ITERATION OF SECOND-STAGE AND COMPOSITE PROCEDURES IN THE CURRENT POPULATION SURVEY December 2006

Samantha V. Cruz, Edwin L. Robison, and Tamara Sue Zimmerman, Bureau of Labor Statistics
Samantha V. Cruz, BLS, 2 Massachusetts Avenue NE, Washington, DC 20212
Cruz.Samantha@bls.gov

KEY WORDS: iterative proportional fitting, raking, weighting

ABSTRACT
Labor force estimates of the U.S. civilian noninstitutional population (CNP) are derived through a number of weighting steps in the Current Population Survey (CPS) estimation procedure. Currently, the weight for each interviewed person includes a second-stage ratio adjustment. This second-stage estimation process uses Iterative Proportional Fitting (IPF), or raking, in order to simultaneously match three sets of independent population controls and to create second-stage weights. Upon completion of the second-stage process, the weights are subjected to a composite estimation process which also uses a three-way rake to match composite population controls. This paper explores more complex raking methods to determine if a single set of weights can be produced that simultaneously preserves all second-stage and composite controls.

I. THE CPS SAMPLE

The CPS is a multistage probability sample of about 60,000 eligible households. It is jointly sponsored by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), and is the Nation's main source of labor force statistics for the entire CNP population. Economic statistics such as the Nation's unemployment rate and employment and earnings data are released on a monthly basis.

An independent sample is selected in each of the 50 states and the District of Columbia. This monthly sample is split into eight panels, or rotation groups of households. Each rotation group is itself a representative sample of the U.S. population.

A housing unit in a rotation group is interviewed for four consecutive months, out for eight, and then interviewed for another four months before exiting the sample permanently. In a given month one panel each is being interviewed for the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th time, or "month-in sample". This rotation sampling scheme (4-8-4) ensures a 75% month-to-month sample overlap. The month-to-month overlap induces positive correlation among CPS estimates for different months. This correlation is exploited via composite estimation to improve monthly estimates and to improve estimates of month-to-month change.

II. OVERVIEW OF WEIGHTING PROCEDURES

In order to produce national and state labor force estimates, a weight for each person is created through a series of weighting steps:

- Preparation of simple, unbiased base weights (and some special weights) which are the inverses of household sampling probabilities
- Household nonresponse adjustment
- First-stage ratio adjustment (reduces variance due to PSU sampling)
- National and State Coverage adjustments
- Second-stage weighting procedure
- Composite weighting procedure

Both the second-stage weighting procedure and the composite weighting procedure use raking or Iterative Proportional Fitting (IPF) methods that force weighted estimates to match specified control totals (Deming and Stephen, 1940). The second-stage procedure uses CNP controls defined for demographic groups. The composite procedure uses a subset of labor force estimates made using a composite estimation formula (see Composite Weighting Procedure section).

Composite weighting partially unravels the second-stage weighting process. Thus, most of the second-stage CNP controls are no longer matched when composite weights when used. While published estimates are derived from the composite weights, for some analyses second-stage weights may be preferable. This is particularly true for employment estimates within second-stage cells, where month-to-month changes will be more stable than estimates

using weights that are developed after combining cells in the composite procedure.

SECOND-STAGE WEIGHTING PROCEDURE

The second-stage (SS) weighting procedure reduces variances by controlling population estimates to independent estimates of the current population. The procedure also reduces bias due to coverage errors.

SS weighting is an IPF procedure that consists of three steps: a state step, an ethnicity step, and a race step. At each step the estimates are raked and forced to match a set of civilian noninstitutional population controls that are derived externally to the CPS (Technical Paper 63, 2000). The SS CNP population controls are prepared by projecting forward the population figures from the most recent decennial census. While the controls are actually estimates, they are derived independently of the CPS and provide information for adjusting sample estimates.

For each state, ethnicity, and race step of the SS weighting procedure, there is a matching set of independent SS population controls:

1. CNP for 50 states, substates, and DC by sex and age (0-15, 16-44, 45+)

2. National CNP for 26 Hispanic and 26 Non-Hispanic age-sex categories

3. National CNP for 56 White, 36 Black, and 32 "residual race" age-sex categories.

At each iteration of each raking step, an adjustment factor for each cell is computed and applied to the weights of persons in the cell. The adjustment factor for a cell is its population control divided by the weighted cell estimate based on all prior weight adjustments. After each raking step, weighted CPS estimates exactly match the cell population controls, but controls for the cells of previous raking steps no longer match. But with each rake, these differences decrease. After 10 iterations of the three-way (state/sex/age, ethnicity/sex/age, race/sex/age/) rake, the estimates of each cell have converged to the population controls for each cell. Although an estimate of level for any characteristic in the CPS can be computed by summing the SS weights for all sample cases that have that particular set of characteristics, official CPS labor force estimates are based on the weights produced in the composite weighting procedure.

