A METHOD OF OBTAINING GROUP WEIGHTS FROM INDIVIDUAL WEIGHTS IN A SURVEY

Edward Gbur, U.S. Bureau of the Census

1. Introduction

In many surveys individuals can be naturally grouped into units. The grouping may result from the nature of the sampling design or from the nature of the information collected by the survey. For example, in a demographic survey the sampling frame may consist of housing units from which a sample is selected. Estimates may be required for individuals as well as for households and subhousehold units such as families.

Survey weights are usually calculated at the lowest level, here the sampled individual. Estimated levels for personal characteristics are obtained by summing the weights associated with all individuals possessing the characteristic of interest. To obtain estimates for characteristics defined at the group level, a weight must be assigned to each unit (group of individuals). A commonly used assignment method designates one individual in each unit, hereafter referred to as a principal person, to represent the unit and uses that individual's weight in group tabulations. This procedure can lead to inconsistencies in the tabulated results from the sample if the person's weight is used. To avoid these contradictions an additional adjustment is often made to the principal person's weight to obtain a new weight for use in group level estimates.

In this paper we describe several alternative methods of obtaining group weights. The Current Population Survey will be used to illustrate the methods. Other examples where similar procedures could be applied include the National Crime Survey and the Consumer Expenditure Surveys.

2. The CPS Example

The Current Population Survey (CPS) is a longitudinal address survey conducted by the Census Bureau for the Bureau of Labor Statistics. It produces information on a wide variety of characteristics of the U.S. labor force. CPS utilizes a stratified multistage cluster design coupled with a rotating panel structure. Estimates are produced at the individual, family, and household levels. Each sampled individual is assigned a weight derived from the reciprocal of the housing unit's probability of selection, multiplied by a set of factors obtained from post-stratification to account for noninterviews, sampling of PSUs, and the age-race-sex distribution of the population.

If these person weights are used for family and household tabulations, internal inconsistencies can result. For example, for married, spouse present (MSP) families, the sum of all husbands' weights should equal the sum of all wives' weights since they both estimate the number of MSP families. However, the two sums differ when the person weights are used. The difference arises from the age-race-sex post-stratification adjustment. To avoid such inconsistencies an additional adjustment is made to the person weights to create a family weight for use in family and household estimates. A summary of the steps in this final stage is given below. The rationale for the procedure is discussed in Bureau of the Census Technical Paper 40 (1978).

Step 1: In MSP households, change the husband's final person weight to the wife's final person weight.

Step 2: In MSP households, for each age-race category calculate

$$f_{ar} = \Sigma(\text{female MSP weight}) \,/\, \Sigma(\text{male MSP weight}) .$$

Step 3: For other male heads (OMH), multiply the final person weight by the appropriate factor f_{ar}.

Step 4: For each male age-race category, calculate

$$f^*_{ar} = [\Sigma(\text{original CPS weight}) - [\Sigma(\text{adjusted weights for male MSP and OMH})] \,/\, [\Sigma(\text{original CPS weight}) - \Sigma(\text{original CPS weights for male MSP and OMH})] .$$

Step 5: For all other males (AOM) for each age-race category, multiply the final person weight by the appropriate factor f^*_{ar}.

All of the steps in the scheme are performed on the entire sample rather than on each panel (rotation group) separately. All family and household tabulations in CPS use the adjusted (or family) weight of the head of the family or household. The head is defined to be either the husband or wife in a MSP family and is a self-declared adult in all other households and families.

3. Alternative Approaches

Three alternatives to the general procedure described above were considered. Each will be described in the context of the CPS example and in general terms wherever possible.

The procedure in Section 2 essentially designates one person (the head) to represent the entire group and modifies his or her final person weight to produce a weight which supposedly represents the entire group. The first proposal is to construct a group weight which is a function of the characteristics of the group as a collection of individuals rather than as a function of the designated individual within the group. This new group weight is not attached to a particular individual but is used for tabulations of all group related items.

In the CPS example, the construction of such a group weight for families would begin with the usual CPS basic weight, noninterview adjustment, and first stage ratio adjustment (for the sampling of PSUs in each stratum). These adjustments would require knowledge of the race of either the head or the householder but do not require the use of that individual's person weight. One or more other ratio adjust-

ments based on the household's characteristics would follow these stages. Examples of such factors would be

(i) residence status (e.g., SMSA vs. non-SMSA; urban vs. rural),
(ii) tenure status (own, rent, no cash rent),
(iii) household structure (primary family only, primary and secondary families, primary family and unrelated persons, etc.),
(iv) characteristics of the primary family (e.g., family size; racial composition - black, nonblack, biracial).

The particular factor or factors used would be determined by further research.

