AN ALTERNATIVE METHOD OF CONTROLLING CURRENT POPULATION SURVEY ESTIMATES TO POPULATION COUNTS

KENNON R. COPELAND, FLORA KWOK PEITZMEIER AND C. EASLEY HOY
BUREAU OF LABOR STATISTICS

1. INTRODUCTION

The Current Population Survey (CPS) produces monthly labor force and related estimates for the total U.S. civilian noninstitutional population, subaggregated by race, ethnicity (Hispanic/nonHispanic orgin), sex, age, and state. In addition, estimates for a number of population subdomains (e.g., families, veterans, wage and salary earners, persons not in the labor force) are also produced on either a monthly or quarterly basis. The data for all these estimates are obtained in monthly interviews of a sample of housing units in the U.S.

The CPS employs raking ratio estimation (RRE) to derive the weights used to tabulate total U.S. estimates. The estimates for the population subdomains make use of weights derived from adjustment procedures built on top of the weights used for the full population data. These adjustment procedures are used to accomplish one or more of the following goals: 1) provide a single weight for a family unit; 2) provide consistency between estimates for the subdomain and estimates for the full population; 3) provide agreement with a set of independent population controls.

An alternative estimation technique to integrate adjustment to population controls with family weighting, using a generalized least squares (GLS) approach, has been proposed and investigated for use in the Consumer Expenditure Survey (Zieschang, 1986). This estimation technique could potentially be used to integrate the various CPS weighting procedures, and possibly could alleviate undesirable aspects associated with the current system.

Given the perceived potential of the GLS procedure, an investigation into the application of the GLS procedure for CPS was begun. As a first step in the investigation, the current CPS RRE procedure was compared to a GLS estimation procedure using CPS data from July 1983 and July 1984.

2. BACKGROUND

A. CPS Sample Design and Weighting

The CPS is a multi-stage probability sample of approximately 60,000 housing units in the U.S. Each month a rotating sample of 8 panels (called rotation groups) of housing units is interviewed, with demographic and labor force data being collected for all civilian occupants of the sample housing units.

CPS employs several weighting and estimation procedures to derive national and state estimates from the survey data:

- Weighting to account for probability of selection;
- Adjustment for noninterview;
- Adjustment for sampling of PSUs (called first-stage ratio adjustment);
- Adjustment to independent population estimates (called second-stage ratio adjustment);
- Composite estimation;
- Seasonal adjustment.

The second-stage ratio adjustment is the procedure investigated in this first step of the research into the application of alternative adjustment procedures for the CPS.

B. Undesirable Aspects of the CPS Weighting System

Perhaps foremost among these is the fact that while changes have been made in the various components of the weighting system over time, rarely is an evaluation made of the impact the change may have on other stages of weighting or of the comparability of estimates produced from various stages of weighting.

The use of the various population subdomain weighting procedures has several undesirable aspects, among them: 1) the possibility of large differences in the weights used for the full population data and the subdomain data for a sample record; 2) inconsistency or noncomparability of data derived by the various adjustment procedures; and 3) multiple weights for sample records (as many as 5 weights).

3. RAKING RATIO ESTIMATION

A. The Raking Ratio Estimation (RRE) Procedure

RRE is a procedure to iteratively adjust sample data to known (outside) marginal totals. Consider a three-way table of (weighted) sample counts $[n_{ijk}]$ for which adjusted sample counts $[\tilde{n}_{ijk}]$ are desired under the condition that

$$\text{(A)} \qquad \sum_{j,k} n_{ijk} = m_{i\cdot\cdot}$$

(B) $$\sum_{i,k} n_{ijk} = m_{\cdot j \cdot}$$

(C) $$\sum_{i,j} n_{ijk} = m_{\cdot\cdot k}$$

be satisfied simultaneously

where $m_{i\cdot\cdot}$ = known (control) row total

$m_{\cdot j \cdot}$ = known (control) column total

$m_{\cdot\cdot k}$ = known (control) 'layer' total

The RRE procedure proportionately ratio adjusts the sample data each way (e.g., row, column, and layer) of the table in successive steps, as follows:

(1) Ratio adjustment by row

$$n_{ijk}^{(1i)} = (m_{i\cdot\cdot} / n_{i\cdot\cdot})\, n_{ijk}$$
$$= a_i^{(1)} n_{ijk}$$

(2) Ratio adjustment by column

$$n_{ijk}^{(1j)} = (m_{\cdot j\cdot} / n_{\cdot j \cdot}^{(1i)})\, n_{ijk}^{(1i)}$$
$$= b_j^{(1)} n_{ijk}^{(1i)}$$
$$= a_i^{(1)} b_j^{(1)} n_{ijk}$$

(3) Ratio adjustment by layer

$$n_{ijk}^{(1k)} = (m_{\cdot\cdot k} / n_{\cdot\cdot k}^{(1j)})\, n_{ijk}^{(1j)}$$
$$= d_k^{(1)} n_{ijk}^{(1j)}$$
$$= a_i^{(1)} b_j^{(1)} d_k^{(1)} n_{ijk}$$

where $n_{i\cdot\cdot}$ = sample row total

$n_{\cdot j\cdot}$ = sample column total

$n_{\cdot\cdot k}$ = sample layer total

The completion of the three adjustment steps constitutes one iteration of the raking process. The three steps are repeated substituting the current value of $n_{ijk}^{(hk)}$ (adjusted sample count following the 3rd way rake of the hth iteration) for n_{ijk} in step (1) each time until a prespecified number of iterations are completed or until conditions (A), (B), and (C) are satisfied simultaneously to some specified degree of closeness. The final $[n_{ijk}^{(hk)}]$ is taken as $[\tilde{n}_{ijk}]$.

If adjustment of the sample record weights is desired, the factor is (assuming g iterations)

$$F_{ijk} = n_{ijk}^{(gk)} / n_{ijk}$$
$$= \prod_{h=1}^{g} a_i^{(h)} b_j^{(h)} d_k^{(h)}$$

The record weights prior to RRE are multiplied by the appropriate F_{ijk} to obtain the adjusted weight.

B. **Properties of RRE**

1. Under simple random sampling, the $[\tilde{n}_{ijk}]$ resulting from RRE are best asymptotically normal (BAN) estimates (Ireland and Kullback).
2. RRE, while producing biased estimates, can sometimes be effective in reducing the mean square error of survey estimates (Hanson), although this is not guaranteed.
3. Although not guaranteed, convergence should occur (Oh and Scheuren, 1978-a) and it has in many CPS and other survey applications of RRE. However, the best way to verify convergence is to run the RRE procedure. To achieve convergence, one must be 'reasonable' when setting up the constraints on the sample, especially with small sample size.
4. All sample records within the same estimation cell (ijk) receive the same adjustment factor F_{ijk}.
5. When convergence is achieved, RRE minimizes the statistic

   $$\sum_{ijk} p_{ijk} \ln (p_{ijk} / \pi_{ijk})$$

   where: $p_{ijk} = \tilde{n}_{ijk} / \tilde{n}_{\cdot\cdot\cdot}$

   $\pi_{ijk} = n_{ijk} / n_{\cdot\cdot\cdot}$

   for all adjustment procedures satisfying the constraint conditions (Ireland and Kullback).

C. **Application of RRE in the CPS**

The CPS second-stage ratio adjustment procedure currently uses a three-way RRE. The procedure adjusts the weights for sample records within a rotation group (RG) following first-stage adjustment so as to control the sample estimates for a number of geographic and demographic groups of the population to independently derived estimates of the population in each of these categories. The second-stage ratio adjustment procedure uses state, age/sex/ethnicity, and age/sex/race census-based estimates of the current population as controls.

