Use of Administrative Data in SIPP Longitudinal Estimation

Vicki J. Huggins and Robert E. Fay, U.S. Bureau of the Census[1]
Vicki J. Huggins, Stat. Meth. Div., U.S. Bur. of the Cen., Wash., DC 20233

Keywords: raking estimation, survey estimation, variance estimation, weighting

1. Introduction

Ratio estimation, a basic strategy to improve the precision of survey estimates, has been the initial inspiration for a variety of more complex estimators employed in current practice. The demographic surveys conducted by the Census Bureau, including the Survey of Income and Program Participation (SIPP), use population controls in their weighting procedures in some form of ratio or raking ratio estimation. Although relatively complex estimators are employed for these surveys, the control totals to which the weighted survey estimates have been adjusted generally have been confined to a relatively restricted set of characteristics, typically derived by updating information from the previous decennial census.

This paper reports initial efforts to evaluate the merits of incorporating other information in the estimation. Specifically, the application concerns the use of data from Internal Revenue Service (IRS) individual income tax files in the estimation procedures for the SIPP, a survey primarily focused on income and related characteristics of persons and families.

The current controls used in SIPP longitudinal estimation represent a cross-classification of age, race, sex and householder/nonhouseholder status. A larger survey, the Current Population Survey (CPS), provides these controls. Since the CPS itself has weighting controls based on post-censal estimates of age, race, and sex, the current estimates from the SIPP may also be characterized as controlled to demographic estimates by age, race, and sex. Use of controls in SIPP estimation reduces the mean square error for many characteristics, primarily by reducing the sampling variance of the estimator, although arguably by reducing bias as well. The purpose of our research is to determine whether additional adjustment to controls derived from administrative income data could further significantly improve SIPP longitudinal estimates of income and program participation.

As we will comment in the concluding section, several difficulties require resolution before such a methodology can be implemented as an official source for SIPP estimates. Among the problems are inconsistencies between the SIPP and IRS universes and issues of missing data. Consequently, our initial study focused on whether such efforts were justified by any expected reductions in variance. The preliminary study evaluated the apparent improvements in variance by adjusting to IRS controls only. The overall effect of a more likely candidate estimator, i.e., one that combines the administrative controls with the demographic controls now in use, awaits further research.

We note that even though this research was conducted in the context of improving longitudinal estimates, for example, calendar year estimates from the SIPP, the methodology is applicable to cross-sectional estimation as well. Specifically, SIPP cross-sectional estimates for characteristics collected in a single interview for a four-month period may also benefit from a similar approach. Applications to CPS characteristics may also be appropriate, including income statistics from the Annual Demographic Supplement conducted in March of each year.

The next section describes the assumptions of the approach and the feasibility of using administrative income data. The succeeding section details the raking ratio estimation used to control SIPP estimates to administrative income data. Available SIPP data and the description of the research procedure follows.

Our preliminary results suggest substantial potential improvements for some characteristics, particularly with respect to statistics on income. The largest gains appear for statistics that depend heavily on the middle and upper end of the income distribution -- mean and median income -- but results for a poverty measure are also encouraging. The improvements for Blacks, and particularly for Hispanics, are less but still notable. The procedure also reduces variances for estimates of Food Stamp recipiency, but yields a mixed outcome for AFDC recipiency.

The final section reviews limitations of the present study and recommendations for future research. We suggest that the preliminary results provide a sufficient rationale for continued investigation.

2. Assumptions and Feasibility of Using Administrative Income Data

Initially, we considered several administrative sources: IRS, Social Security, Food Stamps and AFDC files. The accessibility, timing and size of these files limited us to using only IRS data, however. For example, the current availability of Food Stamp files only at the State level imposes severe operational difficulties in assembling a national file. Also, special permission must be granted to obtain some program data files. The wide coverage by the IRS file of the adult population made it the logical starting point.

