INCOME STRATIFICATION IN PANEL SURVEYS: ISSUES IN DESIGN AND ESTIMATION

John L. Czajka, Mathematica Policy Research, Inc.
600 Maryland Avenue, S.W., Suite 550, Washington, DC 20024-2512

KEYWORDS: Sample Selection; Weighting

1. INTRODUCTION

Panel surveys or administrative record panels that include among their primary objectives the collection of data on personal or household income over time frequently employ some form of differential selection by income level. Over time the incomes of panel sample members can change--perhaps dramatically. With the accumulation of changes at the micro level, the income composition of the panel sample may change as well. Why is this significant? If the users of panel data limited their research to those questions that a panel is designed to address, then changes in the income composition of the sample need not present serious problems. But users do not limit their research to such questions--namely, questions that can be answered by: (1) longitudinal analyses that (2) start at the base period. Instead, researchers and policymakers often use panel data to develop cross-sectional estimates for the panel out-years or to conduct longitudinal analyses that begin well after the initial interview.

This paper addresses several issues related to the design and use of panel surveys that employ differential selection by income level. Data from the 1985 Sales of Capital Assets (SOCA) Panel, a sample of administrative records compiled by the U.S. Internal Revenue Service (IRS), are used to provide empirical illustrations of alternative strategies for panel survey design and estimation.

2. THE SOCA PANEL

The 1985 SOCA Panel is a subsample of the 1985 Statistics of Income (SOI) sample of individual tax returns. Each year the SOI Division of the IRS selects a large, representative sample of the tax returns processed during the calendar year. The sample is highly stratified by income, and returns are selected by simple random sampling, in effect, within each stratum. Every return in the population is "looked at," assigned to a stratum, and then either selected or not, based on the comparison of a random number to the stratum sampling rate. The 1985 sample design included 33 strata, with sampling rates varying from .03 percent in the lowest income class to 100 percent in the highest income class and certain specialized classes of returns. Sample selection yielded 121,418 returns (IRS 1988).

The SOCA Panel consists of 12,980 filing units selected as a stratified probability sample of this much larger cross-sectional sample. High income strata were subsampled at higher rates than low income strata. Average selection probabilities by stratum ranged from .0014 percent in the lowest income stratum to between 20 and 39 percent in the SOI certainty strata.

From these 12,980 base year returns, all primary and secondary taxpayers, identified by their social security numbers (SSNs), were designated as members of the 1985 SOCA Panel. In each subsequent processing year, every return processed by the IRS that contained a SOCA SSN in either the primary or secondary position was selected for inclusion in the SOCA Panel. In addition to the data items captured for the annual SOI sample, records of individual transactions from sales of capital assets were collected as well. These transaction data generate interest in using the SOCA data for annual cross-sectional estimation as well as longitudinal analysis.

3. INCOME DYNAMICS IN A PANEL SAMPLE

For a panel of any duration long enough to be "interesting," the dynamics of personal and household income will act to change the distribution of sample households with respect to income level. Two phenomena induce such "panel drift" generally, although others may be operative in special cases. First, the panel ages, and this implies a drift toward higher income levels over time. Second, the statistical phenomenon of regression to the mean produces a contraction of the distribution over time, with the most extreme observations tending to show the greatest movement toward the center.

Differential selection may amplify the changes induced by these phenomena. Oversampling one class relative to another alters the gross flows between them. Consider, for example, a design that oversamples the tails of the income distribution. The greater the differentiation among sampling rates between the tails and the neighboring strata, the more the units exiting the tails will outnumber the units moving *into* the tails. Over time, therefore, differential selection with respect to income class will affect the net change in class size, the composition of classes (in terms of the class of origin), and the variation in base year selection within classes.

Of what significance is panel drift? Essentially, if there are important benefits to selecting a base year sample that deviates from the composition of the population that it represents, then change in the composition of the sample over time implies some reduction in these benefits. In designing a panel sample with differential selection probabilities by income level, therefore, it is important to take into consideration prospective changes in the income composition of the sample over time. Otherwise the useful life of the panel may be foreshortened.

