COMPARISON OF ALTERNATIVE METHODS FOR HOUSEHOLD ESTIMATION

Charles H. Alexander and Michael J. Roebuck, U.S. Bureau of the Census

I. INTRODUCTION

This paper deals with a problem of making estimates for a population based on a sample, when there is exogenous information about how many units of certain kinds are present in the population. The problem is how to assign a "survey weight" to each sample household (or other group of persons, such as a "consumer unit") when the number of persons in the population in each of several age/race/sex cells is known.

Let N be the number of households in the population. A sample of K households is selected from the population. In one particular survey of interest, the Consumer Expenditure Survey (CE), the sample is a two-stage sample of addresses, with stratification at the first stage and systematic selection at the second; in other applications a cluster of addresses may be selected at the second stage.

Some households may be unintentionally left out of the sampling frame, for a variety of reasons. Assume that the frame is a fixed set of all the N_f households which potentially can come into the sample, where $N_f \leq N$. Also persons in sample households can be missed.

Assume that a method exists for assigning a vector of "unbiased" survey weights $\underline{S} = (S_1, . . S_K)$ to the sample households.

Definition: The weights $\underline{S}$ will be said to be "unbiased" if and only if for any set B of households in the frame,

$$E \left(\sum_{k \in B} S_k \right) = \#(B),$$

where the summation is over the sample households in B and "#(B)" denotes the number of households (in the frame) in the set B.

"Expectation" is used in the sense of the mean over all possible samples, each weighted by the probability of obtaining that sample. Note that the "unbiased" weights are unbiased only with respect to the frame; they may still lead to biased estimates for the population because of undercoverage by the frame.

Let the persons in the full population be divided into J age/race/sex cells and assume that the number of persons P_j in each cell is known. As an example, for CE there are 48 "post-stratification" cells: 12 age groups, the two sexes, and 2 race groups. Usually only persons eligible for the survey are used in the weighting. For CE only persons age 14 or older are included.

In some applications, there are household cells for which the number of households H_c in each cell is known. For example these cells might correspond to household size (1,2,3,4+) crossed by form of tenure (renter/owner).

The problem is how the unbiased survey weights may be altered to take into account the known "control totals" $\underline{P}$ and $\underline{H}$, giving a new vector of weights $\underline{W}$, so as to reduce the bias and variance of survey estimates calculated using the new weights.

This paper will concentrate on a class of "constrained minimum distance methods" in which $\underline{W}$ is chosen to be as close as possible, in some sense, to the original weights $\underline{S}$, subject to the constraint that the new weights give estimates which are exactly consistent with the control totals. The focus of the paper is how various constrained minimum distance methods compare under different assumptions about the coverage of the population by the frame. To facilitate the comparisons, a method of generating a range of models consistent with potentially available data on coverage is developed. The method is illustrated with hypothetical values for the "potentially available" data.

This paper will deal only with the case of known (P_j), without any household controls (H_c). This problem has immediate interest for CE weighting. Household control counts based on the 1980 census have not been used to date for CE, but are being investigated. (Das Gupta, et al (1986)).

II. CONSTRAINED MINIMUM DISTANCE WEIGHTING METHODS

Let there be K sample units (households or consumer units) with unbiased weights $S_1 . . . , S_K$. Suppose that there are J age/race/sex cells with known cell populations $P_1, . . . , P_J$.

Let $A = (a_{kj})$ be a matrix with a_{kj} equal to the number of persons in the kth sample household who are in the j^{th} post-stratification cell. (Assume $a_{k+} \geq 1$).

Constrained minimum distance methods find weights $W_1, . . . , W_K$ such that $D(\underline{W}, \underline{S})$ is minimized subject to

$$(2.1) \qquad \sum_k a_{kj} W_k = P_j \qquad \text{for } j = 1, . . . J,$$

where $D(\underline{W}, \underline{S})$ is some measure of the difference between $\underline{W} = (W_1, . . . , W_K)$ and $\underline{S} = (S_1, . . . , S_K)$. The known initial weights $\underline{S}$ are treated as fixed constants in computing $\underline{W}$.

Six criterion functions will be considered. Three are based on a summation over persons and three over households. In both cases the methods involve a generalized least squares (GLS) objective function, a minimum discriminant information (MDI) function, and a "maximum likelihood estimation", (MLE) criterion.

The criterion functions are:

$$D_1 (\underline{W}, \underline{S}) = \sum_k (W_k - S_k)^2 / S_k$$

$$D_2 (\underline{W}, \underline{S}) = \sum_k a_{k+} (W_k - S_k)^2 / S_k$$

$$D_3 (\underline{W}, \underline{S}) = S_+ - W_+ + \sum_k W_k \ln (W_k / S_k)$$

$$D_4\ (\underline{W},\underline{S}) = \sum_k a_{k+} S_k - \sum_k a_{k+} W_k + \sum_k a_{k+} W_k \ln (W_k/S_k)$$

$$D_5\ (\underline{W},\underline{S}) = W_+ - S_+ - \sum_k S_k \ln (W_k/S_k)$$

$$D_6\ (\underline{W},\underline{S}) = \sum_k a_{k+} W_k - \sum_k a_{k+} S_k - \sum_k a_{k+} S_k \ln (W_k/S_k)$$

D_1 (HH-level GSL) has been studied in Luery (1980) and Zieschang (1986). Roman (1982) discusses extensions to person and household controls and mentions D_2 (person-level GLS). D_4 (person level MDI) was proposed and tested extensively in Scheuren, et al (1981) for both person and HH controls. This general approach goes back further to Stephan (1942) and Oh and Scheuren (1978). D_3 (HH-level MDI), D_5 (HH-level MLE), and D_6 (person-level MLE) are suggested here as obvious possible alternatives. Fienberg (1986) notes that all of these methods are special cases of a parametric family of criterion functions described in Cressie and Read (1984).

