this assumption is violated.

The examples given below partly support this conjecture. The constrained distance methods do not do as well as the simplified principal person method under certain assumptions about undercoverage. Under other assumptions, some of the methods may do quite well. The author concludes that it is necessary to know more about the nature of survey undercoverage before judging that any of these methods is superior to the principal person method. Oh and Scheuren (1978b) raise some related issues about mean square error of the raking estimator when there is undercoverage.

Two examples will be presented, representing two extreme forms of undercoverage. The first ("household undercoverage example") will assume that there is a uniform 10% undercoverage of all households, but that there is no within-household undercoverage. The second example ("within-household undercoverage example") assumes a 10% undercoverage of males due to within-household undercoverage in households where there are both males and females, and undercoverage of all-male households. For single-person households, any "within-household undercoverage" means that the whole household is missed.

In example 1, there is a 10% under-representation of all types of households in the sample. For a sufficiently large sample, this would obviously be due to systematic undercoverage, rather than sampling error. Applying the constrained minimum distance methods and the principal person method to this example gives the total adjusted weights for each household type shown in the last four columns of Table 1.

Note that the GLS-P, MDI-P, and MLE-P methods all bring the adjusted weight up to the actual population value. Thus, these methods give "unbiased" weights. Since all persons have a second-stage factor of 1/.9, the principal person method also achieves this result.

Table 1
Household Undercoverage Example:
Description of Population and Sample

			Total Weight (W_t) for Methods:			
Type & description	Actual Population	Total Initial Weights	GLS-H	MDI-H	MLE-H	GLS-P MDI-P MLE-P Prin. Pers.
1: F	25,000	22,500	23,785	23,745	23,704	25,000
2: M	15,000	13,500	14,120	14,097	14,075	15,000
3: FF	7,000	6,300	7,020	7,016	7,013	7,000
4: FM	40,000	36,000	39,708	39,672	39,632	40,000
5: MM	5,000	4,500	4,913	4,906	4,900	5,000
6: FFM	12,000	10,800	12,529	12,506	12,594	12,000
7: FMM	12,000	10,800	12,408	12,428	12,449	12,000
8: FFF	0	0	0	0	0	0
9: MMM	0	0	0	0	0	0
Total	116,000	104,400	114,483	114,370	114,367	116,000

Control Totals:	Number of Females	=	115,000
	Number of Males	=	101,000
Initial Weighted Person Counts:	Females	=	103,500
	Males	=	90,900

The other methods, GLS-H, MDI-H, and MLE-H, all give substantially too little weight to one-person households and too much to the three-person households. Intuitively, this makes sense; since these methods do not allow for systematic undercoverage and must explain the shortage of sample persons as sampling error, the obvious explanation is that the sample has a below-average number of large households, due to chance. The better performance of MLE-P makes some sense, since it starts out with a multinomial sampling model which allows sampling of persons without regard to households.

Practically speaking, this example reflects very poorly on the GLS-H, MDI-H, and MLE-H methods. Even uniform undercoverage would cause these methods to distort the distribution of household sizes. Worse, the distortion goes opposite from what is commonly assumed about differential household coverage, namely that small households are more likely to be missed than large ones, so that small households need relatively higher weights, not relatively lower weights.

The second example will emphasize within-household undercoverage of males. The situation is more complicated than in the previous example, because a household may have an apparent composition type different than its actual type. For example, a household which actually consists of a male and a female may appear to be a single-person household. The actual and apparent type will be indicated by modifying our previous notation. For example, a FM household in which the male is missed will be denoted F[M]. A [M] household or [MM] household is missed entirely. Table 2 describes the hypothetical data. The actual population is the same as in the previous example.