4. PANEL SAMPLE DESIGN

Czajka and Schirm (1993) outlined several approaches to designing a stratified panel sample with allowance for change in composition. These techniques can be grouped under three general strategies. The first involves selecting a panel at the middle of its intended life, then collecting data retrospectively and prospectively. The second involves backcasting from a desired mid- or end-life composition to a base year composition that will generate the desired end result. This strategy presumes a knowledge of transition probabilities between classes. For income and a number of other characteristics there exists fairly extensive knowledge about transitions, owing to earlier panel studies. The third strategy involves selection based on longitudinal characteristics. In other words, if it were possible to stratify on properties of units over the life of the panel, rather than just a single point in time, how might one do so? Examples of longitudinal characteristics include cumulative or permanent income, conditional transitions, and events. To implement a design based on such characteristics requires an ability to assign class membership on the basis of data available at selection.

The second general strategy is illustrated here with data from the SOCA Panel. The distribution of 1991 returns by 1991 AGI class within each 1985 AGI class was used to estimate the 1991 sample counts, by AGI, that would result from a given base year sample design--that is, sample sizes by 1985 AGI class. Table 1 presents the results of four simulations using alternative 1985 sample designs with 13,000 returns.

The first simulation assumes that equal numbers of returns are drawn from all nine strata (to produce whole numbers for sample counts, one additional return is allocated to each of four strata). This design yields a 1991 sample of 13,026 returns with the distribution shown in the table. With this design the highest income class drops from 1,444 returns in 1985 to 509 returns in 1991 while the lowest income class grows from 1,444 to 1,725 returns.

Table 1. 1985 Base Year Sample Allocation and Projected 1991 Sample Distribution by AGI Class under Alternative Base Year Designs

Absolute AGI ($1,000s)	1985 base year allocation	Projected 1991 distribution
	Equal stratum size	
0 to < 25	1,444	1,725
25 to < 50	1,445	1,754
50 to < 100	1,444	2,419
100 to < 200	1,445	1,786
200 to < 500	1,444	2,063
500 to < 1,000	1,445	1,159
1,000 to < 2,000	1,444	964
2,000 to < 5,000	1,445	647
5,000 or more	1,444	509
Total	13,000	13,026
	Probability proportional to size	
0 to < 25	8,603	5,504
25 to < 50	3,306	3,772
50 to < 100	906	2,208
100 to < 200	135	355
200 to < 500	39	83
500 to < 1,000	8	21
1,000 to < 2,000	2	7
2,000 to < 5,000	1	1
5,000 or more	0	0
Total	13,000	11,951
	Mock SOI design	
0 to < 25	2,328	2,310
25 to < 50	1,953	2,207
50 to < 100	1,638	2,658
100 to < 200	1,364	1,684
200 to < 500	1,995	1,923
500 to < 1,000	1,788	961
1,000 to < 2,000	1,153	705
2,000 to < 5,000	604	305
5,000 or more	177	152
Total	13,000	12,905
	Mock SOCA design	
0 to < 25	1,620	1,902
25 to < 50	1,598	1,957
50 to < 100	1,649	2,622
100 to < 200	1,513	1,873
200 to < 500	2,764	2,248
500 to < 1,000	1,503	1,043
1,000 to < 2,000	1,090	757
2,000 to < 5,000	935	375
5,000 or more	328	208
Total	13,000	12,985

The second simulation assumes that the nine income strata are sampled with probability proportional to size. This design yields no 1985 returns in the top income class and only one return in the next lower income class. The final 1991 sample totals 11,951 returns, and the lowest income class declines in size while the $50,000 to $100,000 class more than doubles in size.

The third and fourth simulations approximate the 1985 SOI cross-sectional sample design and the SOCA Panel design, respectively. The mock SOCA design yields a 1991 sample that is very close to the base year sample in size while the mock SOI design yields 80 fewer returns. The mock SOCA design generates more returns than the mock SOI design in each of the top three income classes, but it requires proportionately larger numbers of base year returns in the top two classes to do so.