Remarks:

1. **"Person-level" vs "Household-level".** Note that each "household-level" criterion is the summation over all sample households of a term involving the household weights S_k and W_k. Similarly, each person-level criterion could also be rewritten as the summation over all sample persons, with each sample person being assigned his/her household's weight. For example, in the expression for D_2, each of the a_{k+} persons in the kth household contributes a term $(W_k - S_k)^2/S_k$. Summing over all sample persons gives $D_2(\underline{W},\underline{S})$.

2. **A Property of the Criterion Function.** In each case $D_i(\underline{W},\underline{S})$ is non-negative and is equal to zero if and only if $\underline{W} = \underline{S}$. This is easily established by examining the first and second partial derivatives of $D_i(\underline{W},\underline{S})$ with respect to W_k.

3. **Equivalent Expressions.** The MDI criterion used in earlier work has been altered slightly in the present paper to obtain the property in Remark 2, without altering the resulting weights W_k. In particular, under the constraint (2.1), minimizing $D_4\ (\underline{W},\underline{S})$ is equivalent to minimizing

$$D_4^*(\underline{W},\underline{S}) = \sum_k a_{k+} W_k \ln(W_k/S_k).$$

Indeed assuming (2.1) the extra terms are the expression

$\sum_k a_{k+}\ S_k - \sum_k a_{k+} W_k = \sum_k a_{k+} S_k - P_+$ which is constant.

Similarly, minimizing $D_5(\underline{W},\underline{S})$ subject to (2.1) is equivalent to minimizing

$$D_5^*(\underline{W},\underline{S}) = W_+ - \sum_k S_k \ln (W_k/S_k).$$

Equivalent expressions for D_3 and D_6 are

$$D_3^*(\underline{W},\underline{S}) = - W_+ + \sum_k W_k \ln(W_k/S_k)$$

$$D_6^*(\underline{W},\underline{S}) = - \sum_k a_{k_+}\ S_k \ln(W_k/S_k)$$

4. **Present Method.** The household weighting method presently used for many Census Bureau household surveys is some form of the "principal person" method. In the basic principal person method the final weight assigned to the k^{th} sample household is

$W_k = S_k\ (P_{j(k)}\ /\ \sum_i a_{ij(k)}\ S_i)$, where some index $j(k)$ is chosen so that $a_{kj(k)} > 0$. The expression in parentheses is the "post-stratification" or "ratio-adjustment" factor for the j(k) person cell. Thus the initial weight S_k is multiplied by the post-stratification factor for one of the persons in the kth household. A rule is needed for determining which of the cells included in the household to choose as j(k). In this paper a modified principal person method is used, in which j(k) will be chosen to be the "best-covered" cell in the household, i.e., j(k) will be that value of j such that $a_{kj} > 0$ and the ratio

$P_j/ \sum_i a_{ij}\ S_i$ is minimized over all j with $a_{kj} > 0$.

In practice, the choice of the "best-covered" cell would be made based on historical evidence from past surveys, rather than on the present sample itself.

The principal person method as actually used for Census Bureau surveys is more difficult to simulate, because the choice of principal person depends in part on who the interviewer first encounters at the household. However, the general idea is to select a principal person in a group which is well covered by the frame. To this extent, the modified principal person method used in this paper is similar to the actual principal person method used for some surveys. (For other surveys, such as CE, there are additional variations on this basic method. Additionally, some surveys use estimates from the Current Population Survey (CPS) as their household control vector $\underline{H}$.)

III. COMPUTATION OF THE WEIGHTS.

The least squares methods D_1 and D_2 have closed-form expressions for $\underline{W}$, providing the constraints are feasible. The other methods require iterative solutions. Assuming the constraints are feasible, the convergence of the solution to method D_4 can be proved using the results of Darroch and Ratcliff (1972), as is shown in Luery (1980).

The other three methods have unique solutions whose form can be found using Lagrange multipliers, assuming the constraints are feasible. The algorithms given below for D_3, D_5, and D_6 are new. They have converged to solutions of the appropriate form in all examples considered so far, but we do not have a general proof of convergence. Note that as long as expressions for $\underline{W}$ can be found, having the proper form and satisfying the constraints (2.1), these $\underline{W}$ values necessarily minimize the criterion function subject to the constraints. The methods for D_3 and D_5 are suggested by the cyclic coordinate descent method for finding the root of an equation, and the method for D_6 is based on an analogy with the solution for D_4. An alternative algorithm related to D_5 is given in Haber and Brown (1986) for a different problem, for which the method is proved to converge when the constraints are feasible. Other related work is Fagan and Greenberg (1985).

Examples to illustrate the calculations are given in Section VI.

