CONTINUING RESEARCH ON USE OF ADMINISTRATIVE DATA IN SIPP LONGITUDINAL ESTIMATION

Suzanne M. Dorinski, U.S. Bureau of the Census[1]
U.S. Bureau of the Census, Washington, DC 20233

Key Words: raking ratio estimation, IRS income data, population controls

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 1984 sample that could be matched to IRS records. They adjusted the matched records to IRS-reported age, race, sex, and adjusted gross income. They did not control the nonmatched sample. Their adjustment produced a reduction in variances for most income and program participation variables.

Subsequent research by Dorinski and Huang [1994] 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. We found significant variance reduction, over previous research that did not adjust non-matched cases, for many of the variables examined.

Final results indicated 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.

Due to some of the adverse results for Black, Hispanic, and program participation estimates, we decided to research the methodology on a more recent panel before adding it to the current SIPP weighting procedure. We chose the 1990 SIPP panel because it contained an oversample of households headed by Blacks, Hispanics, or females with no spouse present living with children under age 18. We focused on the respondents for calendar year 1990.

The next section outlines the methodology used. The succeeding sections discuss the differences from the 1984 panel research, variance results and effects of the new weighting on the bias. The final section presents recommendations.

METHODOLOGY

The Census Bureau matched the 1990 SIPP panel file to the 1990 IRS Tax Year file. SIPP respondents 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) matched. We attached IRS extract data to the SIPP file. Approximately 55% 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.

When trying to use administrative records, several bias issues need to be resolved. The SIPP universe and the IRS universe are not equivalent. 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. Members of the military file tax returns, but aren't necessarily part of the SIPP universe. 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. Persons with incomes below the minimum filing requirements do not have to file tax returns. Previous research indicated that the total bias is no more than 2.4 percent for estimates of total population.

Since we are matching on SSN, we need to be aware of biases that may occur when respondents

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. I wish to thank Robert Fay, George Train and Karen King for their work on this project; Peter Siegel, Jeff Hayes, Easley Hoy and Ram Chakrabarty for their comments on this paper; and Sandy Carnegie for her help in preparing the final version.

refuse to provide SSN. Collection of SSNs is optional in SIPP. Respondents who refuse to provide SSN cannot be matched to IRS returns. Records for respondents who provide SSNs, say they don't know it, or claim not to have one are sent to the Social Security Administration for verification. Name, date of birth, race, and sex are used in the verification process. Records showing an SSN are sent through a computer match. The records that fail the computer match are then sent through a manual match. Records of respondents claiming not to know or have an SSN are also sent through a manual match.

Previous research suggests that weights are overadjusted for respondents who match to IRS returns. The overadjustment then caused the weights for nonmatched respondents to be underadjusted. Since we don't get SSNs for respondents who refuse to provide it, and the match to the IRS returns depends on SSN, we looked at demographic subgroups to see if any particular subgroup is more likely to refuse to provide SSN. The overall refusal rate was 5.1% for the 1990 panel. We defined "more likely to refuse" to be a rate of 5.6% or above. The rates are shown in Table 1. Personal total income of $20,000 to $30,000 per year had the highest SSN refusal rate.

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 last digit of SSN) to Social Security Administration records. From this file the age, race, and sex of the primary filers can be determined.

Census staff prepared tables from the 20-percent IRS sample as controls. The 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 records (age, race and sex). We prepared separate tables for each type of return (joint, single, and (nonjoint) household). We used these tables 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. We then calculated estimates of selected SIPP characteristics 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.

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.

DIFFERENCES FROM 1984 PANEL RESEARCH

One key difference from previous research is the weight used in the raking. The research on the 1984 panel used the final SIPP weight in the raking to IRS (for matched) and CPS (for nonmatched) controls. The SIPP final weight is initial weight * sample cut adjustment factor * noninterview adjustment factor * second stage adjustment factor.