IRS prepares the individual tax files primarily for administrative rather than statistical purposes. The records on the file represent returns, indexed by the Social Security number of the primary filer, rather than persons. The Census Bureau, exclusively for statistical purposes, matches a 20-percent sample of IRS returns, sampled according to Social Security number, to Social Security records to determine the age, race, and sex of the primary filer. Hispanic surnames are also identified on the basis of a standard list of such names; this determination is made by computer and is not necessarily consistent with how these persons would classify themselves in a census or survey. The 20-percent matched file offers the potential for tabulations of demographic characteristics of the primary filer with characteristics reported on the tax forms, such as adjusted gross income and number of exemptions.

Simply for the sake of economy, we employed a subset file, representing one percent of the total IRS file, in place of the 20-percent file. Presumably, the 20-percent file may be substituted later for the one-percent file should these procedures be fully implemented. Although a 20-percent file is still a sample, the effect of sampling variability is nil relative to the sampling variances of SIPP characteristics. Consequently, we treat all IRS totals as if they were free from sampling error.

Except for a small percentage of adults in the sample, SIPP respondents provided Social Security number, and this key was employed to match to the IRS file for both primary and secondary (i.e., spouse on a joint return) filers. The match to IRS was conducted whether or not the sample persons also fell into the 20-percent sample, including the same coding for Hispanic surname.

There are some inconsistencies between the SIPP and IRS universes, to the extent that some IRS returns represent persons not in the SIPP universe. The refusal to prove Social Security number by a few percent of the SIPP sample is also a matter of concern. Overlooking such difficulties for the moment, however, the SIPP sample, once matched to the IRS file, may be used to estimate cross-classifications by age, race, sex, and Hispanic surname status of the primarily filer by income and other characteristics from the return. Analogous cross-classifications are available from the one-percent, and potentially the 20-percent, IRS samples. The situation suggests that some form of estimation, such as ratio estimation or the raking ratio estimation actually applied, could reduce the sampling variance of some or many SIPP characteristics.

The ratio estimation procedures employed for the Census Bureau's current surveys involve all of the sampled cases. For example, all SIPP cases are currently controlled by age, race, sex, and householder/nonhouseholder status. The demographic controls represent the entire universe of the survey, so the involvement of all sample cases is natural. The situation is different with respect to IRS data, however, since many SIPP respondents are legitimately not in the IRS universe. Consequently, the weighting adjustments under study were for only SIPP sample cases linked to an IRS return. The weights of SIPP respondents not linked to a return remained unchanged in this stage of the investigation. Later methods to combine the IRS and demographic controls will presumably again involve all of the SIPP sample cases.

3. Raking Ratio Estimation

We illustrate raking ratio estimation (Brackstone and Rao, 1976) by a simple example. Suppose x_{ij} is a population total and $X_{ij}^{(0)}$ a corresponding weighted sample estimate. If each cell $X_{ij}^{(0)}$ were sufficiently precise, ratio estimation (e.g., Cochran 1977) to x_{ij} within each cell would generally be the estimator of choice. When the sample data are less precise, however, then the raking ratio estimator is a possible alternative. This estimator modifies the sample weights to force consistency between the row and column totals of the sample estimates and the corresponding population totals, $x_{i.}$ and $x_{.j}$, without requiring cell-by-cell consistency. This is done through alternating stages of proportional adjustment of the sample data to the population marginal totals. In a different context, essentially the same procedure appears as the iterative proportional fitting algorithm for contingency tables, which yields maximum-likelihood estimates for hierarchical factorial log-linear models. In more detail, we let $W_{ijk}^{(0)}$ denote the weight attached to sample case k that falls in the $(i,j)^{th}$ cell of the cross-classification. $W_{ijk}^{(0)}$ reflects the inverse of the probability of selection, modified by any previous stages of estimation, such as adjustments for noninterviews. The initial sample estimates of each cell of the cross-classification, X_{ij}, are given by

$$X_{ij}^{(0)} = \sum_k W_{ijk}^{(0)}.$$

The iterative raking ratio estimator is defined recursively at stage t of the iteration by

$$W_{ijk}^{(t)} = W_{ijk}^{(t-1)} \frac{x_{.j}}{\sum_i \sum_k W_{ijk}^{(t-1)}} \qquad t \text{ even},$$

$$W_{ijk}^{(t)} = W_{ijk}^{(t-1)} \frac{x_{i.}}{\sum_j \sum_k W_{ijk}^{(t-1)}} \qquad t \text{ odd}.$$

If t is even, the weighted sample total in each column of the matrix is exactly equal to the known population total for the column, while, if t is odd, this equality is exact for the rows of the matrix.