With the information contained in the SOCA income class transitions between 1985 and 1991 one could attempt to devise a sample design that would yield a given distribution up to six years later.

5. ADJUSTING CROSS-SECTIONAL WEIGHTS FOR NEW UNIT MEMBERS

If a panel unit adds a new member, the cross-sectional weight for that time period (and all subsequent time periods when that member is present) must be adjusted to reflect the new member's independent probability of selection. For a unit of two persons, the maximum number of filers on a tax return, the appropriate weight is given by:

$$(1)\quad w_{12} = 1/[(1/p_1) + (1/p_2) - (1/p_1)^*(1/p_2)]$$

where p_1 is the base year selection probability of the original panel member and p_2 is the base year selection probability of the new unit member. This formulation assumes that the partners' probabilities of selection in the base year are independent under the sample design, so that the product term expresses the partners' joint selection probability. If the new partner was not eligible for selection in the base year, the addition of this new member has no impact on the unit weight; the weight is identical to what it would be if the panel member had remained single.

Unless a panel survey collects retrospective data for all new members, p_2 is generally unknown. When the base year selection probabilities have little variation, the only information required to calculate an approximately correct weight with equation (1) is whether the new member was indeed eligible for selection in the base year. When the base year selection probabilities vary widely, however, a method of handling the unknown selection probabilities of new members is required to calculate correct weights.

To calculate cross-sectional weights for the SIPP and other panel surveys, researchers at the Census Bureau developed a weighting scheme that assigns a person weight of zero to all new members and then calculates the unit weight as the average weight of all persons in the unit. This "equal person" weighting scheme (the term suggested by Kalton and Brick 1994) yields unbiased estimates of population totals, providing that new members who were not eligible for selection in the base year can be identified. The weighting scheme adds variance to population estimates, however, precisely because the new members' selection probabilities are unknown. The amount of variability introduced into the estimates of population totals depends on the variability of the base year selection probabilities and on their relatedness to the characteristics being estimated.

Table 2 shows the weights that would be assigned to a joint tax return filed in year X by a panel member who was single in year $X-1$ if the base year selection probabilities of both partners were known and equation (1) could be applied. The weights reflect base year selection probabilities by income class that are similar to those of the SOCA Panel.

The first column indicates the weights that would be assigned to units with no new members, or with new members who were nonfilers and therefore not eligible for selection in 1985. For a unit with a new member who filed in 1985 (or was presumed to have done so), the equal person weighting scheme would assign a weight equal to one-half the weight in the first column. Generally, this approximates the correct weight if the new spouse belonged to the same income class in 1985 as the panel member. This becomes less true as the panel member's base year selection probability rises because the joint selection probability becomes nonnegligible.

For a panel member selected from a very low income class (high weight) there is potential for the equal person scheme to assign a weight that is very wide of the mark. Among panel members from the lowest 1985 income class, the weights vary from 31,250 (if the spouse belonged to the same 1985 income class) to 4.6 (if the spouse belonged to the highest 1985 income class) whereas the equal person weighting scheme always assigns a weight of 31,250. For a panel member selected from the highest income class, the potential error in the equal person weight assignment is much smaller. Depending on the new spouse's 1985 income class, the correct weights vary between 4.2 and 7.8, with the equal person weighting scheme assigning a weight of 3.9.

Table 2. Tax Return Weights Implied by Combinations of Filing Partners' Base Year Selection Probabilities

Absolute value of panel member's AGI in 1985	Absolute value of new spouse's AGI in 1985 ($1,000s)								
	Non-filer	0 to < 25	25 to < 50	50 to < 100	100 to < 200	200 to < 500	500 to < 1,000	1,000 to < 2,000	2,000 or more
0 to < 25	62500.0	31250.3	16666.9	6493.6	1351.4	387.5	51.2	16.3	7.8
25 to < 50	22727.3	16666.9	11363.9	5494.7	1302.1	383.3	51.1	16.3	7.8
50 to < 100	7246.4	6493.6	5494.7	3623.4	1160.2	370.0	50.9	16.2	7.8
100 to < 200	1381.2	1351.4	1302.1	1160.2	690.6	304.2	49.4	16.1	7.8
200 to < 500	389.9	387.5	383.3	370.0	304.2	195.2	45.4	15.7	7.7
500 to < 1,000	51.2	51.2	51.1	50.9	49.4	45.4	25.9	12.5	6.9
1,000 to < 2,000	16.3	16.3	16.3	16.2	16.1	15.7	12.5	8.4	5.5
2,000 or more	7.8	7.8	7.8	7.8	7.8	7.7	6.9	5.5	4.2