Table 2
Within-household Undercoverage Example:
Description of Population and Sample

Actual Household Type	Apparent Type	Actual Number	Total Initial Weights
1: F	F	25,000	25,000
2: M	M	13,500	13,500
	[M]	1,500	0
3: FF	FF	7,000	7,000
4: FM	FM	36,000	36,000
	F[M]	4,000	4,000
5: MM	MM	4,500	4,500
	[MM]	500	0
6: FMM	FFM	10,800	10,800
	FF[M]	1,200	1,200
7: FMM	FMM	10,800	10,800
	FM[M]	1,200	1,200
8: FFF	FFF	0	0
9: MMM	MMM	0	0
		116,000	114,000
Control Counts:	Number of Females	115,000	
	Number of Males	101,000	
Initial Weighted Person Counts:	Females	115,000	
	Males	90,900	

Note that there is a 10% undercoverage of males, due to missing males within households, or missing all-male households. Each male has a 10% chance of being missed.

Neither column of numbers in table 2 is observed, since there are no household controls. Also the actual household type is not known for the sample units. Thus, the [FM] households appear to be the same as the F households. The data which would be observed are given in Table 3, along with the total initial weight for households which appear to have a given type. The adjusted weights are given for three methods, MLE-H, MLE-P, and principal person. The results for GLS-H and MDI-H are fairly close to MLE-H, and GLS-P and MDI-P are similar to MLE-P, so these other methods are omitted.

The last three columns of Table 3 show the total adjusted weight assigned to each actual household type by the MLE-H, MLE-P, and principal person methods. The principal person weights for each actual household type agree with the population counts for the actual types, shown in the third column of Table 1. In this sense, the principal person weights are unbiased.

This example corresponds to assumptions upon which the simplified principal person is based. The principal person adjusted weights for each actual type of household coincide with the population counts. The one difference is that totally missing [M] or [MM] households are given no weight; however, the weight of the non-missing M or MM households is increased accordingly. The total weighted number of households for the principal person method is equal to the number in the population.

Table 3

Within-household Undercoverage Example: Observed Types and Weights, with Adjusted Weights from Three Methods

		Weight Assigned to Apparent Type			Weight Assigned to Actual Type		
Household Type	Total Initial Weight	MLE-H	MLE-P	Principal Person	MLE-H	MLE-P	Principal Person
F	29,000	27,450	26,973	29,000	23,664	23,253	25,000
M	13,500	14,997	16,338	15,000	14,997	16,338	15,000
FF	8,200	7,368	7,626	8,200	6,290	6,510	7,000
FM	37,200	38,887	39,128	37,200	41,419	41,586	40,000
MM	4,500	5,623	5,446	5,000	5,623	5,446	5,000
FFM	10,800	10,661	10,885	10,800	11,739	12,001	12,000
FMM	10,800	12,605	11,878	10,800	13,859	13,140	12,000
FFF	0	0	0	0	0	0	0
MMM	0	0	0	0	0	0	0
Total	114,000	117,591	118,274	116,000	117,591	118,274	116,000

In this example, the constrained minimum distance methods overestimate the total number of households, but give too little weight to the households without males. In general, too much weight is given to households with males.

It should not be concluded that the principal person method always outperforms the constrained minimum distance methods when there is within-household undercoverage. Under other assumptions about coverage, the principal person method may not do so well. In fact, different versions of the principal person method are used for different surveys, based on various assumptions about coverage. Note also that combinations of the principal person method and raking methods are possible; see Scheuren (1981).

Even in this example, the biased weights assigned by the constrained minimum distance methods could be beneficial for estimating some characteristics. If the households in which males are missed tend to under-report the variable of interest, then giving these households too high a weight may tend to counteract response bias associated with the within-household undercoverage.

The most extreme example of this effect is estimation of the total number of males, in which case the MLE-H and MLE-P weights give estimates which agree with the control totals while the principal person weights do not. However, for household characteristics where there would rarely be reporting errors because of the missed male, such as form of tenure (renter/owner), the biased weights would not be desirable. The performance of the weighting methods in situations like these clearly depends on the nature of the survey undercoverage, and its relationship to the variable being estimated. This is discussed further, with additional examples, in Alexander and Roebuck (1986).