In some applications, the iterations are halted after a specific number of steps, while in others a convergence criterion determines the number of iterations. Convergence is not always assured. Convergence was completely satisfactory in this application to the SIPP, however. The methodology can be extended further to a series of multidimensional control tables. The weights are again computed in a similar fashion through iteration on each control dimension.

4. Data and Issues of Timing

Because of questions on the timing of the availability of the extract IRS files to the Census Bureau, a likely form of implementation for the methodology would be to employ IRS files for the prior year, for example, to adjust calendar year (CY) 1984 SIPP estimates to 1983 IRS data, requiring 1984 CY SIPP sample weights and 1983 CY IRS data. At the time research was initiated, these data were not available.

The Census Bureau had, however, prepared a 12-month longitudinal file, called the 1984 SIPP 3-interview research file, with appropriate longitudinal weights. The SIPP 3-interview file covers the months June 1983-August 1984. Four rotation groups (SIPP quarter samples) were identified for field purposes, and these rotations have overlapping reference periods. For example, rotation group 1 covers the reference months June 1983-May 1984 and rotation group 2 covers the reference months July 1983-June 1984. Therefore, the same twelve months of data are not included for each rotation group on the 3-interview file. Thus, this file of overlapping 12-month periods by rotation was selected for its availability.

A match of 1984 IRS data (close to a 100% IRS file, but missing a few percent of late returns) to a SIPP 1984 first-interview data file had already been completed for other researchers at the Census Bureau. IRS extract data from this file was attached to the 1984 SIPP 3-interview file by matching on Social Security numbers. Approximately 56% of SIPP persons matched to an IRS record. Husbands and wives who filed jointly received the same IRS data. The remaining SIPP population, those who did not match to IRS data, we refer to as nonmatches. These nonmatches are a result of persons who did not file IRS returns, persons who filed but whose return was filed too late for inclusion in the IRS file used in the analysis, and persons for whom Social Security numbers were not available or were incorrect.

The one-percent IRS sample file available for use as controls contains some portions of the U.S. population not covered by the SIPP sample. For example, some institutionalized persons file tax returns, but the SIPP excludes institutionalized persons in its sample. The controls used in this research thus cover a slightly larger population from the SIPP and may introduce some bias. We estimate the maximum amount of bias to be 2.4% for estimates of total population. Similarly, controlling to IRS totals will tend to overestimate SIPP income aggregates. Consequently, we have primarily restricted our analysis on comparisons of variance for means, medians, and similar distributional statistics at this point, recognizing that nonsampling error issues will become important at later stages.

5. Research Procedure

The overall research procedure involves identifying specific SIPP characteristics to undergo the ratio adjustment procedure, preparing controls from IRS data, implementing the estimation, and calculating selected estimates and their variances to analyze the effects of the reweighting.

In order to identify SIPP characteristics to be adjusted, cross-classifications from the matched file of only those SIPP persons who matched to IRS data were computed. The summary tables involved characteristics either available from the IRS data, i.e., adjusted gross income, Hispanic surname, and number of exemptions, or available through a match to Social Security Administration (SSA) records, i.e., age, race, sex. For each type of return: joint, single, and (non-joint) household, marginal tables were identified that could be expected to yield at least 20 SIPP sample cases in each cell of the marginal table to permit reweighting to the IRS controls for the same marginal table.

Analogous cross-classification tables from the IRS one-percent sample were prepared as control tables. Attachment A lists the marginal tables involved in the reweighting by type of return. The SIPP data were proportionally adjusted to each of the sets simultaneously. These reweighted estimates from the SIPP sample agreed with IRS estimates within the level of precision displayed in the marginal tables. Estimates of selected SIPP characteristics were then calculated from the original and reweighted SIPP data sets.