The CPS second-stage ratio adjustment procedure has the following steps:

1. First-way (row) rake by State and RG to independently derived State and D.C. estimates of total population.
2. Second-way (column) rake by age/sex/ethnicity and RG at the national level.
3. Third-way (layer) rake by age/sex/race and RG at the national level.

Repeat the above steps five more times. The final second-stage weight is achieved after the sixth iteration of the three-way rake.

4. GENERALIZED LEAST SQUARES ESTIMATION

A. The Generalized Least Squares (GLS) Estimation Procedure

The GLS procedure adjusts the sample weights from prior stages of weighting by minimizing the weighted squared adjustments subject to a set of linear 'control' constraints the adjusted weights must satisfy.
The goal is to minimize

$$f(\omega) = (\omega - \Omega)' \; \Omega_0^{-1} \; (\omega - \Omega)$$

subject to the constraint $X'\omega = N$

where ω = (n x 1) vector of desired final weights (ω_i) for each record

Ω = (n x 1) vector of weights (Ω_i) from prior stages of estimation for each record

$_0$ = (n x n) diagonal matrix with the Ω_i's on the diagonal

X = (n x k) design matrix where the columns of the i^{th} row define the control cells in which the i^{th} sample record is contained (i.e., X_{ij} is either 0 or 1)

N = (k x 1) vector of controls for each of the control categories corresponding to the columns of X

The Ω_i's are desired to change as little as possible (and thus reduce the chance of extreme weights) so that ω_i is to be 'closest' to the Ω_i (the original weights) in the least squares sense while minimizing f(ω).

The unique solution to $X'\omega = N$ that minimizes the specified f(ω) was found to be (Luery)

$$\omega = \Omega + \Omega_0 X (X' \Omega_0 X)^{-1} (N - X'\Omega)$$

B. The Properties of GLS Estimation

1. When the cells of the contingency table are all nonempty, the GLS procedure generates (BAN) estimates. (Neyman).
2. The GLS procedure is designed to control each cell, component or dimension of the weighting so that the sum equals the respective control totals ('guaranteed convergence') (Luery).
3. The GLS procedure is designed to minimize the statistic (Luery)

$$\Sigma_i \; (\omega_i - \Omega_i)^2 \; / \; \Omega_i$$

where Ω_i = the adjusted weight for the i^{th} sample record from the previous stages of estimation

ω_i = the final GLS weight for the i^{th} sample record

C. Application of GLS in the CPS

Each dimension that defines a set of controls in the current second-stage adjustment will define a set of linear constraints for the GLS procedure. For CPS, there are three dimensions: row (state), column (age/sex/ethnicity), and layer (age/sex/race). The problem is to satisfy each set of these constraints simultaneously while minimizing the weighted squared adjustments for the sample records within a rotation group.

The function to be minimized is

$$f(\omega) = (\omega - \Omega)' \; \Omega_0^{-1} \; (\omega - \Omega)$$
$$= \Sigma_i \; (W_{2i} - W_{1i})^2 \; / W_{1i}$$

where ω = (n x 1) vector of derived final weights (W_{2i}) for each of the n sample persons (n is around 14,000 for each rotation group)

Ω = (n x 1) vector of first-stage weights (W_{1i}) for each sample person

Ω_0 = (n x n) diagonal matrix with the W_i on the diagonal

subject to $X'\omega = N$

where X = (n x k) design matrix (in CPS, k is 132) whose columns correspond to control cells. The entries of the matrix (X_{ij}) are 0's or 1's indicating for each of the n sample persons which control categories that person is contained in.

N = (k x 1) vector of independent population counts for the control categories, corresponding to the columns of X. These control counts are the same as those used in the second-stage RRE.

The columns of X are required to be linearly independent so that an inverse of the matrix ($X' \Omega_0 X$) is achieveable. Therefore, in setting up matrices X and N for CPS, the control cells (number of columns in matrix X) used were reduced to a set of k=132 linearly independent cells.