Method for D_1: The solution vector $\underline{W} = (W_k)$ is

$$\underline{W} = \underline{S} + DA(A'DA)^{-1}(\underline{P} - A'\underline{S})$$

where $\underline{S} = (S_k)$, $\underline{P} = (P_j)$, $A = (a_{ij})$, and D is the K x K diagonal matrix with the elements of $\underline{S}$ on the main diagonal. The weights from D_1 and D_2 may be negative. Ways of handling this are discussed in Zieschang (1986) and Huang and Fuller (1978)

Method for D_2: The solution has the same form as for D_1, except that D is the K x K diagonal matrix with (S_k/a_{k+}) on the main diagonal.

Method for D_3: The solution is of the form

$$W_k = S_k \prod_j \gamma_j^{a_{kj}}$$

subject to (2.1).

An iterative algorithm for generating such a vector W is as follows.

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

$$\gamma_j(i) = \gamma_j(i-1)\left[1 - (\hat{P}_j(i-1) - P_j) \Big/ \sum_k a_{kj}^2 W_k(i-1) \right],$$

where $\hat{P}_j(i-1) = \sum_k a_{kj} W_k(i-1)$.

$$W_k(i-1) = S_k \prod_j (\gamma_j(i-1))^{a_{kj}}.$$

Method for D_4: As shown by Luery (1980) based on Darroch and Ratcliff (1972), a convergent algorithm which minimizes D_4 subject to (2.1) is

$$W_k(0) = S_k \, (P_+ / \sum_k a_{k+} S_k)$$

$$W_k(i) = W_k(i-1) \prod_j \left(P_j / \sum_k a_{kj} W_k(i-1)\right)^{a_{kj}/a_{k+}}$$

This has a simple interpretation. Note that $W_k(i-1)$ is multiplied by the geometric mean of the "post-stratification" factors for the persons in the kth household. (In the examples we considered, the algorithm converges if the arithmetic or harmonic mean is used, but is not clear whether the resulting limits correspond to any distance function.)

Method for D_5: The solution is of the form

$$W_k = S_k / (1 + \sum_j \lambda_j a_{kj})$$

subject to (2.1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \lambda_j(0) = 0$$

$$\lambda_j(i) = \lambda_j(i-1) + (\hat{P}_j(i-1) - P_j) \Big/ \left(\sum_k (a_{kj} W_k(i-1))^2 / S_k\right)$$

$$W_k(i) = S_k / (1 + \sum_j \lambda_j(i) \, a_{kj})$$

Method for D_6: The solution is of the form

$$W_k = S_k / (\sum_j \lambda_{kj} a_{kj} / a_{k+})$$

subject to (2.1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \lambda_j(0) = 1$$

$$\lambda_j(i) = \lambda_j(i-1) \, \hat{P}_j(i-1) / P_j$$

$$W_k(i) = S_k / (\sum_j \lambda_j(i) \, a_{kj} / a_{k+})$$

IV. SOME PROPERTIES OF THE METHODS

The constrained minimum distance methods give the same increase to the weights of all households of the same "type", defined by the number of persons in the household in each post-stratification cell.

Definition: A household type is the set of all households in the population with a given vector $(a_{k1},\ldots .a_{kJ})$.

For example, the households with type $(2,0,0,\ldots,0)$ would be those composed of exactly two persons in the first cell. If there are T different types represented in the population, the types will be indexed $t=1,\ldots,T$ and the type will be identified with its index t. Let $a_j(t)$ denote the number of persons in cell j in a type t household. Note that if there is within-household undercoverage, a household's apparent type may differ from its actual type.

Lemma: Suppose two sample households k and l have the same type, i.e.,

$$a_{kj} = a_{lj} \quad \text{for } j=1,\ldots,J. \tag{4.1}$$

Then for any of the methods $D_1,\ldots,D_6$,

$$W_k / S_k = W_l / S_l \tag{4.2}$$

Proof: (For D_1) Suppose that $W_1,\ldots, W_k$ minimize D_1 (W,S) subject to (2.1), but that for two households k and l of the same type,

$$W_k/S_k \neq W_l/S_l.$$

Define a new set of weights

$$V_i = W_i \quad \text{for } i \neq k,\ l$$

$$V_k = S_k[(W_k + W_l)/(S_k + S_l)]$$

$$V_l = S_l[(W_k + W_l)/(S_k + S_l)]$$

Since $V_k + V_l = W_k + W_l$, (4.1) implies that the weights V also satisfy (2.1).

Simple calculus shows that the expression

$$(X - S_k)^2 / S_k + (C - X - S_l)^2 / S_l,$$

considered as a function of X, is uniquely minimized by

$$X = C \ S_k / (S_k + S_l).$$

It follows that (letting $C = W_k + W_l$)

$$(V_k - S_k)^2 / S_k + (V_l - S_l)^2 / S_l < (W_k - S_k)^2 / S_k + (W_l - S_l)^2 / S_l$$

Therefore $D_1(\underline{V},\underline{S}) < D_1(\underline{W},\underline{S})$, which contradicts the assumption that W minimizes $D_1(\underline{W},\underline{S})$, proving the Lemma by contradiction. The proofs for $D_2,\ldots,D_6$ are similar.

Theorem: Let W minimize $D_i(\underline{W},\underline{S})$ subject to (2.1). Let the sample be partitioned into T household types, and let $S_1^*,\ldots,S_T^*$ be the total weight of the units of each of the T types. Let W^* minimize $D_i(\underline{W}^*, \underline{S}^*)$ subject to (2.1).