Although the raking ratio estimation was defined in terms of demographic characteristics of the primary filer, we also applied the resulting adjustment to the weight of the primary filer to the secondary filer in SIPP households in instances in which both members of the couple could be obviously linked. Predominantly, the primary filer was the husband, so the weight of his wife would receive the same proportional adjustment as his own weight. Because the adjusted gross income on the joint return represents the combined income of the spouses, this procedure appeared the most effective use of the raking, in preference to adjusting only the primary filer's weight, particularly for individual and family characteristics depending on the combined income of the couple, e.g., poverty status.

Comparisons of variances before and after reweighting were based on a modified form of half-sample replication. Replicate factors were available for use in computing variances for SIPP 3-interview estimates using a half-sample replication technique. These factors essentially create half samples from the SIPP data so that the average squared difference between half-sample and full-sample estimates provides an estimate of the sampling variance of the statistic. Wolter (1985) describes the general methodology, and Dippo, Fay, and Morganstein (1984) illustrate applications of the replicate weighting approach.

To determine the variance implications of this approach then, each replicate-weighted set of SIPP data was independently reweighted. Variances were computed for selected SIPP income estimates according to the original weights and replicate weights, and according to the reweighted weights and reweighted replicate weights.

6. Results

Tables 1 and 2 summarize the principal findings from this preliminary study. Table 1 focuses on the annual individual income for persons who are age 25 or over at the beginning of the first wave of data. The income figures generally represent 12 months of data; however, the incomes for longitudinal sample persons who leave the universe before the end of the 12-month period are simply taken to be the total of their monthly incomes over the period in which they were in the universe. As mentioned earlier, the income figures do not correspond to a calendar year nor, in fact, to any one 12-month period for the entire sample.

Table 1 shows substantial reductions in variance for estimates of income for the population age 25 and over. In particular, the sampling variance of the estimated mean income for the overall population is reduced by an estimated 54 percent. Approximately equal proportional gains in reliability for mean income appear for males and for females. Improvements in the reliability of the income distribution result for both sexes, with the most dramatic gains for males occurring at the relatively higher end of the distribution, especially in estimating the total for either $20,000 or $30,000 and above. The gains for females are more evenly spread among the categories shown.

The adjustments generally benefit the estimates for Blacks, but less consistently. The estimated income distribution for Black females shows little improvement, although the variance of mean income is reduced. Both the income distribution and the mean income show improvements for males and for both sexes combined.

The results for Hispanics are mixed and even less promising than those for Blacks. It is likely that the coarseness of the raking for characteristics by ethnicity, necessitated by relatively small sample sizes, prevented gains of the same magnitude as those for the overall population. Furthermore, Hispanic surname, employed in the adjustments to IRS control totals, may have yielded less variance advantage because of a considerably less than perfect correlation with Hispanic origin reported in the SIPP.

Table 2 presents variance estimates for other characteristics. The first column of the table compares variances for estimates of the percentage of person months in poverty. This statistic is based on comparing the person's family income for a month with 1/12 of the low income poverty cut-off based on the current rate of inflation. This comparison is made separately for each month in which a SIPP sample person with positive longitudinal weight was in the universe. Months in which the person was not in the universe are excluded. Empirically, this measure based on monthly comparisons has been shown to produce a higher rate of poverty than the definition employed in the Current Population Survey (CPS) based on a full year's worth of income, in large part because of temporary short spells of poverty experienced by many people who would not be classified as poor for the year under the concepts of the CPS.

The results for the poverty measure are promising, especially for the overall population and for Blacks. The results for Hispanics are mixed, showing minor gains for the overall Hispanic population age 25+, but not for either sex separately. Nonetheless, the fact that some gains appear in the comparison is encouraging, since it suggests that further elaboration of the multiple raking approach may offer additional improvements. In particular, integration of the raking with adjustments to demographic controls now employed in the current longitudinal weighting procedures may produce additional gains at the lower end of the income distribution, where many are not required to file an IRS form.