For the 1984 panel research, we used (SIPP final weight*IRS adjustment) as the weight for estimates on matched cases, and (SIPP final weight*CPS adjustment) as the weight for estimates on nonmatched cases.

The second stage adjustment factor comes from the second stage ratio estimation performed in longitudinal weighting. The second stage ratio estimation currently used in SIPP weighting is composed of several rakes. For persons age 14 and under, the second stage is a Spanish adjustment, followed by an age adjustment by race and sex.

For persons 15 and above, blacks and nonblacks are handled separately in the second stage ratio estimation. The black and nonblack tables are based on age, sex, and household status. Both blacks and nonblacks are raked to CPS controls, then undergo a Spanish adjustment, then another rake to CPS controls, then another Spanish adjustment. At this point, Spanish origin persons are removed from further processing in the second stage ratio estimation. Both blacks and nonblacks then go through a final raking to CPS controls.

The research on the 1984 panel was done to see if raking to IRS controls was feasible. The results show that raking to IRS controls may improve survey estimates, so our research on the 1990 panel focuses on implementing the raking as we would in current SIPP weighting. If we add the IRS raking to current SIPP weighting procedures, we would probably do the IRS raking at the beginning of the second stage ratio estimation process.

Thus, in the current research, we used the SIPP 1990 pre-second stage weight for matched cases, which is initial weight * sample cut adjustment factor * noninterview adjustment factor.

Due to time constraints, we were unable to control the nonmatched cases to demographic controls, so we used the SIPP final weight for nonmatched cases to produce variance estimates.

We had planned to do the SIPP second stage ratio estimation for the matched cases after the IRS adjustment. Due to time constraints, we weren't able to finish that part of the research either. However, the IRS adjustment is a type of second stage adjustment -- we are raking to controls based on filing status, age, race, sex, Spanish surname, and adjusted gross income. So for matched cases, we used (pre-2nd stage weight*IRS adjustment factor) as the weight to produce variance estimates.

Another difference from previous research was how we derived IRS controls. The 1984 panel was controlled to IRS totals derived from a one-percent sample of IRS returns. In this research, we controlled to a 20-percent sample of IRS returns for increased reliability. The IRS controls from 1984 excluded returns from deceased taxpayers. Respondents who die are still part of the SIPP calendar year weighting, so this research used controls **with** deceased taxpayers.

The 1984 panel research used a 3-interview research file which contained data covering the period June 1983 - August 1984. The time period did not completely overlap with the 1984 IRS tax file. The current research focuses on calendar year 1990 data, which does coincide with the 1990 tax year data.

We used VPLX to compute the estimates and variances of income and program participation variables. VPLX is a computer program written by Robert Fay of the Census Bureau, which calculates the estimates and variances for totals, means, and proportions through replication methods. The system shares techniques of several standard methods of variance estimation and combines them together. (For more information on VPLX, see Fay [1990].)

VARIANCE RESULTS

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

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

If the ratio is 1.00, the adjustment has not changed the variance. If the ratio is less than 1.00, the adjustment has decreased the variance. We defined a ratio of less than 0.95 as useful, while a ratio of greater than 1.05 was not useful.

Table 2 shows reduction in sampling variances for most of the estimates studied. However, it should be noted that the variances for Hispanic females with annual incomes of $20,000 to $30,000 and $30,000+ actually increased. Previous research's problems [Dorinski and Huang 1994] for estimates of Black women with annual incomes of $20,000 to $30,000, $30,000+ and $20,000+ are now resolved.

Table 3 presents variance ratios for the estimated number of recipients for the following government programs: food stamps, Aid to Families with Dependent Children (AFDC), AFDC or General Assistance (AFDC/GA), Veterans' compensation, the Supplemental Food Program for Women, Infants and Children (WIC), Federal Supplemental Security Income (SSI), Social Security, and unemployment compensation. To be a recipient of a program, a person must have received benefits from the program one or more months.

