USE OF ADMINISTRATIVE DATA IN SIPP LONGITUDINAL ESTIMATION

Suzanne M. Dorinski and Hertz Huang[1]
Suzanne M. Dorinski, Demographic Statistical Methods Division, U.S. Bureau of the Census, Washington, DC 20233

Introduction

The Survey of Income and Program Participation (SIPP) currently uses cross-classifications of age, race, sex and householder/nonhouseholder status as controls in longitudinal estimation. The controls come from the Current Population Survey (CPS), which has its own controls based on post-censal estimates of age, race and sex. Previous research by Huggins and Fay [1988] ratio adjusted the SIPP sample that could be matched to Internal Revenue Service (IRS) records but did not control the nonmatched sample. They found a reduction in variances for most income and program participation variables. Subsequent research applied demographic totals based on the CPS controls for age, race, sex and ethnicity, to ratio adjust the estimates based on the SIPP sample that did not match to the IRS records. We combined the nonmatched and matched samples and then calculated estimates along with their variances.

Final results indicate large reductions in variances for many income and income related characteristics, with some variances affected adversely. Some variance estimates for Hispanics and to a lesser extent Blacks increased. Bias of the estimates studied either did not change or increased.

The next section describes the previous research done by Huggins and Fay. The succeeding section outlines the methodology used to ratio adjust the nonmatched sample to CPS controls. The variance results and effects of the new weighting on bias follow. The final section presents recommendations for further research.

Background

Previous researchers, Huggins and Fay [1988], matched the SIPP 1984 3-interview research file to a 1984 IRS file. The SIPP 1984 3-interview research file is a 12 month longitudinal file with appropriate longitudinal weights, covering June 1983 - August 1984. SIPP respondents were matched to the 100-percent IRS file through their social security number (SSN). Both primary and secondary filers (i.e., spouse on a joint return) were matched. IRS extract data was then attached to the SIPP file. 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 included persons who did not file IRS returns, persons who filed too late, and persons for whom SSNs were not available or were not correct.

Many issues are unresolved. There are differences between the SIPP universe and the IRS universe. Some IRS returns represent persons not in the SIPP universe. For example, some institutionalized persons file tax returns, but the SIPP excludes institutionalized persons in its sample. Many SIPP respondents are legitimately not in the IRS universe. Children with no income of their own do not file income tax returns, yet may be SIPP respondents. These differences introduce a bias, but it is thought to be no more than 2.4 percent for estimates of total population. Thus, the initial study focused on whether the approach was justified.

The IRS files contain returns indexed by the SSN of the primary filer. Strictly for statistical purposes, the Census Bureau matches a 20-percent sample of IRS returns (sampled according to SSN) to Social Security Administration records. From this file the age, race, and sex of the primary filers can be determined. Simply for the sake of economy, the researchers used a subset file, representing one percent of the total IRS file, to create controls for the raking ratio adjustment. The 20-percent file may be substituted for increased reliability for the one percent file should these procedures be implemented.

Huggins and Fay prepared cross-classification tables using the SIPP respondents that matched to the IRS file. These tables involved characteristics either available from the IRS file (adjusted gross income, Hispanic surname, and number of exemptions) or through a match to the Social Security Administration

---

[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 Vicki Huggins, Robert Fay, David Adams, James Clement, James Lessard and Kevin Cooper for their work on this project; Raj Singh, James Hartman, John Bushery and Jay Kim for their comments on this paper; and Sandy Carnegie for her help in preparing the final version.

Hispanic surname, and number of exemptions) or through a match to the Social Security Administration records (age, race, or sex). For each type of return (joint, single, and (non-joint) household), they identified marginal tables that could be expected to yield at least 20 SIPP sample cases in each cell. Analogous tables from the one percent IRS sample were prepared as control tables. These control tables were used to proportionally adjust the SIPP data to each set simultaneously using an iterative raking procedure. (For more information on the raking procedure, see Huggins and Fay [1988].) The weights of SIPP respondents not linked to a return remained unchanged. Estimates of selected SIPP characteristics were then calculated from the original SIPP data and the reweighted SIPP data.