Then for each household type t,

$$W_t^* = \sum_{k\in t} W_k. \tag{4.3}$$

Proof: (For D_1) For any vector of T weights $\underline{W}^*$, satisfying (2.1), there exists a unique vector of K weights $\underline{W}$ which satisfy (4.2) and (4.3). These weights are

$$W_k = S_k W_t^* / S_t^* \quad \text{for } K \in t. \tag{4.4}$$

It is easy to see that these weights $\underline{W}$ satisfy (2.1) and that $D_1(\underline{W},\underline{S}) = D_1(\underline{W}^*,/\underline{S}^*)$.

Indeed,

$$\begin{aligned}\sum_{k=1}^{K} a_{jk} W_k &= \sum_t \sum_{k\in t} a_{kj} W_k \\ &= \sum_t (W_t^* / S_t^*) \sum_{k\in t} a_{kj} S_k \\ &= \sum_t a_j(t) (W_t^* / S_t^*) \sum_{k\in t} S_t \\ &= \sum_t a_j(t)\, W_t^* = P_j\end{aligned}$$

and

$$\begin{aligned}D_1(\underline{W},\underline{S}) &= \sum_k (W_k - S_k)^2 / S_k \\ &= \sum_t \sum_{k\in t} (S_k W_t^*/S_t^* - S_k)^2 / S_k \\ &= \sum_t \sum_{k\in t} S_k (W_t^* - S_t^*)^2/ (S_t^*)^2 \\ &= \sum_t (W_t^* - S_t^*)^2/ S_t^* = D_1(W^*,S^*)\end{aligned}$$

Conversely, for any weights $\underline{W}$ which satisfy (4.2), the weights $\underline{W}^*$ defined by (4.3) are related to $\underline{W}$ by (4.4). Thus, there is a one-to-one correspondence between weights W which satisfy (4.2) and weights $\underline{W}^*$, such that $D_1(\underline{W},\underline{S}) = D_1(\underline{W}^*,\underline{S}^*)$. By the Lemma, the minimum over all $\underline{W}$ is the same as the minimum over all $\underline{W}$ which satisfy (4.2). It follows that the weights

$\underline{W}$ which minimize $D_1(\underline{W},\underline{S})$ correspond to the weights $\underline{W}^*$ which minimize $D_1(\underline{W}^*, \underline{S}^*)$.

The proofs for $D_2,\ldots,D_6$ are similar.

This result can be used to simplify the computations, since the number of terms is now equal to the number of household types rather than the number of sample households.

The expression $D_5(\underline{W}^*, \underline{S}^*)$ has a specific interpretation in terms of maximum likelihood. Suppose there are T household types in the population. A simple random sample of size n is selected with replacement. Let p_t be the probability that a given unit is of type t, and let x_t be the observed number of type t units in the sample.

Then the log-likelihood function $L(p_1,\ldots,p_T / x_1,\ldots,x_T)$ is (up to a constant)

$$\sum_{t=1}^{T} x_t \ln (p_t/x_t)$$

Let $S^*_t = (N/n)\, x_t$ and $W^*_t = N\hat{p}_t$, where $\hat{p}_t$ is the maximum likelihood estimator. Then maximizing (4.5), subject to $\Sigma\, p_t = 1$ and (2.1), is equivalent to minimizing

$$-\sum_{t=1}^{T} S^*_t \ln (W^*_t/S^*_t)$$

subject to $W^*_+ = S^*_+$ and (2.1). Under these constraints (4.6) is the same as $D_5(\underline{W}^*, \underline{S}^*)$.

Thus, the D_5 weights correspond to maximum likelihood estimation under simple multinomial sampling. Under this model, D_1 and D_3 are asymptotically equivalent to D_5. This model assumes no systematic undercoverage of households. Since these estimators are optimal (as far as maximizing the likelihood function) under this model, it may be expected that they will not do as well when systematic undercoverage is present. Since D_1 and D_3 are similar to D_5, this comment may also apply to them.

Method D_6 similarly corresponds to maximum likelihood, for multinomial sampling of persons. Let p_{jt} be the probability that a given sampled person is in cell j and in a household of type t. It can be shown that D_6 corresponds to the maximum likelihood estimates constrained by (2.1) and by the assumption that $p_{1t} = \ldots = p_{Jt}$ for each type t, for all cells j represented in that household type.

Although D_5 makes more sense in terms of the actual sample selection, D_6 may perform better when there is systematic undercoverage. An example of this occurs when the sample is subject to uniform undercoverage of all types of households. In this case, examples show that D_6, along with D_2, D_4, and the principal person method, weight the sample to represent the population's household types exactly, while D_5, D_1, and D_3 give too little weight to small units and too much to large units. (See Alexander (1986)).

V. A MODEL FOR COVERAGE

Although the methods D_1, D_3, or D_5 are in some sense approximately optimal when there is no systematic undercoverage, their properties in the presence of undercoverage are not so clear. To address this issue, a model for coverage is needed. A general model is the following:

For $t = 1,\ldots,T$ and $s = 1,\ldots, T$, let

Z_t = number of households in the population with actual type t

Z_+ = total number of households in the population

$K(t,s)$ = The probability that a type t household would have apparent type s if designated for interview

Let s = 0 correspond to a missing unit, so $K(t,0)$ is the probability that a household type t is missed from the frame

Note that $\sum_t Z_t\, K(t,s)$ is the expected number of households of apparent type s.

