REDESIGN OF CURRENT POPULATION SURVEY RAKING TO CONTROL TOTALS

Edwin L. Robison, Martha Duff, and Brandon Schneider, Bureau of Labor Statistics
Harland Shoemaker, Bureau of the Census
Brandon Schneider, BLS-OEUS-SMD Room 4985, 2 Massachusetts Avenue NE, Washington, DC 20212

Key Words: Raking, Iterative Proportional Fitting, Weighting, Current Population Survey

Abstract

Weighting for the monthly Current Population Survey includes an iterative proportional fitting procedure ("second-stage ratio adjustment" or "raking") that forces CPS estimates to match three sets of population control totals. The procedure now in place successively computes ratio adjustments to weights so that estimates are forced to match three sets of controls representing the Civilian Noninstitutional Population of the United States: 1) state controls; 2) national ethnicity x gender x age controls; and 3) national race x gender x age controls. The monthly control totals are treated as fixed constants, and are derived from models that update decennial demographic census totals using administrative data sources. A new process to be implemented in January 2003 retains the three-way iteration, but is designed based on a reevaluation of data analysis needs, convergence properties, and survey coverage. Fixed cells replace an on-the-fly collapsing technique, six gender x age controls are used for each state in the first step of the iterative procedure, Asian is added as a new race, and age categories are harmonized between successive steps of the iterative procedure to improve convergence. Two non-iterated coverage adjustment steps, national and state, were added. The new steps better account for known national interactions between ethnicity and race coverage and known differences in race coverage among the states.

Background

The Current Population Survey is jointly sponsored by the Bureau of Labor Statistics and the Bureau of the Census. The BLS/Census CPS Weighting Group was formed in October 1999 to address weighting issues. CPS weighting includes modules for noninterview adjustment, first-stage ratio adjustment, second-stage ratio adjustment, and benchmarking to composite estimates. Of particular concern were revisions to the weighting needed because of changes in the race and ethnicity questions that will be implemented for the first time in the January 2003 Current Population Survey. We felt this provided an opportunity to look for statistical improvements we could make in the weighting, and at the same time simplify software development and maintenance. Improvement efforts concentrated on second-stage weighting, since that procedure is most affected by changes in the race and ethnicity questions, and since second-stage adjustment dominates the other adjustments in the CPS weighting process. (Modifying the composite weighting procedure, immediately following second-stage weighting in the weighting process, was also a high priority. See *Redesign of the Current Population Survey Composite Weighting* by the same authors as this paper.)

In determining what changes to make to second-stage weighting, a number of factors were considered, including:

- BLS plans for publishing revised race categories at the state and national level (Asian added to core data releases)
- Making control cell definitions more consistent with composite weighting and more consistent across the second-stage weighting steps (state, ethnicity, and race)
- Pre-collapsing small cells to eliminate the need of the current "on-the-fly" collapsing algorithm that produces inconsistent results over time
- Providing more stable monthly estimates for population subgroups of interest to users (In particular, there was a request for demographic population controls within each state.)
- Confidence in the population controls for various age, race, ethnic, and geographic categories
- Possible changes in race reporting patterns over time
- Simplifying development and maintenance of the weighting software

Description of Current Second-Stage Weighting

The Current Population Survey is a rotating panel survey, obtaining responses from about 50,000 households each month. The primary product of the CPS is labor force data for the Civilian Noninstitutional Population. A given monthly sample is divided into eight panels or rotation groups of households. There is a scheme of panel replacement for the next month where one panel is permanently dropped and replaced by a new panel, and one panel is temporarily dropped for eight months and replaced by a returning panel. In adjacent months, six panels are in common. In a given month one panel each is being interviewed for the 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}, and 8^{th} time ("month-in-sample"). There are known biases in labor force data associated with the month-in-sample.

Second-stage estimation in the current CPS weighting scheme solely consists of an iterative proportional fitting or raking procedure with three steps: a state step, an ethnicity step, and a race step. The current methodology lacks any adjustment to account for interactions between state, ethnicity, and race. Each step has population control totals that are estimates of CNP/8, one-eighth of the Civilian Noninstitutional Population. The population controls are essentially 1990 decennial Census estimates, updated to the current time using models and a variety of data sources. (Adjustments for census undercount are included.) A divisor of eight is used since each of the eight monthly panels or rotation groups is processed separately. The ethnicity and race steps account for juveniles as well as adults, but the state step is limited to adults 16 years of age and over (CNP16+).