Pending further research on survey coverage and its effect on weighting, what recommendations can be made? Among the constrained minimum distance methods considered in this paper, GLS-H, MDI-H, and MLE-H seem unattractive because of their failure to adjust correctly for uniform undercoverage of households. This is in spite of the fact that, if there were no undercoverage, MLE-H seems to be based on a more sensible model than MLE-P, since households rather than persons are the ultimate sampling unit.

The possibility of negative weights raises questions about the appropriateness of GLS-P, even though in some practical applications (such as Zieschang 1986b) there are very few negative weights, so that they could be replaced by positive weights with little effect on the estimates. That leaves MDI-P and MLE-P. Our results give little basis for choosing between these methods. Computational considerations tend to favor the "raking" method MDI-P. Based on limited experience with the algorithms of Section 3, the MLE methods converge more slowly than the MDI methods. Further, there has been considerable research into ways to improve the efficiency of raking for large-scale applications, such as Ireland and Scheuren (1975). Taking all this into account, the raking method, MDI-P, seems to be the most promising of the constrained minimum distance methods.

The constrained minimum distance methods give household weights which are consistent with control totals for person, unlike the principal person method. However, the superiority of the constrained minimum difference methods over the principal person method as an adjustment for undercoverage is far from obvious. Undercoverage is an essential part of the survey weighting problem. The principal person method is an ad hoc solution to the undercoverage problem, based on some very simplistic assumptions about coverage. However, as seen in Section 4, the constrained minimum difference methods may be viewed as "optimal" (i.e., maximum likelihood or the asymptotic equivalent) estimators under models which assume perfect coverage. The choice is thus between an optimal solution to the wrong problem and an ad hoc solution to what may or may not be the right problem. Clearly more research is needed.

6. SOME AREAS FOR FURTHER RESEARCH

6.1 Household Control Totals

If independent estimates of the number of households of different kinds were available, then ordinary post-stratification could be used for household estimates. Household controls by size of household are being investigated, based on updating 1980 census results (Das Gupta *et al.* 1986). The availability of household controls would fundamentally change our ability to deal with the household weighting problem.

Even with household controls, it might be beneficial to also incorporate person controls. The household controls are not likely to include detailed information on the age, race, and sex of the household members. The use of raking to simultaneously control the estimates to independent controls for persons and households is developed by Scheuren (1981), using an estimate of the total number of households. Zieschang (1986a) describes how similar adjustments may be made using generalized least squares.

Household controls clearly have great potential for adjusting for differential coverage of various types of households. There still may be problems is dealing with within-household undercoverage, since this may lead to errors in determining the true household size, which would cause sample households to be placed in the wrong post-stratification cell.

6.2 Research Concerning Coverage

Coverage of persons is measured fairly well by comparing the initial survey estimates $\hat{N}_j$ to the control totals N_j. It is difficult to determine how much of this undercoverage is due to missing entire households and how much is due to missed persons within households. Additional information could be obtained by comparing initial weighted household estimates to household controls, once these controls become available. In the meantime, 1980 survey estimates by type of household could be compared to the corresponding 1980 census counts.

Even with this additional information, it is not possible to completely distinguish household undercoverage from within-household undercoverage, without making additional assumptions. Alexander and Roebuck (1986) present some preliminary suggestions about how a range of coverage models might be fit to census and survey data. An alternative approach would be to include coverage parameters in a multinomial sampling model such as those described for the MLE-H or MLE-P weighting methods. Other approaches to modelling coverage are presented in Wolter (1986).

6.3 Estimation of Variances

Methods for estimating variances of the weighted estimators have not been investigated for most of the constrained minimum distance methods. For raking estimators, some methods are available; see Arora and Brackstone (1977), Bankier (1978) and Fan *et al.* (1981).