Table 3 shows reduction in sampling variances for most of the estimates examined. Note that previous problems with estimates for Hispanics receiving food stamps, Hispanics receiving AFDC, Hispanics and Hispanic females receiving AFDC or General Assistance, Hispanics receiving WIC benefits, Blacks receiving Social Security, and Black men receiving unemployment compensation have been resolved.

Several demographic estimates are presented in Table 4. We found reduction in sampling variances for most of the estimates examined. Note that previous problems [Dorinski and Huang 1994] with estimates for Hispanic males ever married, males and Hispanics ever divorced, total population (male and female) ever separated and Hispanic females ever separated have been resolved. However, the adjustment has increased variances for estimates of Blacks ever separated.

Certain unemployment and employment characteristics are presented in Table 5. We found reduction in sampling variances for most of the estimates examined. Previous problems [Dorinski and Huang 1994] with estimates for Hispanics and unemployment estimates for Black males have been resolved.

From Table 6, we see that variance estimates for Hispanics have been improved. However, estimates of Blacks and Hispanics ever receiving property income continue to suffer from increased variances. The adjustment has also increased the variance for estimates of females ever disabled.

Finally, in Table 7, 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 studied. Previous problems [Dorinski and Huang 1994] for estimates of Hispanics below poverty all 12 months and percentage of months Hispanic males spent in poverty have been resolved. However, the variance for females below poverty all 12 months has increased.

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 covered by Social Security, the population covered by AFDC, the population covered by food stamps, and the population covered by SSI.

We studied SIPP estimates of persons covered by Social Security each month during 1990. The estimates before and after adjustment are not significantly different at the 0.10 level.

We looked at SIPP estimates of persons covered by AFDC. The estimates before and after adjustment are not statistically different.

Table 8 shows SIPP estimates of persons covered by food stamps. The before and after adjustment estimates are statistically different. The adjustment appears to have reduced the bias of the estimates.

We studied SIPP estimates of persons covered by SSI. The estimates before and after adjustment are not statistically different.

RECOMMENDATIONS

We recommend that the research on the adjustment of the matched population to IRS controls and the unmatched population to adjusted Census/CPS controls continue. The results so far look promising. If further results are still good, we may try to adapt the methodology to routine SIPP production weighting. However, the methods will have to be greatly simplified for production.

There are still several bias issues that need to be addressed before the system could be adopted. We still haven't found an adequate way to adjust the IRS controls to exclude the military and institutionalized filers, who aren't a part of SIPP's universe. The future quality of SSA's race data is uncertain. [In some states, SSNs are now assigned at birth from the generation of the birth certificate, but the states are treating race as confidential data, so SSA isn't getting the race of the individual linked to the SSN.] Hence we may not be able to depend on SSA-reported race in the IRS raking. We would like to have SSN refusals go through the manual search SSN validation, but that may not be possible due to privacy/confidentiality concerns.

The Committee on National Statistics has recommended that SIPP become the official vehicle for measuring poverty in the United States. If and when that happens, there may be a debate about the true benefit of this adjustment if the results hurt our poverty estimates. We may have to look for other ways to adjust the poor and near-poor cases.

REFERENCES

Dorinski, Suzanne M. and Hertz Huang (1994), "Use of Administrative Data in SIPP Longitudinal Estimation," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 656-661.

Fay, Robert E. (1990), "VPLX: Variance Estimates for Complex Samples," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 266-271.

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.

Table 1. SSN Refusal Rates for SIPP 1990 Wave 1 Respondents

Demographic characteristic	Refusal rate
Black	6.1%
Age 40-49	6.0%
Age 50-59	6.0%
Age 60-69	6.3%
Personal earnings $10,000 - $20,000 per year	5.8%
Personal earnings $30,000+ per year	5.9%
Personal total income $10,000 - $20,000 per year	9.1%
Personal total income $20,000 - $30,000 per year	9.5%