1. The current state step categorizes the CPS sample observations into 51 cells: a single cell for each state and the District of Columbia. Each cell is controlled to its CNP16+ (each panel to CNP16+/8).

2. Following the state step, the current ethnicity step categorizes the same observations into 19 cells: 14 Hispanic gender x age cells and five non-Hispanic age cells, gender combined. Each cell is controlled to its CNP (each panel to CNP/8).
   - Hispanic age categories: 0-5, 6-13, 14, 15, 16-19, 20-29, 30-49, and 50+ (14 and 15 combine gender)
   - Non-Hispanic age categories: 0-5, 6-13, 14, 15, and 16+

3. The last step of the iterative procedure, the race step, categorizes the observations into 118 cells: 42 Black gender x age cells, 66 White gender x age cells, and 10 Other gender x age cells. Each cell is controlled to its CNP (each panel to CNP/8).
   - Black age categories: 0-1, 2-3, 4-5, 6-7, 8-9, 10-11, 12-13, 14, 15, 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, and 65+
   - White age categories: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10-11, 12-13, 14, 15, 16, 17, 18, 19, 20-24, 25-26, 27-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-67, 68-69, 70-74, and 75+
   - Other age categories: 0-5, 6-13, 14, 15, 16-44, and 45+ (14 and 15 combine gender)

The term "Other" for race is a residual category that includes American Indian, Aleut, Eskimo, Asian, and Pacific Islander. We cannot now distinguish Asians from Native Hawaiians and Other Pacific Islanders, but this will be possible when new race data collection procedures are implemented January 2003.

The steps are iterated separately – six times – for each of the eight CPS panels. For each cell k of each step, a simple adjustment is computed for each of the eight panels using adjusted observation weights w_{kij} from the previous step. The multiplicitive panel adjustments are then applied to the weight of every observation in the cell. The adjusted observation weights w_{kij} ' are used in the next step of the iterative procedure.

$$ADJ_{ki} = (CNPk/8)/ \sum_j w_{kij}$$
$$w_{kij}' = ADJ_{ki} * w_{kij}$$

The iterations are needed since the execution of each step slightly imbalances the previous steps. After the race step, for example, weighted CPS estimates of population for a given rotation group no longer match the population control by ethnicity/gender/age or by state. After cycling through the process six times, CPS estimates of population for each rotation group nearly match all three sets of controls. That is, the iterative raking process converges to the three sets of population controls. (Also, if you use all eight panels, CPS estimates of population nearly match the desired Civilian Noninstitutional Populations.)

Cell Collapsing -- Cells with few respondents have the potential of having large weight adjustments. The current second-stage weighting procedure has a collapsing algorithm. For each age cell in the ethnicity and race steps, a preliminary adjustment factor is computed for each rotation group. If the factor lies outside of the acceptable range 0.6-2.0 (or if the cell has no respondents), then the cell is collapsed with one or more adjacent age cells in its rotation group. In a typical month about 10 cells require collapsing, and different panels generally will not have the same collapsing. The collapsed cells are not always the same cells from one month to the next. There is no collapsing for the state step.

Redesigned Second-Stage Weighting Procedure

Second-stage estimation in the redesigned CPS weighting scheme, to be implemented in January 2003, includes two new preliminary steps, a national-level coverage step and a state-level coverage step, that are followed by an iterative raking procedure similar to the current methodology. The national-level coverage step was designed to account for the interaction between ethnicity and race, and the state-level coverage step was designed to account for differences in state race coverage relative to national coverage. In the state-level coverage step and the state step of the iterative procedure California and New York are split into substate areas: Los Angeles-Long Beach metropolitan

area and the balance of California, and New York City and the balance of New York, respectively. In the national-level coverage step and the race step of the iteration procedure Asian is introduced as a new race. Each adjustment of the redesigned procedure consists of a fixed number of cells – collapsing is eliminated. Each step, excluding specific cells of the state coverage step, has population control totals that are estimates of CNP/4, one-fourth of the Civilian Noninstitutional Population. The eight monthly panels are paired to increase cell counts, allowing more demographic detail. The population controls are 2000 decennial Census estimates, updated to the current time using models and a variety of data sources.

Specifications follow for cells of the national-level coverage step (A), the state-level coverage step (B), and the three steps (1-3) of the redesigned iterative procedure.