The unique solution to $X'\omega = N$ that minimizes f(ω) is, as given earlier

$$\omega = \Omega + \Omega_0 X (X' \Omega_0 X)^{-1} (N - X'\Omega)$$

5. DATA PRODUCED

A. Estimates

Estimates were tabulated, using the final weights derived from RRE and GLS, for both months for the following characteristics: race x sex x labor force status (civilian labor force, employed, unemployed, employment rate, not in labor force); ethnicity x sex x labor force status; state x race x labor force status; and state x ethnicity x labor force status.

Standard errors and CVs for these estimates were derived using a random group estimator, with the sample rotation groups as random groups.

In addition to tabulating estimates, standard errors, and CVs for both RRE and GLS, the relative difference of these tabulations were also calculated, where the relative difference is defined as

$$(Y_{gls} - Y_{rre}) \ / \ Y_{rre}$$

where Y_{rre} = estimate of Y based on the weights derived through the use of RRE

Y_{gls} = estimate of Y based on the weights derived through the use of GLS

B. Month-in-Sample Indexes

Month in sample indexes, which measure the relative bias between rotation groups, defined as

$$(8Y_k \ / \ Y) \times 100$$

were calculated for both July 1983 and July 1984 based upon both the RRE estimates and the GLS estimates for: labor force status x sex; labor force status x race, and labor force status x ethnicity.

C. Measures of Closeness

Fagan and Greenberg (1985) listed four measures of closeness for evaluating adjustment procedures. These procedures are:

Measure A = $\Sigma_i \ W_{2i} \ln(W_{2i} \ / \ W_{1i})$
(Raking Measure)

Measure B = $\Sigma_i \ -W_{1i} \ln(W_{2i} \ / \ W_{1i})$
(Maximum Likelihood Measure)

Measure C = $\Sigma_i \ (W_{2i} - W_{1i})^2 \ / \ W_{2i}$
(Minimum Chi-Square Measure)

Measure D = $\Sigma_i \ (W_{2i} - W_{1i})^2 \ / \ W_{1i}$
(GLS Measure)

where W_{1i} = weight for sample record i prior to adjustment

W_{2i} = final weight for sample record i following adjustment

Values for each of the four measures were calculated using both the RRE and GLS final weights. The measures were calculated for the total sample in each rotation group for both months, and by race, by ethnicity, and by sex within each rotation group.

D. Distribution of Factors

Sample records in the same control category for one marginal (state, age/sex/ethnicity, or age/sex/race) but in different control categories for another marginal most likely will not receive the same adjustment factor. (This is true for both RRE and GLS.)

For each set of independent population controls, ratios C/E (the inverse of the coverage rate), were C is the independent control and E is the sample estimate based on the first-stage weights, can be derived. These ratios represent the factor which would be applied to the first-stage weights of sample records if ratio adjustment was carried out for only one set of controls and are referred to here as one-way factors.

In order to compare the RRE and GLS procedures relative to the application of the adjustment factors, the following statistic was generated for both RRE and GLS:

$$R_i = (W_{2i} \ / \ W_{1i}) \ / \ (C \ / \ E)_i$$

$$= \frac{\text{factor applied to sample unit i}}{\text{one-way factor for sample unit i}}$$

Histograms showing the distribution of R for RRE and GLS were drawn for the following demographic classes: race x sex; ethnicity x sex.

The factors applied to the sample records by RRE and GLS were also compared by looking at the distribution of the one-way factors C/E relative to the ratio of the factors applied by the two adjustment procedures

$$RRE \ / \ GLS = [(W_{2i} \ / \ W_{1i})_{rre}] \ / \ (W_{2i} \ / \ W_{1i})_{gls}]$$

This ratio indicates the difference between the factors applied to a sample weight by the RRE and GLS procedures. The distribution of each set (state, age/sex/ethnicity, age/sex/race) of one-way factors C/E was charted for the total sample, and for sample records for which the ratio RRE / GLS was less than 0.95 and greater than 1.05.