Table 2. 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	.75*	.67*	.79*	.54*	.56*	.62*
Male	.75*	.76*	.81*	.63*	.62*	.65*
Female	.65*	.76*	.82*	.61*	.62*	.71*
Black	.79*	.73*	.90*	.78*	.58*	.87*
Male	.87*	.76*	.93*	.81*	.73*	.90*
Female	.62*	.73*	.95	.90*	.61*	.72*
Hispanic	.73*	.88*	.66*	.83*	.66*	.72*
Male	.76*	.99	.78*	.83*	.69*	.73*
Female	.82*	.91*	1.12	1.11	1.03	.85*

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

Program Participation

	Recipient for One or More Months							
	FOOD	AFDC	AFDC/GA	Vets	WIC	SSI	OASDI	UNEMP
Total	.91*	.93*	.94*	.97	.83*	.97	.42*	.94*
Males	1.02	.96	.98	.99	-	.98	.44*	1.11
Females	.90*	.96	.95	.98	.83*	1.03	.49*	.89*
Black	.79*	.81*	.80*	1.03	.78*	.97	.75*	.85*
Males	.91*	.82*	.84*	1.13	-	.95	.76*	.87*
Females	.79*	.89*	.85*	.99	.81*	.98	.76*	.86*
Hispanic	.85*	.81*	.87*	.96	.76*	.82*	.73*	.99
Males	.74*	.86*	1.08	1.00	-	.71*	.70*	.85*
Females	.86*	.79*	.76*	.86*	.77*	.84*	.82*	.99

Table 4. Ratios of Estimated Variances After and Before Adjustments to Administrative Data

Marital Status

	Ever Married	Ever Divorced	Ever Separated
Total	.58*	.88*	.89*
Males	.71*	.93*	.94*
Females	.47*	.82*	.94*
Black	.77*	.95	1.15
Males	.84*	1.05	1.14
Females	.76*	1.00	1.08
Hispanic	.80*	.90*	.96
Males	.74*	1.02	1.17
Females	.94*	.89*	.85*

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

Table 5. **Ratios of Estimated Variances After and Before Adjustments to Administrative Data**

	Unemp 1	Unemp 2	Emp 1	Emp 2
Total	.91*	.94*	.71*	.70*
Males	.92*	.94*	.62*	.61*
Females	.89*	.93*	.79*	.79*
Black	.84*	.90*	.85*	.85*
Males	.93*	.99	.81*	.82*
Females	.77*	.86*	.93*	.93*
Hispanic	.82*	.79*	.87*	.90*
Males	.71*	.67*	.79*	.79*
Females	.96	.94*	.86*	.88*

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 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) or (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 6. **Ratios of Estimated Variances After and Before Adjustments to Administrative Data**

	Ever Disabled	Ever Received Wages or Salary	Ever Received Property Income
Total	.99	.73*	.86*
Males	.93*	.75*	.79*
Females	1.09	.75*	.97
Black	.92*	.86*	1.07
Males	.91*	.80*	1.09
Females	1.00	.92*	1.22
Hispanic	.86*	.93*	1.24
Males	.76*	1.00	1.36
Females	.92*	.86*	1.11

Table 7. **Ratios of Estimated Variances After and Before Adjustments to Administrative Data**

	Below Poverty for All 12 Months	Below Poverty for At Least One Month	Months in Poverty
Total	1.01	.83*	.83*
Males	.88*	1.00	.86*
Females	1.06	.78*	.87*
Black	.88*	.76*	.76*
Males	.87*	.87*	.80*
Females	.89*	.76*	.79*
Hispanic	.80*	.78*	.73*
Males	.68*	.81*	.69*
Females	.84*	.80*	.80*

Table 8. **SIPP Estimates of Persons Covered by Food Stamps (Numbers in Thousands)**

MONTH	BEFORE ADJUSTMENT	AFTER ADJUSTMENT	BENCHMARK	AS PERCENT OF BENCHMARK	
				BEFORE	AFTER
**JAN	16,251	16,668	19,849	82%	84%
**NOV	16,937	17,320	21,294	80%	81%
**DEC	16,865	17,202	21,687	78%	79%

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

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