COMPOSITE WEIGHTING PROCEDURE

After a SS weight is computed for each record, adult person weights are further adjusted using a composite weighting procedure. Composite (Cmp) weighting utilizes information collected monthly from the full sample, and is performed only for sample persons aged 16 and above. The composite weighting mimics the SS weighting, where sample person weights are raked to force their sums to equal control totals. Instead of independent population controls however, composite CPS labor force estimates are used as controls and are computed using a composite estimation formula. The CPS composite estimator for a labor force total in month t is given by

$$Y_t' = (1-K)\hat{Y}_t + K(Y'_{t-1} + \Delta_t) + A\hat{b}_t$$

where

$$\hat{Y}_t = \sum_{i=1}^{8} x_{t,i}$$

$$\Delta_t = \frac{4}{3}\sum_{i \in S}(x_{t,i} - x_{t-1,i-1}) \text{ and}$$

$$\hat{b}_t = \sum_{i \notin S} x_{t,i} - \frac{1}{3}\sum_{i \in S} x_{t,i}$$

i = 1,2,…,8 month in sample
$x_{t,i}$ = sum of weights after second-stage ratio adjustment of respondents in month t, and month-in-sample i with characteristic of interest
S = {2,3,4,6,7,8} sample continuing from previous month
K = 0.4 for unemployed
0.7 for employed
A = 0.3 for unemployed
0.4 for employed

Like the SS weighting procedure, there are three steps (state, ethnicity, and race) in the composite weighting procedure, with three corresponding composite population controls:

1. State (by sex-age categories)

2. Ethnicity (Hispanic/Non-Hispanic, by sex-age categories)

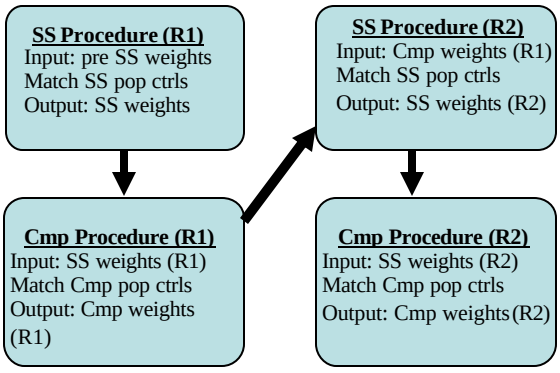
3. Race (Black/White/Asian & Res, by sex-age categories).

After 10 iterations of the three-way rake in the composite weighting, summed sample person composite weights do not match all of the SS independent population controls. Composite weighting matches SS CNP controls when the cells in the two procedures use the same demographic definitions (further split by labor force). This occurs for some cells in the ethnicity and race steps. But, more commonly, second-stage age groups are combined for composite weighting. When composite weights are summed for the second-stage cells, most CNP controls are missed by 1000 or more (Robison et al. 2002). The Proposed Methodology section details how a more complex raking schema can be implemented in order to simultaneously preserve both the second-stage and composite population controls.

III. PROPOSED METHODOLOGY

Conserving all the second-stage population controls can be achieved with a more complex iterative process of raking the second-stage and composite weighting steps. Figure 1 shows a straightforward flowchart of the process for the first iteration of SS-Cmp cycle. In order to differentiate between the within-weighting step iteration and the SS-Cmp cycle iteration, the term "round", or "R", will be used to define one cycle of the SS-Cmp iteration.

Figure 1. Flowchart of Iterated SS and Cmp Complex Weighting Methodology



The following steps explicitly describe the method:

1. Run SS estimation to obtain SS weights.

2. Use the SS weights as inputs to the Cmp procedure, obtaining Cmp weights.

3. Use the Cmp weights as inputs into the SS weighting procedure, obtaining SS weights (R2).

4. Use the SS weights (R2) as inputs to the Cmp weighting procedure, obtaining R2 Cmp weights.

5. Iterate steps 3 and 4 and check to see if the summed Cmp weights converge to SS population controls.

The complex raking method described above only iterates the SS-Cmp cycle and doesn't take into account other steps of the entire weighting procedure described in section II. Two months of data were chosen to test the methodology; 20 rounds (R1-R20) were completed for July 2005 and 10 rounds (R1-R10) for August 2005. Data were provided by the U.S. Census Bureau.