A. The non-iterated national-level coverage step, adjusting for subpopulations prone to under/over coverage, categorizes the CPS sample observations into 126 cells: 26 Hispanic White gender x age cells, four Hispanic non-White gender x age cells, 18 non-Hispanic Asian gender x age cells, 26 non-Hispanic Black gender x age cells, 34 non-Hispanic White gender x age cells, and 18 non-Hispanic Residual gender x age cells. The Hispanic White and non-Hispanic Black cells have identical age breaks, and the non-Hispanic Asian and non-Hispanic Residual cells have identical age breaks.
   - Hispanic White and non-Hispanic Black age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, and 65+
   - Hispanic non-White age categories: 0-15, and 16+
   - Non-Hispanic Asian and non-Hispanic Residual age categories: 0-4, 5-9, 10-15, 16-24, 25-34, 35-44, 45-54, 55-64, and 65+
   - Non-Hispanic White age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-69, 70-74, and 75+

B. The non-iterated state-level coverage step allows more detailed controlling for states/substates with larger numbers of persons of Black race in the sample. All 515 defined cells are controlled to CNP. In some cells the panels are paired and controlled to CNP/4, but for others the controlling is only to CNP with all eight panels combined.
   - Non-Black – In all states/substates six gender x age cells (0-15, 16-44, 45+) are defined. Each of the 318 cells is controlled to its CNP. Except for the District of Columbia, each panel is controlled to its CNP/4.
   - Black – In 26 states/substates six gender x age cells (0-15, 16-44, 45+) are defined. In 12 of these states/substates the panels are paired and each of the 72 cells is controlled to its CNP (each panel pair to CNP/4): New York City, FL, GA, IL, MI, MS, NJ, NC, OH, PA, TX, and the District of Columbia. In the remaining 14 states/substates the eight panels are combined and each of the 84 cells is controlled to its CNP: Los Angeles-Long Beach metropolitan area, the balance of California, the balance of New York, AL, AR, CT, DE, LA, MD, MA, MO, SC, TN, and VA.
   - Black – In 14 of the states with smaller Black race populations two gender cells are defined, age combined. Panels are combined and each of the 28 cells is controlled to its CNP: AK, AZ, CO, KY, OK, IN, KS, MN, NE, NV, RI, WA, WV, and WI.
   - Black – For the remaining 13 states, those with the smallest Black race population, one cell is defined, gender and age combined. Panels are combined and each of the 13 cells is controlled to its CNP: HI, IA, ID, ME, MT, NH, NM, ND, OR, SD, UT, VT, and WY.

1. The first step of the iterative procedure, the state step, categorizes the observations into 318 cells: six gender x age cells for Los Angeles-Long Beach metropolitan area, the balance of California, New York City, the balance of New York, each of the remaining 48 states and the District of Columbia. Each cell is controlled to its CNP (each panel pair to CNP/4).
   - Age categories: 0-15, 16-44, and 45+

2. Following the state step, the ethnicity step categorizes the same observations into 52 cells: 26 Hispanic gender x age cells and 26 non-Hispanic gender x age cells. The Hispanic and non-Hispanic cells have identical age breaks as the Hispanic White and non-Hispanic Black cells in the national-level coverage step. Each cell is controlled to its CNP (each panel pair to CNP/4).
   - Hispanic and non-Hispanic age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, and 65+

3. The last step of the iterative procedure, the race step, categorizes the observations into 86 cells: 26 Black gender x age cells, 34 White gender x age cells, and 26 Asian and Residual combined gender x age cells. The Black, and Asian and Residual combined cells have identical age breaks as the ethnicity cells. The White cells have identical age breaks as the non-Hispanic White cells of the national-level coverage step. Each cell is controlled to its CNP (each panel pair to CNP/4).
    - Black, and Asian and Residual combined age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, and 65+
    - White age categories: 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-62, 63-64, 65-69, 70-74, and 75+

There will be ten iterations of steps 1-3, although convergence to the unit usually takes only six to eight iterations. As in the current methodology, the iteration is needed since the execution of each step slightly imbalances the previous steps. Changing the order of the steps in the iterative procedure has no effect on the final results since the set of equations tend to have a unique solution.

The CPS sample contains enough observations to eliminate the cell collapsing that is necessary in the current methodology. The predefined cells for each step in the redesigned second-stage weighting procedure rarely fall below a frequency count of 80 (20 per panel pair) – the criteria set in the initial redesign phases.