6. RESULTS

(Space limitations do not permit inclusion of the tabulated data prepared for the meetings. These are available upon request.)

A. Estimates

1. Level of Estimates

Weighted labor force estimates based on the current CPS RRE procedure for July 1983 and July 1984 did not show any noticeable differences or trends when subaggregated to the sex x race/ethnicity and state x race/ethnicity level.

2. Standard Error of Estimates

Estimated standard errors for the CPS RRE and the GLS estimates also did not show any noticeable differences or trends at the sex x race/ethnicity level. At the state x race/ethnicity level differences in the estimated standard errors were noticed; it is not known whether these differences are outside sampling variability.

B. Month-in-Sample Indexes

Month-in-sample indexes for labor force x race, labor force x sex, and labor force x Hispanic characteristics were virtually identical for estimates based upon the CPS RRE and GLS procedures.

C. Measures of Closeness

Tabulation of the measures of closeness provided some interesting and, in some cases, puzzling results. The CPS RRE yielded smaller values for all four measures, even Measure D which GLS is designed to minimize. The CPS RRE which resulted in smaller values of measure tended to produce larger values of that measure for subaggregates of minority populations.

Measure A should be minimized through the use of a convergent RRE, and the value of Measure A for the total sample based upon the CPS RRE weights was less than the value of Measure A based upon the GLS weights for all 16 rotation groups for which adjustment was performed.

Although Measure D should be minimized through the use of the GLS procedure, the value of Measure D based upon the GLS weights for the total sample was greater than the value of Measure D for the CPS RRE weights for 11 of the 16 rotation groups. (It should be noted that the GLS procedure minimizes Measure D among the class of adjustment procedures yielding estimates that meet the population controls. Since the CPS RRE did not converge to the population controls, it is not a member of this class.)

Data for Measures B and C show that the weights based upon CPS RRE yielded smaller values for both measures for all 16 rotation groups.

When comparing the measures resulting from the CPS RRE and GLS procedures, the procedure that yielded the smaller measure of closeness for the total sample tended to yield the larger measure of closeness for subaggregates of minority populations (blacks, Hispanics, males). This was true for all four measures of closeness.

D. Distribution of Factors

There appear to be differences in the factors applied as a result of RRE and GLS for those groups of the population which were over or undercovered by the survey.

For both Hispanic males and females, the factors R based upon the GLS procedure had more observations in the interval centered around 1 and fewer observations in the intervals away from 1 than did the Rs based upon the RRE procedure. This may indicate that the GLS procedure tends to yield second-stage factors closer, as compared to RRE, to those which would have been applied had adjustment been made only to age/sex/ethnicity controls. (Similar results, but of smaller degrees, were obtained for black males and black females.)

The distributions of C/E relative to the ratio of the factors applied by RRE and GLS indicate that, for each set of controls (state, age/sex/ethnicity, age/sex/race), sample records from population groups which were over- or undercovered to some extent were more likely to receive factors more than 5% different as a result of RRE and GLS than was the sample as a whole.

E. Cost

Based upon the computer runs made to perform the weighting for the RRE and GLS procedures, the cost to prepare the files and perform the weighting was approximately three times as much for the GLS procedure than it was for the RRE procedure. There was also more storage of files involved with the GLS procedure, which would tend to add to the cost. (The size of the matrices involved for CPS are quite large, with the number of rows for Ω, Ω_0, X, and N being around 14,000 for each RG.) These cost data were based upon programming done by the authors, and therefore might not be indicative of the relative costs which could result from more efficient programming.

7. SUMMARY AND CONCLUSIONS

This investigation was intended to provide some base information for use in the comparison of RRE and GLS as applied to the CPS, and in future investigation into the use of GLS as an alternative to the numerous weighting procedures currently used for CPS.