IV. ANALYSES AND RESULTS

The results are broken out by the three steps in both the second-stage weighting and composite weighting procedures: state, ethnicity and race. First we focus primarily on the July 2005 data. To determine convergence, we use the difference:

$$Diff_i = SSctrl - totalR_i$$

where

SSctrl = second-stage population control,
totalR = summed composite weights for a particular step's characteristics,
i = round number

For information on convergence issues for IPF see Rüschendorf's 1995 paper.

Table 1 shows the results for the state step for males ages 16-44. Five states were randomly chosen: Alaska (AK), Georgia (GA), Louisiana (LA), Pennsylvania (PA), and Tennessee (TN).

Although only the first five rounds are shown in table 1, for the most part the difference between the SS population controls and the summed Cmp weights for each state for males 16-44 decrease in magnitude as each subsequent round of the SS-Cmp cycle is completed. The largest difference is found between the first and second round; this round one difference is the discrepancy that is already present in the current CPS weighting

Table 1. Results for the State Step for Males 16-44 (7-2005)

	Diff1	Diff2	Diff3	Diff4	Diff5
AK	-1282.61	-122.92	-18.22	-6.30	-4.29
GA	-7706.21	-1343.25	-318.53	-185.34	-154.76
LA	2266.37	-371.10	-458.28	-351.25	-263.67
PA	1318.94	195.16	24.16	8.93	7.12
TN	7039.49	738.36	34.08	-51.80	-55.52

procedures. The other states and substates also show similar results to the states shown in table 1.

Figure 2. Results for the Ethnicity Step for Non-Hispanic Males (7-2005)

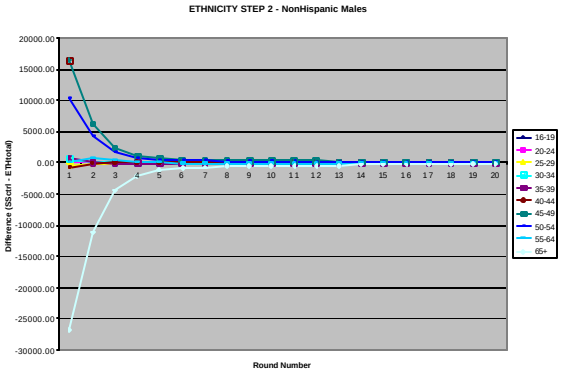


Figure 2 is a graphical representation of results similar to those found in table 1, but for the ethnicity step for Non-Hispanic males. On the x-axis are the completed rounds (up to R20); on the y-axis is the difference between the SS population control and the summed Cmp weights. Each line represents a different age category represented in the composite weighting procedure. It is obvious that as more and more rounds are completed, the summed Cmp weights virtually converge to the corresponding SS population controls; this occurs even before the 20th round is completed.

Figure 3. Results for the Ethnicity Step for Hispanic Females (7-2005)

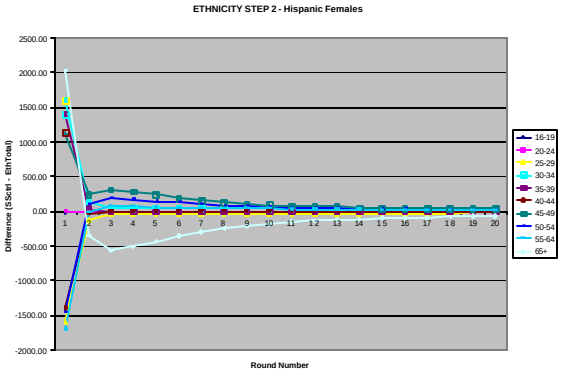


Figure 3 shows the results for the ethnicity step for Hispanic females for July 2005. There is a large initial drop in difference between the SS population control and the summed Cmp weight. Although rounds two and three look show an increase, after round three the discrepancy between the controls and the weights starts to decrease, leading to convergence.

Figure 4. Results for the Race Step for Black Females (7-2005)