Although the raking ratio estimation was defined in terms of demographic characteristics of the primary filer, the primary filer's adjustment was also applied to the weight of the secondary filer in SIPP households where couples could be obviously linked. Thus the weight of the secondary filer (usually the wife) received the same proportional adjustment as the primary filer. Since the adjusted gross income on a joint return represents the combined income of the spouses, this procedure appeared to be the most effective use of the raking compared to adjusting only the primary filer's weight, particularly for individual and family characteristics that depend on the combined income of the couple e.g., poverty status.

Many SIPP respondents are not in the IRS universe; hence the weighting adjustments in Huggins and Fay's study are only for SIPP sample cases linked to a return. For selected SIPP income estimates, they used the ratio of the estimated variances, with and without the IRS adjustment, as a comparison. The variances were calculated using a modified form of half-sample replication. Each replicate-weighted set of SIPP data was independently re-weighted using the raking procedures.

The results, based on respondents age 25 or older, showed considerable variance improvements for most variables. The largest gains appear for statistics that are highly related to the middle and upper end of the income distribution. The adjustments generally benefit the estimates for Blacks, but less consistently. The results for Hispanics are mixed and less promising than those for Blacks.

Methodology

We anticipated that a further reduction in variance could be achieved by ratio adjusting the nonmatched sample to CPS controls. First we estimated cross-classification tables by age, race, sex, and ethnicity for nonmatched respondents from the SIPP 3-interview file. Then we controlled these tables to analogous tables constructed from CPS based controls.

The nonmatched controls were simply the difference between the estimates from the SIPP respondents that matched to the IRS file and the CPS based controls. We ratio adjusted the nonmatched sample to these controls. The nonmatched and matched samples were then combined and the estimates were calculated along with their variance estimates.

In order to compare our results with the previous results, we initially focused on persons age 25 or older. We then applied the same techniques to persons 15 or older, since they are the primary interest of SIPP. In all cases we compared the variance estimates to the current SIPP longitudinal weighting variance estimates. Due to space limitations, only the results for persons 15 or older are presented in the tables.

Variance Results

In order to judge the changes before and after the adjustment, we looked at the following ratio:

$$\frac{\text{(variance after adjustment)}}{\text{(variance before adjustment)}}$$

A ratio of less than 0.95 indicated a significant decrease in the variance estimate after adjustment. A ratio of 0.95 or greater indicated either no change or an increase in the variance estimate after the adjustment.

We decided to examine other variables since (1) there were significant gains for the majority of income related variables and (2) we feared that improving the variances for some variables might increase the variances for other variables.

Table 1 shows reduction in sampling variances for most of the estimates studied. However, it should be noted that the variances for Black females with annual incomes of \$20,000 to \$30,000 and \$30,000+ actually increased. The variance estimates for Hispanics with annual incomes of \$30,000+, Hispanic males with annual incomes of \$10,000 to \$20,000 and mean income of Hispanic females were also affected.

Table 2 presents variance ratios for the estimated number of recipients for the following government programs: food stamps (FOOD), Aid to Families with Dependent Children (AFDC), AFDC or General Assistance (AFDC/GA), Veterans' compensation (VET), the Supplemental Food Program for Women, Infants and Children (WIC), Federal Supplemental Security Income (SSI), Social Security (OASDI), and unemployment compensation (UNEMP). To be a

recipient of a program, a person must have received benefits from the program one or more months.

Table 2 shows reduction in sampling variances for about half of the estimates examined. However, estimates of Hispanics receiving food stamps, Hispanics receiving AFDC, Hispanics receiving AFDC or General Assistance, Hispanics receiving WIC benefits, Blacks receiving Social Security and Blacks receiving unemployment compensation are among the estimates that either did not show a reduction in sampling variances or experienced an increase in sampling variances.

Several demographic estimates are presented in Table 3. We found reduction in sampling variances for about half of the estimates examined. However, estimates of the percentage of Hispanics ever married, divorced, or separated, and estimates of the percentage of total males and total females ever separated are included in the estimates that either did not show a reduction in sampling variances or had an increase in sampling variances.