The advantage of this approach is that the unit's characteristics determine the weight used in tabulations concerning the unit as a group of individuals. It is more intuitively appealing that, for example, an estimate of the number of black families with a single parent and one or more children depends on factors such as those listed above rather than on the age, race, sex, and marital status of the person designated as the head.

There are two major drawbacks to this approach. First, control counts may not be easily obtainable for the additional factors to be used for ratio adjustments. The use of some type of administrative records may be the only source of independent control counts. Second, there is no guarantee that new inconsistencies will not arise when tabulation results based on group weights are compared with those based on person weights. It is this second drawback which makes the implementation of this type of procedure unlikely.

The second proposal retains the idea of an additional modification of the final person weight. The motivation for this alternative is the arbitrary assignment of the wife's final person weight to the husband in MSP families. Under this proposal the final CPS person weight calculation would not be changed.

For MSP families, independent control counts for the joint age distribution of the husbands' and wives by racial category (e.g., both black, both nonblack, biracial) would be used to form the usual ratio adjustment factor "control count/sample total" for each cell. This would eliminate the arbitrariness of the present MSP adjustment. Controlling to the joint age distribution would ensure the consistency of the sums of the husbands' and wives' family weights.

In theory the independent controls for the joint age distribution by race for MSP families could be obtained by updating census estimates. At each time point the distribution would be updated to account for

(i) aging of the couples,
(ii) additions (marriages, immigration),
(iii) deletions (divorces, separations, deaths, migration).

The practical problem is the quality and timeliness of the data necessary for updating. The required information for marriages and divorces is not available for all states. Preliminary estimates from available data have recently been published for 1982. Final estimates are available for 1979. Death records are available from all states and since 1981 should contain the necessary age, race, sex, and marital status information. The time lag for death records is not quite as long. The missing data, the quality of the available data, and the lack of timeliness of the final estimates make this approach infeasible at the present time. In the future this source of control counts may be more useful than at present.

The final and most promising proposal is based on the application of the generalized least squares (GLS) principle to obtain ratio adjustment factors for each sampled individual. It represents a second alternative which adjusts the final person weight to obtain a group weight. The idea of adjusting frequency table entries using the method of least squares can be traced to Deming and Stephan (1940). We have taken the adjustment idea one step further and used the adjusted entries to form post-stratification ratio factors.

Our final proposal can be briefly summarized as follows. Each sampled individual is cross-classified by two or more factors (usually those used in the present adjustment of the person weight) and the sum of the final person weights of the individuals in each cell is obtained. The method of constrained generalized least squares is used to find new cell totals which have the property that they are as close as possible (in the sense of squared error) to the original cell totals and satisfy the constraints which have been imposed on the cross-classification to achieve internal consistency in the survey tabulations. The new adjusted cell totals can be used to form ratio adjustment factors to apply to the final person weights to obtain group weights.

In the context of the CPS example, each sampled individual is cross-classified by race, sex, age, and household relationship status (MSP, O*H, AO*, where * is M or F). For each race category, two tables are constructed as shown in Figure 1. The table entries are the sums of the final CPS person weights of all sampled individuals in that cell. The known age distributions by race for males form independent column control counts for the male tables. There are corresponding row controls for the female tables. There are no controls for the rows in the male tables nor for the columns in the female tables. The upper portion of the male tables and the leftmost portion of the female tables form two estimates of the joint age distribution for MSP by race. The sum over race for a cell in this portion of the male tables should be identical to the corresponding sum in the female tables. Summing over race takes into account biracial marriages. In general these corresponding sums will not be equal when final person weights are used.

The method of GLS is used to produce "adjusted" cell totals for all four tables simultaneously subject to the marginal age distribution constraints and the cell by cell MSP equality constraints. These adjusted cell totals are used to compute ratio adjustment factors for each cell. These factors can then be applied to the final CPS person weight of

Figure 1

Table for Males* (by Race)

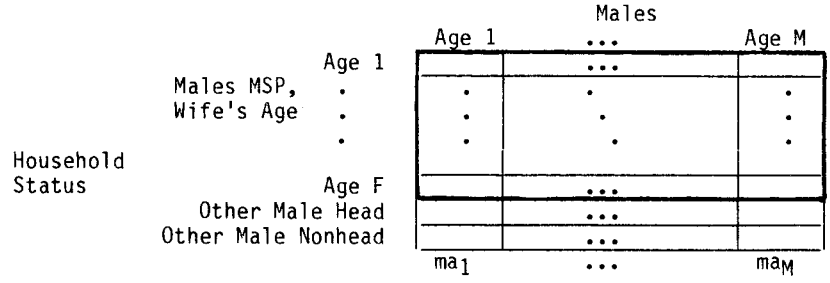
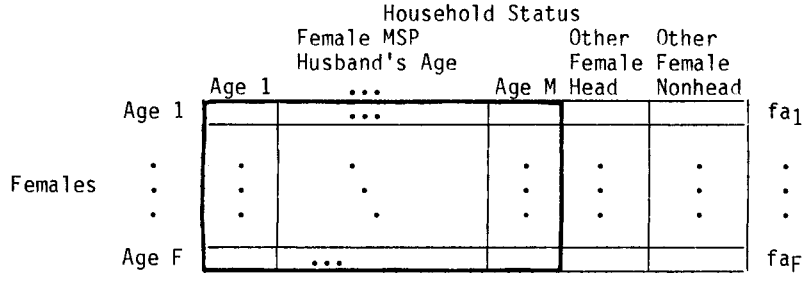