It will be assumed that $\sum_{s=0}^{T} K(t,s) = 1$, and that $K(t,s) = 0$ if $a_j(s) > a_j(t)$ for any j. Thus, there is assumed to be no systematic "overcoverage" of either persons or households. In practice, overcoverage of households is possible if errors in sampling give the same unit duplicate chances of selection. Failure to identify persons "with usual residence elsewhere" can lead to person overcoverage. Although these errors do occur to some extent, they are thought to be much less common than instances of undercoverage.

VI. ESTIMATION OF THE COVERAGE PARAMETERS

The values Z_t are estimated relatively accurately every ten years by the decennial census. The most useful data about household coverage from the periodic surveys are the "household coverage ratios" at the time of the decennial census,

$$C_t = \hat{Z}_t \,/\, Z_t,$$

where $\hat{Z}_t$ is the sum of the unbiased survey weights of sample households of apparent type t.

Unfortunately, a given set of household coverage ratios may be consistent with very different assumptions about $\{K(t,s)\}$. Two approaches to solving this problem will be taken. The first is to try a range of parametric models which can be estimated from $\underline{Z}$ and $\hat{\underline{Z}}$. The second approach incorporates additional information based on intermediate assumptions about various aspects of coverage.

a. Estimation of a range of models based on Z and $\hat{Z}$.

The distribution of $\hat{\underline{Z}}$ is related to $\{K(t,s)\}$ by the T equations

$$(6.1) \qquad E(\hat{Z}_s) = \sum_t Z_t \; K(t,s).$$

Thus, a parametric model for $\{K(t,s)\}$ with at most T parameters is needed. The parameters can be estimated from (6.1) using least squares. Alternatively, maximum likelihood can be used, if a specific distribution of $\hat{\underline{Z}}$ is assumed. Our experience with different models is limited so far, but the following two models for T=8 show how various assumptions can be incorporated into estimable models. These saturated linear models may be estimated by solving (6.1), substituting $\hat{Z}_s$ for $E(\hat{Z}_s)$.

The approach will be illustrated with J=2, representing males and females, assuming at most two persons of the same sex in an household. The household type will be described symbolically; for example "MF" will denote units with one male and one female. (j = 1 denotes "male" and j = 2 denotes "female".)

Model 1: (Low household undercoverage.)

This model has eight parameters (a,b,d,c,e,f,g,h) related to $\{K(t,s)\}$ as shown in Table 1

This model assumes that only M or F households may be missing. No more than one person in a household may be missing. The probability that one male is missing is allowed to vary according to the type of household. To reduce the number of parameters, the probability that one female is missing within an household is assumed to be constant.

Model 2: (No within-household undercoverage).

This model assumes $K(t,s) = 0$ unless $t = s$. Thus, there are eight parameters: $K(t,t)$, for $t=1,\ldots,8$.

The estimates from these two models for illustrative values of Z and $\hat{Z}$ are shown in Tables 2, 3, and 4.

b. More realistic model based on CPS Data (Model 3)

An additional model was considered. Data for the model were derived from the 1984 CPS and the 1983 and 1984 CE. For simplicity, and comparability with available data, the model restricted itself to households with at most three persons. Household composition was determined for household size and sex of household members.

As before, the subscript j took on only two values, 1 for males and 2 for females. The subscript t took on the values 1 through 9, plus a tenth for a household missed completely. Results of the comparisons for this model are available from the authors. Numbers of households were derived from Table 21 of the 1984 CPS report on household and family characteristics. (U.S. Bureau of the Census, 1985) As of 1984, there were an estimated 61,978,000 households with three or fewer persons; this figure was used for the base of the household estimates in the model. Breakdowns of numbers of households by size also agreed with the CPS report; to the extent that the report does not break down household composition by sex, assumptions were made based on the information provided in the report about married-couple family households with male and female heads.

The values described as Z_t in the previous section were granted under these assumptions. Values of the C_j's (person coverage ratios) are estimated from weighting output from the 1984 CE Quarterly (Interview) Survey. The full year's data from this survey indicate that the coverage rate for persons is about 90%, compared to updated decennial census counts of the number of persons of each sex. Coverage for males was somewhat lower and coverage for females somewhat higher than this figure.

Values of K(s,t) were computed based on the assumption that household and within-household undercoverage were about equal; i.e., if coverage of persons within a particular household type (say, a three-person household) was assumed to be about 90%, then about 5% of the undercoverage was assumed to be from completely missed households and the other 5% of the undercoverage was within households. Obviously, this assumption has an effect on the results; different assumptions can be made and their results analyzed.

The within-household component of undercoverage was further broken down by subjective assumptions of how often a particular type of household would fail to report persons by sex and number. To continue the example, of the 5% of undercoverage attibuted to persons missed within households of size 3, 4% of this might be attributed to missing one person and the remainder (1%) was from missing two people (on the assumption that a household was more likely to underreport one person than more). These raw undercoverage figures were then corrected for the number of persons in the household in order to retain the proper overall coverage ratio.

In future work, we intend to estimate further models using CE or CPS and census data, using a greater variety of person characteristics. Clearly, the complexity of the problem increases with the number of person cells. A realistic goal may be to include two races, both sexes, and three age categories, with a limit of two persons per household in each cell.