Certain unemployment and employment characteristics are presented in Table 4. We found reduction in sampling variances for about half of the estimates examined. Note that employment and unemployment characteristics for Hispanics and unemployment characteristics for Black males and total Blacks had increases in sampling variances.

From Table 5, we see that the variance estimates for ever-disabled and ever-received wages or salary have decreased significantly for total population and for Blacks, while increasing for Hispanics. For ever-received property income, only the variance estimate for Black males has decreased.

Finally, in Table 6, the variables (1) all 12 months in poverty, (2) percentage below poverty for at least one month, and (3) percentage of months in poverty were examined. The variances showed overall improvement except for most Hispanic characteristics.

Effects on Bias

While the primary focus of the research had been on reducing the variance of SIPP estimates, we also wanted to see what effect the adjustment had on the bias. The estimates previously discussed do not have easily obtainable benchmarks, so we looked at different estimates to analyze the effects on bias. We looked at monthly estimates of the population 15+ covered by Social Security, the population covered by AFDC, the population covered by food stamps, and the population 15+ covered by SSI.

We derived benchmarks for the estimates by following the methodology outlined in a report done by Czajka, Doyle, Walker, Whitmore and Citro [1982]. That report documents benchmarks for income and labor force statistics from the 1979 Income Survey Development Program research panel, a precursor to SIPP. The general method for deriving benchmarks is to get the administrative record totals and make the necessary adjustments to the SIPP population. Administrative record totals for Social Security, AFDC and SSI are published in the Social Security Bulletin. Administrative record totals for food stamps are published by the Food and Nutrition Service of the U.S. Department of Agriculture. To adjust the administrative record totals to the SIPP population, one subtracts beneficiaries living outside the U.S. and institutionalized beneficiaries. Some of the beneficiaries shown in the administrative record totals die before they receive the program benefit for a particular month, so there is also an adjustment made to reflect deaths.

Vaughn [1989] reported the data quality of the income estimates in the 1984 SIPP panel. His report analyzed quarterly estimates which were calculated using cross-sectional weights. While the results of this paper will be slightly different since longitudinal weights were used for this analysis, the same types of results and trends should be seen.

For Social Security coverage, the before-adjustment and after-adjustment estimates are significantly different. The adjustment has increased the bias of the estimates slightly. Table 7 shows the before and after comparisons. Vaughn reported that the SIPP estimates of Social Security recipients ranged between 96 and 99 percent of the benchmark.

The before-adjustment and after-adjustment estimates for AFDC coverage and for food stamp coverage are not significantly different. Hence the adjustment has not changed the bias of those estimates and differences are not analytically important.

For SSI coverage, the before-adjustment and after-adjustment estimates are significantly different and the adjustment appears to have increased the bias of the estimates. Table 8 shows the before and after comparisons. Vaughn found that the SIPP estimates of SSI recipients averaged 97 percent of the benchmark, yet there was an upward trend over time.

What Went Wrong?

We thought that the varied results were due to differences in the subpopulations represented in the estimates examined. Overall, SIPP records matched to IRS records 56% of the time. However, the match rate varied quite a bit for subpopulations. 15% of the AFDC recipients matched to IRS records, while 24% of the food stamp recipients matched. 50% of the

Social Security recipients matched to IRS records, while only 8% of the SSI recipients matched.

We assumed that the reason the adjustments (to IRS controls for the matched sample, to CPS controls for the nonmatched sample) did not significantly change the AFDC and food stamp estimates was because relatively few of those recipients matched to IRS records. But relatively few SSI recipients matched to IRS records, yet the estimates for SSI did change significantly.

Further investigation revealed that the weights of matched SSI recipients were increasing after the adjustments, while the weights of nonmatched SSI recipients were decreasing. We had not expected this kind of result. Weights of matched Social Security recipients also increased after the adjustments, while the weights of nonmatched Social Security recipients decreased. However, the effect was not as pronounced as for SSI recipients. The changes in the weights were less dramatic for AFDC and food stamp recipients.