The new race/ethnicity data collection that will be implemented January 2003 allows multi-race reporting. Also, it will become possible to distinguish Asians from Native Hawaiians and Other Pacific Islanders. White, Black, and Asian cells in the specification exclude multi-race reporters. The Residual race includes observations not categorized strictly as Asian, Black, or White: Native Hawaiian and Other Pacific Islander; American Indian, Aleut, and Eskimo; and Multi-Race. [To develop and test the new weighting procedures, data files with the current race categories were used: White, Black, Asian and Pacific Islander, and a residual that is American Indian, Aleut, and Eskimo. Note that this residual based on current race is quite a bit smaller than the new residual will be since it does not have Multi-Race or Native Hawaiian and Other Pacific Islander.]

Cell definitions are consistent with the needs of the composite weighting procedure that come after second-stage weighting. In particular, age breaks are defined consistently between the two procedures in order to minimize the extent to which composite weighting "undoes" the second-stage population controlling.

Observations on the Second-Stage Redesign

More Cell Detail – Compared to the current procedure, there is more cell detail in second-stage weighting. Substate areas in California and New York are used. Many state controls are used (instead of just a single CNP16+ control) in order to improve analysis of month-to-month changes of labor force for state demographic groups. At the national level, some cells are defined for Asian race. In general, finer age breaks are used at the national level than for the current procedure. (There was actually a reduction in national White and Black age detail, particularly among juveniles. This was mainly in response to unwanted month-to-month data movements for age cohorts caused by fluctuations in population controls for narrowly-defined age ranges.)

Panel Pairing -- Without pairing, the national demographic detail would have to be reduced (particularly for Hispanic, Asian and Black) and a split by race would be possible in only a few states. (The current procedure rakes each panel separately.) As a result of pairing, small increases in variances on topside estimates are possible. All panels cannot be collapsed in the iterative process, since composite weighting follows second stage weighting. The structure of our composite estimator requires that the incoming panels (rotation groups that are in months-in-sample 1 and 5) be kept separate from the others because of known month-in-sample biases.

No Cell Collapsing -- Software development and maintenance are simplified by eliminating an entire complex process. Cells in the coverage steps and iterative procedure are "pre-collapsed" and the large majority have more than 120 observations (30 per panel pair). Testing showed that variances started to appreciably increase when many cell sizes fell below 120 observations. Should a cell happen to have zero observations, no weight adjustment calculation is made for that cell. If this were to occur it would most likely be seen in Black cells of the state coverage step. This is only a possibility in states with the smallest samples and the lowest Black populations. A cell monitoring system is being developed to detect when survey conditions change and cells grow too small.

Coverage and Interactions – All of the variables used to define cells (state, ethnicity, race, gender, and age) are known to influence coverage. There are certainly

interactions between the variables. A straightforward iterative procedure lets each step act independently. It cannot properly take into account more complex interactions between variables without crossing them. In practical terms, with an iterative procedure like ours it is important for a first pass to get things right and handle the interactions in an acceptable way. If the first pass introduces problems it is risky, since there is no guarantee that later iterations will remedy the problems. No attempt is made at this time to address other coverage problems (for example: urban versus rural; state by ethnicity).

Convergence -- Consistently defined age breaks are, more than any other factor, the key to fast convergence in the iterative procedure. For the redesign, White in the race step has 17 age breaks, and all other defined age breaks are logical collapsings of those 17. The "magic number" is the 13 age breaks used in the ethnicity step and in the race step for Black and Residual. We would have liked to further split out Asian in the race step -- but neither Asian nor the very small residual could support 13 age breaks -- and in testing convergence slowed to a crawl. Incidentally, convergence is now assured to 26 national gender-by-age population totals. Inconsistencies slow convergence for the current procedure, and in six iterations several population controls are missed by hundreds. Due to inconsistencies, only 10 national gender-by-age population controls are matched. The coverage steps also affect convergence – the national coverage step speeds convergence, but the state coverage step slightly slows convergence.

Iterative Nature of the Second-Stage Weighting Procedure – The iterative nature of the proposed second-stage and composite weighting procedures is similar to the present procedures. Iterative proportional fitting provides larger cells and allows us to match many more control totals than we could match with a non-iterated system that crossed all variables (geography x ethnicity x race x gender x age). The ethnicity and race steps tie into important core tabulations and the state step ties into the important Local Area Unemployment Statistics program. Matching control totals in general 1) lowers variances somewhat on current month estimates and 2) substantially lowers variances on important month-to-month comparisons. Similar benefits can be obtained for other subpopulations even when iteration is not possible, as shown by the analysis in the next section.