For any of the methods, replication methods for estimating the variance could be applied. These methods have been shown to give reasonable results under fairly general conditions; see for example Krewski and Rao (1985). It remains to be determined whether these conditions can be applied to the constrained minimum distance methods.

6.4 Computational Issues

Zieschang (1986b) has applied the generalized least squares methods to the U.S. Consumer Expenditure Surveys. Scheuren (1981) describes a large-scale application of the raking method to household weighting. The maximum likelihood constrained minimum distance algorithms (MLE-H and MLE-P) have not been tried on large-scale problems of this kind. If they were to be used in actual survey weighting, research may be needed to improve their computational efficiency.

ACKNOWLEDGMENTS

The author would like to thank Michael J. Roebuck for his assistance with portions of this research, and also the associate editor and referees for their helpful comments. The author is also indebted to Brenda Kelly for her diligence in typing this manuscript.

REFERENCES

ALEXANDER, C.H. (1986). The present Consumer Expenditure Surveys weighting method. In *Population Controls in Weighting Sample Units,* Section 1. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-32.

ALEXANDER, C.H., and ROEBUCK, M.J. (1986). Comparison of alternative methods for household estimation. *Proceedings of the Section on Survey Research, American Statistical Association,* 54-64.

ARORA, H.R., and BRACKSTONE, G.J. (1977). An Investigation of the Properties of Raking Ratio Estimates: II. With cluster sampling. *Survey Methodology,* 4, 232-252.

BANKIER, M.D. (1978). An estimate of the efficiency of raking ratio estimators under simple random sampling. *Survey Methodology,* 4, 115-124.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā,* Series C., 41, 97-114.

BISHOP, Y.M.M., FIENBERG, S.W., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

CRESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society,* Series B, 46, 440-464.

DARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics,* 63, 1470-1480.

DAS GUPTA, P., GIBSON, C., HERRIOT, R.A., LAMAS, E., and ZITTER M. (1986). New approaches to estimating households and their characteristics for states and counties. Paper presented at the 1986 annual meeting of the Population Association of America.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *Annals of Mathematical Statistics,* 11, 427-444.

FAGAN, J.T., and GREENBERG, B. (1985). Algorithms for making tables additive: raking, maximum likelihood, and minimum chi-square. U.S. Bureau of the Census, Statistical Research Division Report Series No. Census/SRD/RR-85/12.

FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., and THOMPSON, J.H. (1981). 1980 census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 176-181.

FIENBERG, S.E. (1986). Comments on some estimation problems in the Consumer Expenditure Surveys. In *Population Controls in Weighting Sample Units.* Section 5. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-12.

HABER, M., and BROWN, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association,* 81, 477-482.

HUANG, E.T., and FULLER, W. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of Social Statistics Section, American Statistical Association,* 300-305.

IRELAND, C.T., and SCHEUREN, F.J. (1975). The rake's progress, *Computer Programs for Contingency Table Analysis.* Washington, D.C.: The George Washington University, 155-216.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* 9, 1010-1019.

LUERY, D.M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association,* 325-330.

OH, H.L., and SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 716-722.

OH, H.L., and SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 723-725.

PUGH, R.E., TYLER, B.S., and GEORGE, S. (1976). Computer-based procedure for N-dimensional adjustment of data - NJUST. U.S. Social Security Administration, Staff Paper No. 24.

SCHEUREN, F.J. (1981). Methods of estimation for the 1973 Exact Match Study. *Studies from Interagency Data Linkages, Report No. 10.,* U.S. Department of Health and Human Services, U.S. Social Security Administration, 9-122.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics,* 13, 166-178.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association,* 81, 338-346.

ZIESCHANG, K.D. (1986a). Generalized least squares: an alternative to principal person weighting. In *Population Controls in Weighting Sample Units,* Section 2. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-41.

ZIESCHANG, K.D. (1986b). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 64-71.