VII. COMPARISON OF THE METHODS

Each weighting method leads to an assignment of household weights $W_1,\ldots,W_K$ and corresponding total weights $W_1^*,\ldots, W_T^*$ for the T household types. Table 6 gives the weights W^*

for our hypothetical example. (Here W_t^* is the total weight given to units of apparent type t.)

Four ways have been discussed for trying to decide which assignment of weights is the best.

The first two ways of comparing or testing weighting methods do not require assumptions about the population parameters $\{K(t,s)\}$. The last two require such assumptions. The fourth test also requires additional assumptions about the variables the survey is designed to measure.

1. Compare the total estimated number of households W_+^* with the actual number of household Z_+. For the hypothetical data used to generate Models 1 and 2, this comparison is shown in Table 7. In the table, the methods are listed from best to worst according to this comparison.

2. Compare the individual values W_t^* to the actual values Z_t. This can be done by calculating D_i $(\underline{W}^*, \underline{Z})$ for any of the six distance measures introduced in Section III. This comparison is shown in Table 8 for all six difference measures. In this example, the rankings are the same for all the difference measures.

3. Assume a given model $\{K(t,s)\}$. Then compare $Z_t K(t,s)$ with the weighted total W_{ts}^*, defined to be the total weight given to sample units of actual type t which have apparent type s. This comparison may be made with any of the difference measures D_i. This test is not included in the illustration.

4. In addition to the assumptions about $\{K(t,s)\}$, make assumptions about the distribution of the variable of interest to the survey, for example expenditures. If $E(t,s)$ is the estimated mean expenditure for those households of actual type t which appear to have type s (using the unbiased survey weights), then the mean expenditure estimated by the constrained minimum distance method is

$$\left(\sum_{s,t} W_{ts}^* \; E(t,s) \right) \Big/ \sum_{s,t} W_{ts}^*$$

This may be compared to the actual population mean expenditure under the chosen assumptions about undercoverage and expenditures. For models 1 and 2, the comparison under one set of expenditure assumptions is made in Tables 9 and 10.

The four ways of testing the weighting methods have different strengths and weaknesses. The first two tests require no assumptions about coverage or expenditures. If the weights fail these tests, then there is a problem with the weights. However, even if the weights pass these two tests, there may be hidden problems because of within-household undercoverage. Passing these tests means that the total weights assigned to units of each apparent type agrees with the population count of units which actually are of that type. There may well be important differences between units of the same apparent type, if their actual type is different. There may also be important differences between units of the same actual type with different apparent types; failure to report persons to the interviewer may be related to the socioeconomic circumstances of the household and to these characteristics of the persons. This leads to the third test, which requires that each (t,s) combination be correctly weighed.

The third test does not guarantee that weighted expenditure estimates will be accurate. One problem is that missing units must be left out of the test. Indeed, for sample units of apparent type zero (missing units), there is no weighted total to be compared to Z_t $K(t,0)$. The second problem is that no adjustment is made for underreporting of expenditures. For some types of expenditures, such as clothing or personal items, undercoverage of persons is likely to imply underreporting of household expenditures.

(A possible solution to the first problem with Test 3 is to compare each W_{ts}^*, for $s \neq 0$, with

$$(7.1) \quad Z_t \; K(t,s) \; / \; (1 - K(t,0)).$$

This in effect allocates the missing type t units evenly across the nonmissing type t units. This allocation would be reasonable if the missing units had similar expenditure characteristics to the average of all units of their actual type.)

The fourth test goes further and examines the effect of the weighting on expenditure estimates under certain assumptions. A set of weights may pass this test by making compensating errors, for example giving too much weight to some units with higher-than-average expenditures to compensate for underreporting of expenditures by some other units. Clearly, a set of weights which passes the test for one type of expenditures may fail to pass the test for other characteristics of interest. Since only the apparent type of units in the survey is known, the assumptions about $\{K(t,s)\}$ and $\{E(t,s)\}$ are difficult to verify. It is necessary to seek weights which do well under a variety of plausible assumptions.

The above comparisons were also made for the data from Model 3. This work used methods D_1 and D_2, and a variant of D_4 which uses the arithmetic mean in place of the geometric mean. In comparing the "true" total number of households for this model (61,978,000) with the estimates from the various methods, it was found that the methods uniformly overestimated the number of households. The person-level methods tended to overestimate total households more than the household-level method. In the case of D_1, this seemed to be due to the many larger households in the model which were reported as smaller due to undercoverage, which results in estimates of the number of one-person households being greater than the true value even before the weights under the different methods are applied. In the case of D_2 and D_4, the overestimate of households seemed to be due to the fact that the initial weights are "weighted up" on the basis of sex alone, without regard to household size. Such overestimates of the number of households have not been seen in using D_1 on actual CE data. (Zieschang, 1986)

Estimates of mean expenditures for model 3 were derived under two different assumptions of reporting: (1) that each household would report actual total expenditures regardless of the number of persons missed in that household ("Reported as Actual"); and (2) that each household would report expenditures only for those members reported as being in the household ("Reported Only"). The "Reported as Actual" assumption uniformly produced overestimates of mean expenditures, probably due to the higher weights given larger households and the fact that true expenditures are carried through all household sizes. The "Reported Only" assumption uniformly produced underestimates of mean expenditures, perhaps because the underreporting of persons carries through the expenditure reporting. The expenditure assumptions were based on data in U.S. Department of Labor (1986).