The results obtained at the macro level do not indicate any difference in the estimates obtained from the RRE and GLS procedures. The measures of closeness indicated that the CPS RRE made smaller changes to the sample weights to meet the control constraints than did the GLS. The CPS RRE tended to produce larger measures for subaggregates of minority populations. (This result is being further investigated.) Based on the work done in this investigation, it does appear that the RRE is less expensive to run on the CPS second-stage adjustment than is the GLS.

The results obtained at the micro level do indicate that there are differences in the factors applied to the first-stage weights by the two procedures. These differences may possibly be related to the coverage. (This is being further investigated.)

A number of other questions and issues have arisen in conducting this investigation. Among them are:

1. How sensitive is the GLS procedure to changes that are planned or may be made in the future to the CPS population counts and cell definitions?
2. If the GLS procedure were adopted for the CPS, how would collapsing procedures be efficiently incorporated?
3. Can the GLS procedure encompass the use of composited full population estimates as controls for one or more sets of constraints?
4. What are the effects of using biased estimates (resulting from the noninterview and first-stage adjustments) on the GLS procedure (which uses assumptions of unbiased estimates)?
5. Some of the properties of the GLS procedure such as bias and MSE reduction should be further investigated.
6. Can an efficient processing system be developed for the GLS procedure as applied to the CPS to allow monthly production?
7. Algorithms for adjustment procedures that minimize the Maximum Likelihood Measure and the Minimum Chi-Square Measure should be developed and compared against the RRE and GLS procedures.
8. How would a multivariate RRE procedure be applied to the CPS weighting procedures, and how would it compare to the univariate RRE and the GLS?
9. There are currently plans for using replicate weighting to allow variance estimation in the CPS. The issues of variance estimation in conjunction with the use of GLS (i.e., is there an efficient method) should be investigated.

The next step in the investigation into the application of GLS to CPS should extend its use to combining the second-stage adjustment and principal person weighting and/or combining the second-stage adjustment and veterans' weighting. Each application presents different problems which would have to be addressed before GLS could be considered for use in the CPS, such as the need for equal weights within sample housing units for the tabulation of family data; and the need to control population subdomain estimates to composited full population data.

ACKNOWLEDGEMENTS

The authors wish to thank Fritz Scheuren of IRS for his helpful comments in review of this paper and William Winkler of EIA who discussed this paper at the Chicago meetings. The paper was typed by Tawanna Jackson.

REFERENCES

Fagan, J. and Greenberg, B. (1985). "Algorithms for Making Tables Additive: Raking, Maximum Likelihood, and Minimum Chi-Square." SRD Report Series, Census/SRD/RR-85/15, Bureau of the Census.

Gbur, E. (1984). "The Current Population Survey Family Weighting Procedure and Some Alternatives," Draft Report. U.S. Bureau of the Census.

Hanson, R.H. (1978). The Current Population Survey Design and Methodology, Technical Paper 40, U.S. Bureau of the Census.

Ireland, C.T., and Kullback, S. (1968). "Contingency Tables With Given Marginals." Biometrika 55, 179-188.

Luery, D. (1980). "An Alternative to Principal Person Weighting." Draft Internal Memorandum. U.S. Bureau of the Census.

Neyman, J. (1949). "Contribution to the Theory of the X^2 Test". Proceedings 1^{st} Berkeley Symposium.

Oh, H.L., and Scheuren, F. (1978a). "Some Unresolved Application Issues in Raking Ratio Estimation". Proceedings of the American Statistical Association. Section on Survey Research Methods.

(1978b). "Multivariate Raking Ratio Estimation in the 1973 Exact Match Study." Proceedings of the American Statistical Association. Section on Survey Research Methods.

Scheuren, F. (1981). "Methods of Estimation for the 1983 Exact Match Study." SSA Publication No. 13-11750 10-24.

Zieschang, K.D. (1985). "Consumer Expenditure Survey: GLS Weighting Study." Draft Report. Bureau of Labor Statistics.

(1986). "Generalized Least Squares: An Alternative to Principal Person Weighting." Presented at the BLS Conference on Population Controls in Weighting Sample Units.