Box Plot Analysis of the Coverage Steps

National-Level Coverage Step – This non-iterated step helps correct for interactions between race and ethnicity coverage that proved impossible to address in our iterative procedure. For example, research discovered gross undercoverage of Non-Black Hispanics that can be corrected for in this step but not in the iterative steps. Without the national-level coverage step non-Hispanic Asians (shown in Figure 1 below), non-Hispanic Blacks, non-Hispanic Residuals, and Hispanic Whites tend to be overestimated at the end of the second-stage iterative procedure; whereas, non-Hispanic Whites, and Hispanic Asians, Hispanic Blacks, and Hispanic Residuals tend to be underestimated. The step has 18 controls for non-Hispanic Asians (gender by 9 age breaks; over 95 percent of all Asians included). Although Asian controls were unworkable in the iterative race step, a reasonable degree of control is made possible by the coverage step.

Figure 1 visually illustrates the improvement for non-Hispanic Asians using July 2000 data. The box plots for the 18 non-Hispanic Asian controls are scaled for easy viewing:

- O is the ideal when the estimate equals the control
- Positive values (1 - control/estimate) are shown if the estimate is greater than the control
- Negative values (estimate/control - 1) are shown if the estimate is less than the control

The box plots are for:

- First -- controls compared to first-stage estimates (the starting point for second-stage weighting)
- Null -- controls compared to estimates from iteration, but without any coverage steps
- A -- controls compared to estimates from iteration, but only with the national-level coverage step
- AB -- controls compared to estimates from the redesigned second-stage procedure
- B – controls compared to estimates from iteration, but only with the state-level coverage step

It is clear that the iterative process alone (Null) improves non-Hispanic Asian coverage, but including the national-level coverage step (A) draws the estimates closer to the controls and reduces the spread. The state-level coverage step has little effect on the non-hispanic Asian estimates. The same general picture emerges every month for all ethnicity/race subpopulation that are specifically used in the national-level coverage step.

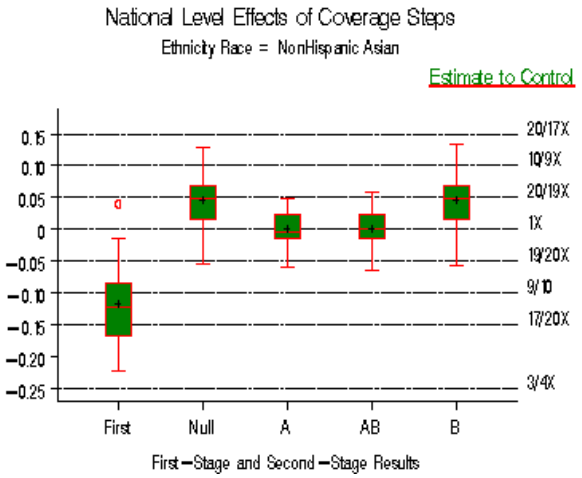

Figure 1.

State-Level Coverage Step – This non-iterated step adjusts for state differences in gender/age/race coverage. It proved impossible to include race in an iterated state step.

Figure 2 visually illustrates the value of the state-level coverage step using July 2000 data. The box plots compare estimates to controls for 197 Black cells. Without a state-level coverage adjustment, some estimates are quite far off from the controls. With the step, almost all estimates are within 5 percent of the controls at the end of the second-stage weighting procedure.

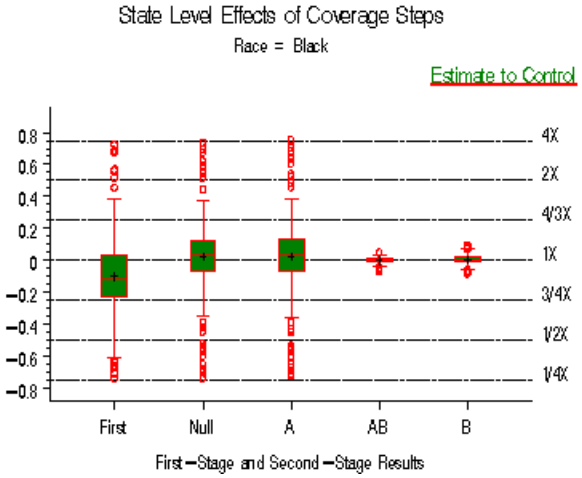

Figure 2.

Disclaimer

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

References:

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)

Robison, Edwin (2001). "Proposal for Redesigned CPS Weighting, Second-Stage and Composite Procedures. "