Table for Females* (by Race)



*The MSP categories of each table are set off by dark borders.

each individual in the cell to obtain the individual's family weight.

4. The Generalized Least Squares Approach to CPS

The constrained, generalized least squares method outlined in the previous section can be formalized as follows for the CPS example. Define T_{ijkM} to be the observed total in the $(i,j)^{th}$ cell of the male table for race k. Define T_{ijkF} similarly for females. Note that the subscripts i and j have switched definitions for males and females. That is, for the male tables, i represents the household relationship status and j the individual's age category. For the female tables, i represents the age category and j the household relationship status. There is, however, consistency in the MSP portion of the tables and this use of subscript notation simplifies the presentation of the GLS calculations.

Let τ_{ijkM} and τ_{ijkF} be the adjusted cell totals corresponding to T_{ijkM} and T_{ijkF}, respectively. These are the parameters to be estimated. The GLS method finds the τ-values which minimize

$$Q(\tau) = \sum_{k=1}^{2} \Big[\sum_{i=1}^{F+2} \sum_{j=1}^{M} w_{ijkM} (T_{ijkM} - \tau_{ijkM})^2 + \sum_{i=1}^{F} \sum_{j=1}^{M+2} w_{ijkF} (T_{ijkF} - \tau_{ijkF})^2 \Big], \quad (1)$$

subject to the constraints

$$\sum_{i=1}^{F+2} \tau_{ijkM} = ma_{jk} \quad \text{for } j=1, \ldots, M; \; k=1,2,$$

$$\sum_{j=1}^{M+2} \tau_{ijkF} = fa_{ik} \quad \text{for } i=1, \ldots, F; \; k=1,2, \quad (2)$$

$$\tau_{ij1M} + \tau_{ij2M} = \tau_{ij1F} + \tau_{ij2F} \quad \text{for } i=1, \ldots, F; \; j=1, \ldots, M,$$

where w_{ijkM} and w_{ijkF} are weights whose determination will be discussed later and the ma_{jk} and fa_{ik} are obtained from the known age distributions of males and females by race, respectively.

Technically the constraints (2) should be supplemented by the inequality constraints

$$\tau_{ijkM} \geq 0 \text{ and } \tau_{ijkF} \geq 0 \text{ for all } i, j, k.$$

Enforcing these inequality constraints makes the minimization of (1) much more difficult. In addition, they only need to be considered if the minimization of (1) subject to (2) produces estimates which violate them.

The model corresponding to the sum of squares function $Q(\tau)$ in (1) subject to the constraints (2) can be written in matrix notation as

$$T = X\tau + e,$$
$$A\tau = c \ , \qquad (3)$$

where

$$T = [T_{111M}, T_{211M}, \ldots, T_{(F+2)M2M}, T_{111F}, \ldots, T_{F(M+2)2F}]' \ ,$$

and τ is the parameter vector with entries in the order corresponding to that in T, X is the design matrix for the model, A is the coefficient matrix for the constraints, c is the vector of constants for the constraints, and the random vector $e \sim (0,\Sigma)$.

Let W be a general weight matrix. Then $Q(\tau)$ can be written in matrix form as

$$Q(\tau) = (T - X\tau)' W^{-1} (T - X\tau) \ .$$

The constrained GLS estimator of τ is given by

$$\hat{\tau}_A = \hat{\tau} + (X'W^{-1}X)^{-1} A' [A(X'W^{-1}X)^{-1} A']^{-1} \cdot (c - A\hat{\tau}) \ , \qquad (4)$$

where $\hat{\tau}$ is the unconstrained GLS estimator; i.e.,

$$\hat{\tau} = (X'W^{-1}X)^{-1} X'W^{-1}T.$$

For the CPS example, the design matrix X is an identity matrix so that

$$\hat{\tau} = WW^{-1}T$$
$$= T$$

and the expression for the constrained estimator $\hat{\tau}_A$ in (4) reduces to

$$\hat{\tau}_A = T + WA'(AWA')^{-1}(c - AT) \ . \qquad (5)$$

There are several possible weight matrices W which could be used. In our original formulation of the CPS problem W was taken to be diagonal. In general, a nondiagonal matrix W can be considered. Deming and Stephan (1940) used a diagonal weight matrix in which the diagonal entries were the reciprocals of the observed cell totals. An observed zero would be replaced by a prespecified large positive constant. Alternatively, it may be desirable to have the weights reflect the covariance structure of the observed totals T. In this case, $W = \Sigma$. A third alternative would allow the weights to reflect survey related factors such as coverage rates for the various cells.