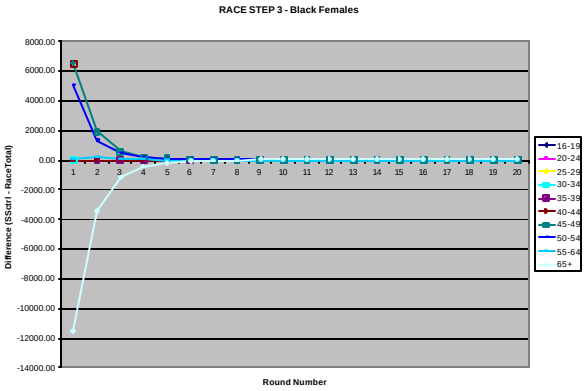
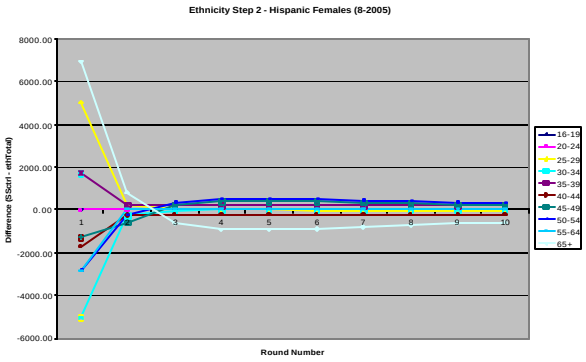


Figure 4 illustrates convergence for the race step for Black females. 20 rounds were completed, though near convergence is met close to rounds five and six. These results are similar to those found in figure 2.

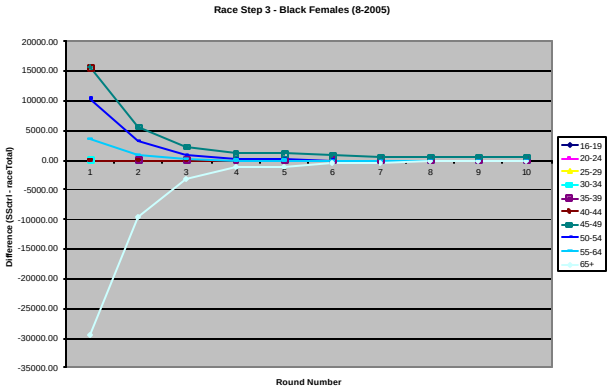
July 2005 data seem to show convergence before the 20th round is completed; thus, only ten rounds were completed for the August 2005 data. The results for the August data are similar to those found in July. Figure 5 shows the results for the ethnicity step for Hispanic females for August 2005. Similar to the corresponding July data, there is a large initial difference, an increase between rounds two and three, and then a slow decline towards convergence towards round ten.

Figure 5. Results for the Ethnicity Step for Hispanic Females (8-2005)



Lastly, figure 6 shows the first ten rounds for the race step, for Black females for August. These results are identical to the results obtained for the corresponding July data: by the tenth round, the summed Cmp weights virtually match the SS independent population controls.

Figure 6. Results for the Race Step for Black Females (8-2005)



V. SUMMARY

Inherent to each of the second-stage and composite weighting procedures are specific advantages dealing with variance and bias issues. Although the composite weighting partially unravels second-stage weighting, there is no reason why a complex raking algorithm cannot be implemented to simultaneously preserve both sets of population controls. Remembering that the largest drop in differences occurs from the first round (already present discrepancy) to the second round for all three demographic steps, the CPS labor force estimates would benefit from even one SS-Cmp cycle beyond what is presently done. And because CPS labor force estimates are published on the order of thousands, these differences would be virtually negligible.

Not only does this research suggests that this idea to be feasible for the CPS estimation design, but also shows a preliminary glimpse of what this algorithm could look like.

VI. FUTURE WORK

Future efforts include examining whether consistent results are found using alternate months of data, preferably not consecutive months. Also, it would be interesting to investigate variances of the estimates, and how they would be affected if a complex raking method were implemented. Further, we would like to extend this research into incorporating the State and National coverage steps in the complex raking method. A combination of the coverage steps and the second-stage and composite weighting steps could possibly lead to faster convergence.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

REFERENCES

Bureau of Labor Statistics and U.S. Census Bureau, *Current Population Survey: Design and Methodology, Technical Paper 66* (unpublished, 2006).

Bureau of Labor Statistics and U.S. Census Bureau (2000), *Current Population Survey: Design and Methodology, Technical Paper 63* (www.bls.census.gov/cps/tp/tp63.htm).

Deming, W.E. and Stephan, F. F. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *Ann. Math. Statist*. Vol. 11, 427-444.

Lent, Janice, S. M. Miller, P. J. Cantwell, and M. Duff (1999). "Effects of Composite Weights on Some Estimates from the Current Population Survey," *Journal of Official Statistics*, Vol. 15, No. 3, pp.431-448.

Robison, Edwin, M. Duff, B. Schneider, and H. Shoemaker (2002). "Redesign of Current Population Survey Composite Weighting." *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 2924-2929.

Rüschendorf, L. (1995). "Convergence of the Iterative Proportional Fitting Procedure." *The Annals of Statistics*, Vol. 23, No. 4, 1160-1174.