VIII. SUMMARY

This paper compares six methods of assigning survey weights to households, constrained to be consistent with known counts of the number of persons in different person cells. Three of the methods (D_1, D_2, and D_4) have been investigated previously, and the others (D_3, D_5, and D_6) are added in this paper to round out the picture. Numerical results suggest that the three household-level methods (D_1, D_3, D_5) give nearly identical results, as do the three person-level methods (D_2, D_4, D_6).

Section IV contains results which may help in understanding what the methods actually do. Method D_5 corresponds to maximum likelihood estimation (subject to the constraints on person counts) for multinomial sampling, where the "cells" which define the multinomial random variables correspond to different household "types". A household type is defined by the number of persons in the household in each of the person cells. Method D_6 has a similar interpretation under another model for the sampling. Thus, it can be expected that methods D_1, D_3, and D_5 will do a good job reducing variance when multinomial sampling is a good model, in particular when there is no systematic undercoverage. However, the methods are not specifically suited to correcting for systematic undercoverage. The relationship of undercoverage to D_2, D_4, and D_6 needs further theoretical study.

A general approach to empirical studies of the effect of systematic undercoverage on these methods is described in Section V, VI, and VII, and illustrated with some hypothetical data. Each model parameter corresponds to the probability that a household of one type (as defined above) is missing or appears to have a different type. Under various assumptions about the relationships of these probabilities, the parameters may be estimated from decennial census counts and weighted survey estimates for the household types. A simplified version of the present principal person method is suggested so that it can be included in these model-based empirical studies.

To compare the various methods, it may not be sufficient to examine the weighted total number of households, or even the total weight given to all sample households of different apparent types. It may be necessary to make assumptions about the joint distribution of apparent and actual household types and about the distribution of the variable of interest, such as expenditures. (The third test in Section VII as modified by (7.1) may suggest a way of making comparisons with less detailed assumptions about expenditures. This requires further study.)

Future research on CE weighting will investigate the use of independent estimates of the number of households by household size and possibly other characteristics. Scheuren, et al (1981), Roman (1983) and Zieschang (1986) have described ways of doing this using household and person controls simultaneously. It may or may not be possible to construct such estimates for the household types described in Section IV. Depending on how the independent household estimates are constructed, the problem may be complicated by inconsistency between the independent household and person estimates. The coverage models of Section V can also be applied to evaluate the performance of these weighting methods under various assumptions about coverage and expenditures.

ACKNOWLEDGEMENT

The authors wish to thank Ms. Brenda Kelly for her typing and other assistance in preparing this manuscript.

REFERENCES

Alexander, C.H. (1986). "Alternatives to the Generalized Least Squares and Principal Person Methods for Consumer Expenditure Surveys Weighting." Census Bureau internal report dated February, 1986.

Cressie, N and Read, T.R.C. (1984). "Multinomial Goodness-of-Fit Tests." Journal of the Royal Statistical Society (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", presented at the 1986 annual meeting of the Population Association of America.

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.

Fienberg, S.E., (1986). "Comments on Some Estimation Problems in the Consmer Expenditure Surveys". **Population Controls in Weighting Sample Units**, collected papers from a conference sponsored by the Bureau of Labor Statistics, March 18, 1986.

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," **ASA Proceedings of Social Statistics Section**, 300-305.

Luery, D. (1980). "An Alternative to Principal Person Weighting". Census Bureau internal memorandum, dated July 8, 1980.

Oh, H.L. and F. Scheuren (1978). "Multivariate Raking Ratio Estimation in the 1973 Exact Match Study." **ASA Proceedings of Survey Methods Research Section**, 175-182.

Roman, A.M. (1983). "The Consumer Expenditure Surveys - Specifications for Weighting Research." Internal Census Bureau memorandum dated May 13, 1983.

Scheuren, F., Oh, H.L., Vogel, L. and Yuskavage, R. (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, Social Security Administration, publication No. 13-11750.

Stephen, F.F. (1942), "An Iterative Methods of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known." **Annals of Mathematical Statistics**, 13, 166-178.

Zieschang, K.D. (1986). "Generalized Least Squares: An Alternative to Principal Person Weighting". **Population Controls in Weighting Sample Units**, collected papers from a conference sponsored by the Bureau of Labor Statistics, March 18, 1986.

U.S. Bureau of the Census (1985). Current Population Report, Series P-20, No. 398, **Household and Family Characteristics: March 1984**, U.S. Government Printing Office, Washington, D.C., 1985

U.S. Department of Labor, Bureau of Labor Statistics (1986). Bulletin 2246, **Consumer Expenditure Survey: Interview Survey, 1982-83**, U.S. Government Printing Office, Washington, D.C., 1986.