In the model (3) there are $2M(F+2) + 2F(M+2)$ parameters and $2M + 2F + MF$ constraints. In the current CPS weighting procedure, F=17 and M=17 so that there are 1292 parameters and 357 independent constraints. This yields 935 independent parameters to be estimated. The vector τ_A has 1292 entries. The matrix AWA' to be inverted in (5) is a 357×357 matrix. If W is a diagonal matrix, then the block structure of the coefficient matrix A can be exploited to avoid the direct calculation of $(AWA')^{-1}$. A derivation is sketched in the Appendix. The only nondiagonal matrix which needs to be inverted has dimensions $2(M+F) = 68$.

5. Conclusion

Several methods have been presented for obtaining group weights from individual weights. The most promising of these methods is based on constrained generalized least squares. The CPS is used to illustrate the method.

In this paper GLS was used in the formation of ratio factors for an additional post-stratification ratio adjustment. However, a GLS based adjustment could theoretically be used as a replacement for the entire weighting procedure to produce a common weight (the group weight) for all individuals within a group. Each unit would be cross-classified by the usual control factors, which would also provide the entries of the constant vector in the constraints. Unfortunately, the response vector would contain an entry for each unit in the sample and matrix manipulations would be required for matrices with dimensions equal to the number of groups in the sample. This would probably cause computational problems which would make the method impractical.

References

Deming, W.E. and F.F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. Annals of Math. Stat., 11, 427-444.

Graybill, F.A. (1983). Matrices with Applications in Statistics. second edition. Wadsworth, Inc., Belmont, CA.

U.S. Bureau of the Census (1978). The Current Population Survey: Design and Methodology. Technical paper 40.

Appendix

The matrix A in the model (3) can be written in block form where the division of the rows corresponds to the three sets of constraints (2) and the division of the columns corresponds to the four tables in Figure 1.

Let

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ 0 & 0 & A_{23} & 0 \\ 0 & A_{32} & 0 & 0 \\ 0 & 0 & 0 & A_{44} \\ A_{51} & A_{52} & A_{53} & A_{54} \end{bmatrix} \ ,$$

where

$$A_{11} = A_{23} = 1_{F+2} \otimes I_{M\times M} \ ,$$

$$A_{32} = A_{44} = I_{F\times F} \otimes 1_{M+2} \ ,$$

$$A_{51} = -A_{53} = [\, I_{F\times F} \ , \ 0_{F\times 2}] \otimes I_{M\times M} \ ,$$

$$A_{52} = -A_{54} = I_{F\times F} \otimes [\, I_{M\times M} \ , \ 0_{M\times 2}] \ ,$$

and I represents the identity matrix, 1 a vector of ones, 0 a matrix of zeros, and $\otimes$ the Kronecker product (cf., Graybill (1983), Section 8.8).

Assume that W is a diagonal matrix which has been partitioned to correspond to the four tables in Figure 1; i.e.,

$$W = \text{diag}\,[W_{11}, W_{12}, W_{21}, W_{22}] \ ,$$

where W_{ij} is the weight matrix associated with the table defined for race i and sex j. Then

$$AWA' = \begin{bmatrix} V_{11} & 0 & 0 & 0 & V_{15} \\ 0 & V_{22} & 0 & 0 & V_{25} \\ 0 & 0 & V_{33} & 0 & V_{35} \\ 0 & 0 & 0 & V_{44} & V_{45} \\ V_{15}' & V_{25}' & V_{35}' & V_{45}' & V_{55} \end{bmatrix} \ ,$$

where the V_{ij} are simple functions of the A_{ij} and W_{ij}. To obtain $(AWA')^{-1}$, repartition the matrix AWA' as

$$AWA' = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \ ,$$

where

$$U_{11} = \text{diag}\,[\, V_{11}, V_{22}, V_{33}, V_{44}\,] \ ,$$

$$U_{22} = V_{55} \ ,$$

$$U_{21} = [\, V_{15}' \ , V_{25}' \ , V_{35}' \ , V_{45}'] \ ,$$

$$U_{12} = U_{21}' \ .$$

Both U_{11} and U_{22} are diagonal matrices. To obtain $(AWA')^{-1}$, the formulas for the inverse of a partitioned matrix (cf., Graybill (1983), Section 8.2) can be applied to the latter partition of AWA'. Only U_{11} and U_{22} need to be inverted.