Table 2 also presents variance comparisons for recipiency characteristics for three different transfer programs. For purposes of Table 2, persons were considered recipients for a program if they were included for one or more months. The comparison for Food Stamp recipients is relatively mixed, yielding some improvements in the overall distribution and for Blacks, but higher variances for Hispanics. AFDC is even more mixed, with no overall gain and again poorer results for Hispanics.

Table 2 implies substantial improvements for Social Security recipiency for the overall population. In fact, however, this result is primarily an artifact of the variance calculations in this study. The current SIPP longitudinal weights incorporate ratio estimation to control totals including age by sex, but the replicate weights do not properly reflect the variance reduction arising from this aspect of the ratio estimation. On the other hand, the reweighting of the SIPP file to control totals from the IRS explicitly incorporated a control to age totals according to the age of the primary filer. Hence, the estimated variances for the reweighted estimates reflect some effect of the age distribution, while the variance estimates for the original weighting do not. No variance advantage from the reweighting appears for the proportion with Social Security income for the population age 65 and over; indeed, the reweighted estimates appear slightly worse. On the other hand, the reweighted file still has smaller variances for the population 65 and over for Food Stamp recipiency. Consequently, there is reason to discount the results for Social Security recipiency reported in table 2, but the results for Food Stamps and AFDC are probably considerably less affected by the methodological limitations of this preliminary analysis.

7. Discussion and Recommendations for Further Research

In this research, we examined person-level SIPP characteristics and obtained good overall reductions in variances. We think that even greater variance reductions are possible for SIPP estimates at the family and household levels, since the IRS annual gross income values often reflect family and household income. Person, family and household characteristics are all of great importance to the SIPP.

The gains in table 1 for person-level income probably represent an upper bound on the gains for person-level characteristics that should be expected from further refinements of the method. Among the reasons for this are: that the effect of control to IRS values of adjusted gross income probably produces the greatest effect for SIPP income characteristics; that the period covered by the SIPP income variables partially overlaps with the IRS tax year, while, as noted earlier, a more likely form of application of this procedure would be to control to IRS data for the previous year; and that reintroduction of the demographic controls into a combined raking strategy may dampen some of the gains reported here.

As yet, we have not systematically experimented with altering the number and coarseness of the marginal tables, but such experimentation appears desirable. The differences noted earlier between the SIPP and IRS populations will require further adjustments, possibly requiring data from an external source.

Interestingly, the number of single returns for filers in their early 20's was substantially less than expected. Corresponding, joint returns for this age group appeared overrepresented in the SIPP relative to IRS. This suggests that some coverage problems may exist with the early 20's age group. Possibly, the associated bias may be reduced using administrative income data as controls. For the nonmatch population on the lower end of the income distribution, which is typically the group with more coverage problems, this observation is especially important. Since the possibility of using other administrative data for nonmatches will probably be explored, coverage of the lower income groups may be improved with this type of procedure.

Variance reductions for the SIPP longitudinal person-level characteristics examined in this research are encouraging. In addition to improving SIPP longitudinal estimation, SIPP cross-sectional estimation may be improved with a similar procedure. More directly, March CPS income estimates may benefit from the same procedure, since CPS March income estimates are calendar year estimates. Based on our results, further development of this procedure for SIPP longitudinal estimation is justified, and its usefulness to other types of estimation and other surveys appears attractive.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

We wish to thank Rameswar Chakrabarty, Maurice Kniceley, Thomas Moore, and Gary Shapiro for their comments, and Alice Coburn for her help in preparing the final version.

References

Brackstone, G.J. and J.N.K. Rao, "Raking Ratio Estimators", Survey Methodology, Statistics Canada, 2, 1976.

Cochran, William G. Sampling Techniques, New York: John Wiley & Sons, 1977.

Dippo, Cathryn S., Robert E. Fay, and David H. Morganstein (1984), "Computing Variances from Complex Samples with Replicate Weights," Proceedings of the Survey Research Methods Section, Washington DC: American Statistical Association, pp. 495-500.

Wolter, Kirk, Introduction to Variance Estimation, New York: Springer-Verlag, 1985.