Recommendations for Further Research

Since only 56% of the SIPP records matched to an IRS record, we plan to look at further adjustments to deal with the population difference between the SIPP and IRS. For the nonmatched sample, we may try to do an adjustment based on CPS income related data, such as controlling to ever worked, or ever received wages and salary. We may also study variance reductions for other SIPP estimates including estimates at the family and household levels. Other SIPP estimates that may be studied include health care estimates, estimates of program transitions, and program participation spell estimates. Since some ratio factors for Blacks and Hispanics are large, further collapsing should improve variances for these groups.

We may explore using income on IRS records to impute a monthly income instead of using income reported in the interview for those SIPP cases that match to IRS records.

We may also try using the 20 percent sample of IRS records in place of the one percent file used to create the controls.

This research was based on the 1984 Panel of the SIPP. The sample design has changed since the 1984 Panel. For example, the 1990 Panel includes an oversample of households headed by Blacks, Hispanics, or females with no spouse present living with relatives. The 1992 and 1993 Panels are larger than the 1984 Panel. Results for Blacks and Hispanics may be better in more recent panels. We may redo the adjustments with data from one of these panels.

References

Czajka, John L., Pat Doyle, Michael Walker, Richard Whitmore and Constance F. Citro (1982), "Benchmark Estimates for Transfer Income and Labor Force Statistics from the 1979 ISDP Research Panel." A report prepared pursuant to contract HEW-100-79-0129 by Mathematica Policy Research, Inc., Washington, DC.

Huggins, Vicki J. and Robert E. Fay (1988), "Use of Administrative Data in SIPP Longitudinal Estimation," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 354-359.

Vaughn, Denton R. (1989), "Reflections on the Income Estimates from the Initial Panel of the Survey of Income and Program Participation (SIPP)." In Individuals and Families in Transition: Understanding Changes Through Longitudinal Data. Papers presented at the Social Science Research Council Conference in Annapolis, Maryland, March 16-18, 1988, U.S. Bureau of the Census.

Table 1. Ratios of Estimated Variances After and Before Adjustments to Administrative Data

	Annual Income Distribution					
	Loss-$10K	$10K-$20K	$20K-$30K	$30K +	$20K +	Mean Income
Total	.29*	.57*	.67*	.42*	.35*	.39*
Males	.40*	.88*	.96	.41*	.39*	.47*
Females	.37*	.47*	.60*	.82*	.50*	.41*
Black	.42*	.66*	.74*	.81*	.70*	.52*
Males	.50*	.78*	.75*	.81*	.61*	.48*
Females	.52*	.69*	1.32	1.26	1.33	.72*
Hispanic	.67*	.75*	.83*	1.00	.73*	.86*
Males	.86*	.99	.84*	.96	.70*	.82*
Females	.66*	.68*	.80*	1.08	.79*	1.01

* - Indicates significant decrease in variance after adjustment (ratio <0.95)

Table 2. Ratios of Estimated Variances After and Before Adjustments to Administrative Data
Program Participation

	FOOD	AFDC	AFDC/GA	VET	WIC	SSI	OASDI	UNEMP
Total	.80*	1.04	1.03	.87*	.99	.74*	1.21	.84*
Males	.87*	1.00	.96	1.00	-	.73*	1.06	1.07
Females	.80*	1.11	1.07	.83*	1.00	.79*	1.06	.89*
Black	.70*	.73*	.88*	.86*	1.10	.86*	1.16	.97
Males	.83*	.86*	.93*	.89*	-	.99	1.10	1.12
Females	.62*	.81*	.89*	.95	1.10	.86*	1.29	.89*
Hispanic	1.07	1.21	1.30	.83*	1.15	.83*	.88*	1.20
Males	1.09	1.30	1.42	1.00	-	.85*	.83*	1.44
Females	1.02	1.20	1.23	.83*	1.14	1.13	.93*	.85*

Table 3. Ratios of Estimated Variances After and Before Adjustments to Administrative Data
Marital Status