TABLE 1

MODEL 1: PROBABILITIES {K(t,s)}

Actual Type	Apparent Type MMFF	MMF	MFF	MF	MM	FF	M	F	Missing
MMFF	1-a-f	f	a	-	-	-	-	-	-
MMF	-	1-b-f	-	b	f	-	-	-	-
MFF	-	-	1-c-f	f	-	c	-	-	-
MF	-	-	-	1-d-f	-	-	f	d	-
MM	-	-	-	-	1-e	-	e	-	-
FF	-	-	-	-	-	1-f	-	f	-
M	-	-	-	-	-	-	g	-	1-g
F	-	-	-	-	-	-	-	h	1-h

TABLE 2

HYPOTHETICAL VALUES FOR Z AND $\hat{Z}$ IN THOUSANDS

TYPE:	MMFF	MMF	MFF	MF	MM	FF	M	F
Z	5000	6000	6500	20000	3300	4600	9000	12425
$\hat{Z}$	4800	5800	6300	19000	3200	4550	8000	11800

TABLE 3

MODEL 1: POPULATION EXPECTED VALUES FOR Z AND $\hat{Z}$ IN THOUSANDS

Actual Type	Apparent Type MMFF	MMF	MFF	MF	MM	FF	M	F	Missing	Total
MMFF	4800	140	60	-	-	-	-	-	-	5000
MMF	-	5600	-	172	167	-	-	-	-	6000
MFF	-	-	6240	182	-	79	-	-	-	6500
MF	-	-	-	18646	-	-	559	795	-	20000
MM	-	-	-	-	3033	-	268	-	-	3300
FF	-	-	-	-	-	4471	-	129	-	4600
M	-	-	-	-	-	-	7173	-	1827	9000
F	-	-	-	-	-	-	-	10877	1548	12425
	4800	5800	6300	19000	3200	4550	8000	11800	3395	

TABLE 4

MODEL 2: POPULATION EXPECTED VALUES (Z_t K(t,s)) IN THOUSANDS

Actual Type	Apparent Type MMFF	MMF	MFF	MF	MM	FF	M	F	Missing	Total
MMFF	4800	-	-	-	-	-	-	-	200	5000
MMF	-	5800	-	-	-	-	-	-	200	6000
MFF	-	-	6300	-	-	-	-	-	200	6500
MF	-	-	-	19000	-	-	-	-	1000	20000
MM	-	-	-	-	3200	-	-	-	100	3300
FF	-	-	-	-	-	4500	-	-	50	4600
M	-	-	-	-	-	-	8000	-	1000	9000
F	-	-	-	-	-	-	-	11800	625	12425
	4800	5800	6300	19000	3200	4550	8000	11800	3575	

Table 5 in the paper handed out in the session has been omitted from the proceedings version.

TABLE 6

ESTIMATES FROM THE SEVEN MODELS

Weighted Estimates (in 1000s) based on $\hat{Z}$, assuming P_1= 64100 and P_2 = 70625

TYPE:	MMFF	MMF	MFF	MF	MM	FF	M	F
$\hat{\underline{Z}}$:	4800	5800	6300	19000	3200	4550	8000	11820
W^* For D_1:	5171	6199	6597	19734	3393	4627	8241	11900
$\underline{W}^*$ for D_3:	5175	6201	6596	19728	3392	4627	8237	11899
$\underline{W}^*$ for D_5:	5179	6202	6596	19722	3392	4626	8233	11898
$\underline{W}^*$ for D_2:	5026	6113	6554	19895	3416	4671	8541	12114
$\underline{W}^*$ for D_4:	5025	6113	6554	19894	3417	4672	8542	12116
$\underline{W}^*$ for D_6:	5025	6113	6553	19893	3417	4672	8543	12117
W^* for princial person method	4993	6033	6553	19763	3368	4733	8420	12273
Assumed $\underline{Z}$:	5000	6000	6500	20000	3300	4600	9000	12425

TABLE 7

COMPARISON OF TOTAL ESTIMATED NUMBER OF HOUSEHOLDS

"Actual":		66825
D_6	:	66334
D_4	:	66332
D_2	:	66330
Princ. Pers.:		66135
D_1	:	65864
D_2	:	65857
D_3	:	65850

TABLE 8

(COMPARISON OF METHODS FOR DATA IN TABLE 2)

DISTANCE ($D_i(W^*, \underline{Z})$) BETWEEN "ACTUAL" AND APPARENT DISTRIBUTION FOR THE SEVEN WEIGHTING METHODS

Weighting Method	Distance Measure D_1	D_2	D_3	D_4	D_5	D_6
D_6	39.4	50.7	19.9	25.5	20.1	25.7
D_6	39.5	50.8	19.9	25.5	20.2	25.8
D_2	39.6	50.9	20.0	25.6	20.2	25.8
Princ. Pers.	47.8	57.1	24.3	28.9	24.7	29.4
D_1	106.3	146.4	54.2	74.0	55.3	74.9
D_3	107.6	148.8	54.8	75.2	55.9	76.1
D_5	108.9	151.3	55.5	76.5	56.6	77.4

TABLE 9

HYPOTHETICAL EXPENDITURES BY ACTUAL x APPARENT HOUSEHOLD TYPE

EXPENDITURES ($1000's)

	MMFF	MMF	MFF	MF	MM	FF	M	F	Missing
MMFF	25	20	20	20	20	20	20	20	20
MMF	-	21	-	17	17	-	17	17	17
MFF	-	-	21	17	-	17	17	17	17
MF	-	-	-	19	-	-	15	15	15
MM	-	-	-	-	19	-	15	-	15
FF	-	-	-	-	-	19	-	15	15
M	-	-	-	-	-	-	11	-	9
F	-	-	-	-	-	-	-	11	9

TABLE 10

WEIGHTED MEAN EXPENDITURE

	Model 1	Model 2
"Actual"	$17,002	$17,102
Princ. Pers.	$17,624	$17,330
D_6	$17,641	$17,345
D_4	$17,641	$17,345
D_2	$17,642	$17,345
D_1	$17,718	$17,413
D_3	$17,719	$17,414
D_5	$17,720	$17,415