Table 1 Ratios of Estimated Variances After and Before Adjustment to Administrative Totals

	Percentages of Income Distribution					Mean
	Loss- $10K	$10K- $20K	$20K- $30K	$30K+	$20K+	Income
Total Age 25+	.49	.80	.58	.41	.38	.46
Males	.53	.93	.70	.38	.35	.46
Females	.48	.58	.61	.78	.54	.49
Black Age 25+	.74	.91	.87	.80	.75	.69
Males	.68	.93	.87	.74	.65	.61
Females	.81	.97	1.15	1.12	1.15	.74
Hispanic Age 25+	1.03	.83	.82	1.01	.69	.83
Males	1.23	.86	.77	.91	.68	.86
Females	.79	.81	.81	1.07	.83	.94

Table 2 Ratios of Estimated Variances After and Before Adjustment to Administrative Totals

	Mos. in Pov.	Food Stamp	AFDC Recip.	Soc. S. Recip.
Total 25+	.74	.89	1.00	.27
Males	.71	1.01	1.13	.27
Females	.80	.81	.99	.40
Black 25+	.71	.76	.89	1.14
Males	.65	.81	1.42	.94
Females	.78	.77	.87	1.21
Hispanic 25+	.89	1.21	1.15	.85
Males	.99	1.13	1.11	.91
Females	1.00	1.23	1.17	.98

Attachment A

A.1 Marginal Tables for the Adjustment of Joint Returns

1. Age2 by AGI2
2. Age2 by Race
3. Age2 by Hispanic
4. Age4 by Number of Exemptions (1-2/3/4 /5/6+)
5. Age4 by AGI1
6. Number of Exemptions (1-2/3/4/5/6+) by Race
7. Number of Exemptions (1-2/3/4/5/6+) by Hispanic
8. Number of Exemptions (1-2/3/4/5/6+) by AGI2
9. AGI2 by Race
10. AGI2 by Hispanic

A.2 Marginal Tables for the Adjustment of Single Returns

1. Age1 by Sex
2. AGI1
3. Age3 by AGI2
4. Age3 by Race by Sex
5. Age3 by Hispanic
6. Age3 by Number of Exemptions (1/2/3+)
7. Hispanic by Sex
8. Number of Exemptions (1/2/3+) by Sex
9. AGI2 by Sex
10. AGI (-$10K/$10-20K/$20K+) by Race by Sex
11. AGI (-$10K/$10-20K/$20K+) by Hispanic
12. AGI (-$10K/$10-20K/$20K+) by Number of Exemptions (1/2/3+)
13. AGI (-$10K/$10-20K/$20K+) by Age3 by Sex
14. Number of Exemptions (1/2/3+) by Race

A.3 Marginal Tables for the Adjustment of Household (Non-joint) Returns

1. Age3 by Race
2. Age3 by Sex
3. AGI2 by Sex
4. AGI2 by Age4
5. AGI2 by Race
6. Hispanic by Sex
7. Number of Exemptions (1-2/3/4+) by Sex
8. AGI (-$10K/$10-20K/$20K+) by Race by Sex
9. AGI (-$10K/$10-20K/$20K+) by Hispanic
10. AGI (-$10K/$10-20K/$20K+) by Number of Exemptions (1-2/3/4+)
11. Number of Exemptions (1-2/3/4+) by Race

Notes: Race (Black/Non-Black),
Hispanic (Hispanic/Non-Hispanic),
AGI1 (Under $2500/$2500-4999/$5000-7499 /$7500-9999/$10-15K/$15-20K /$20-25K/$25-30K/$30-35K/$35-40K /$40-45K/$45-50K/$50-75K/$75K+)
AGI2 (Under $10K/$10-20K/$20K-30K /$30K+)
Age1 (-17/18-24/25-34/35-44/45-54/44-64 /65+)
Age2 (-24/25-34/35-44/45-54/55-64/65+)
Age3 (-24/25-34/35-44/45-54/55+)
Age4 (-34/35-44/45-54/55+)

Age is the age of the primary filer. For joint returns, this person is generally, but not exclusively, the husband.