	% Ever Married	% Ever Divorced	% Ever Separated
Total	.55*	.77*	1.23
Males	.93*	1.11	1.38
Females	.70*	.85*	1.09
Black	.68*	.75*	.80*
Males	.50*	.95	1.18
Females	.84*	.79*	.73*
Hispanic	1.05	1.42	1.28
Males	1.24	1.89	1.28
Females	.85*	1.01	1.07

Table 4. Ratios of Estimated Variances After and Before Adjustments to Administrative Data
Employment/Unemployment Characteristics

	Unemp 1	Unemp 2	Emp 1	Emp 2
Total	.68*	.74*	.70*	.72*
Males	.86*	.88*	.72*	.73*
Females	.89*	.88*	.82*	.83*
Black	1.09	1.02	.74*	.75*
Males	1.33	1.17	.82*	.82*
Females	.84*	.88*	.61*	.60*
Hispanic	1.29	1.09	1.58	1.55
Males	1.70	1.42	1.31	1.27
Females	1.03	1.10	1.85	1.81

* - Indicates significant decrease in variance after adjustment (ratio <0.95)

Unemp 1 an individual is (1) with a job an entire month but missed one or more weeks, spent time on layoff, or (2) with job one or more weeks, spent some time looking or on layoff, or (3) no job during a month, spent entire month looking or on layoff, or (4) no job during month, spent one or more weeks looking or on layoff.

Unemp 2 an individual (1) has no job during a month, or conditions (3) and (4) from Unemp 1.

Emp 1 an individual is with a job an entire month, and worked all weeks.

Emp 2 is Emp 1, or with a job an entire month, and missed one or more weeks with no time on layoff.

Table 5. Ratios of Estimated Variances After and Before Adjustments to Administrative Data
Ability to Work/Income Received

	% Ever Disabled	% Ever Rec'd Wages or Salary	% Ever Rec'd Property Income
Total	.75*	.71*	1.01
Males	.80*	.85*	.97
Females	.79*	.75*	1.22
Black	.82*	.78*	1.14
Males	.95	.80*	.92*
Females	.80*	.72*	1.57
Hispanic	1.23	1.38	1.17
Males	1.23	1.09	1.27
Females	1.21	1.47	.95

* - Indicates significant decrease in variance after adjustment (ratio <0.95)

Table 6. Ratios of Estimated Variances After and Before Adjustments to Administrative Data
Poverty Measures

	% In Poverty All 12 Months	% In Poverty At Least 1 Month	% Months In Poverty
Total	.84*	.75*	.65*
Males	.86*	.74*	.63*
Females	.88*	.80*	.73*
Black	.78*	.62*	.63*
Males	.68*	.63*	.65*
Females	.80*	.68*	.63*
Hispanic	1.08	.90*	.91*
Males	1.06	.92*	1.08
Females	1.23	.96	.98

* - Indicates significant decrease in variance after adjustment (ratio <0.95)

Table 7. SIPP Estimates of Persons 15+ Covered by Social Security (Numbers in Thousands)

MONTH	BEFORE	AFTER	BENCHMARK	AS PERCENT OF BENCHMARK	
				BEFORE	AFTER
* 1	32199	31814	32916	97.8%	96.7%
* 2	32389	32016	32945	98.3%	97.2%
* 11	32646	32348	33287	98.1%	97.2%
* 12	32594	32288	33240	98.1%	97.1%

* Indicates difference between estimates before and after is significantly different at the 0.10 level.

Table 8. SIPP Estimates of Persons 15+ Covered by SSI (Numbers in Thousands)

MONTH	BEFORE	AFTER	BENCHMARK	AS PERCENT OF BENCHMARK	
				BEFORE	AFTER
* 1	3284	2895	3379	97.2%	85.7%
* 2	3311	2917	3390	97.7%	86.1%
* 11	3532	3135	3460	102.1%	90.6%
* 12	3542	3147	3470	102.1%	90.7%

* Indicates difference between estimates before and after is significantly different at the 0.10 level.