Table 2 makes apparent that the amount of variability introduced by the equal person weighting scheme will depend not only on the variability of the base year selection probabilities but also on the correlation between the base year probabilities of panel members and the persons added to their units. If the persons added to units always came from the same base year income class as the original panel members, then the weights assigned by the equal person weighting scheme would lie very close to the correct weights.

To what extent is this condition satisfied? Table 3 displays for each base year income class the probability of a given increase (or reduction) in absolute AGI associated with a transition from single to married.

Among panel members whose base year incomes placed them in the lowest income class, nearly 3.5 percent experienced an increase of \$50,000 or more in conjunction with the transition from single to married. For these units, using the change in AGI as an estimate of the new spouse's base year income, together with equation (1), yields cross-sectional weights between 387.5 and 6,493.6 rather than the 31,250 assigned by the equal person scheme.

As the incomes of panel members rise, the weights implied by the equal person weighting scheme appear to become even less consistent with marriage behavior. For example, among panel members with base year incomes between \$100,000 and \$200,000, fewer than 6 percent experienced changes in income consistent with the new spouse having income in the same range.

This exercise also reveals some of the risks associated with attempting to estimate the 1985 selection probability of the spouse on the basis of the change in AGI between two returns. Table 3 reflects a high frequency of instances in which the net change was negative. While a decline in income may indeed be attributable to the new spouse, the likelihood is greater that such a change is due to a reduction in the panel member's income. The fact that so many such cases occur underscores the fact that even plausible changes often may not accurately reflect the new spouse's income. Of course, even when the change accurately reflects the spouse's current income, this amount becomes over time a less accurate proxy for the spouse's base year income.

The findings presented here provide an argument for collecting from all new members of a panel whatever information is sufficient to estimate their base year selection probabilities. This will make it possible to consider the application of a weighting scheme that introduces less variability than the equal person scheme. In the absence of such data, there may be value in using whatever information is available to *estimate* the base year probabilities of persons who join panel units--particularly units with relatively high weights. Weights calculated on the basis of such information will introduce some bias, but if the base year selection probabilities are sufficiently varied these alternative weights have the potential to produce estimates with lower mean squared error than weights based on the equal person method.

6. REPRESENTATION OF THE POPULATION

If properly weighted, panel data can be used to provide out-year, cross-sectional estimates for the survivors of the base year population that the panel was selected to represent. Frequently, however, there is interest in using panel data to develop out-year estimates for the *entire* population, including segments from which the panel sample includes no observations--namely, persons who were not eligible for selection in the base year and who have not joined units that *were* eligible for selection.

Post-stratification to known population totals is often used with panel data to generate cross-sectional estimates that apply to the full population.

Table 3. Probability of a Given Increase in Absolute Income Associated with Marriage, Conditional on Base Year Income

		Increase in absolute AGI ($1,000s)						
Panel member's base year absolute AGI	Reduction	0 to < 25	25 to < 50	50 to < 100	100 to < 200	200 to < 500	500 to < 1,000	1,000 or more
0 to < 25	6.59	71.52	18.39	2.74	.51	.24	.00	.00
25 to < 50	6.11	43.34	36.43	10.61	1.45	1.54	.00	.51
50 to < 100	12.27	30.61	28.48	14.34	8.91	2.07	1.00	2.33
100 to < 200	33.85	11.04	13.26	22.53	5.92	12.33	1.17	.00
200 to < 500	44.87	17.65	3.60	5.20	7.99	14.46	4.12	2.11
500 to < 1,000	51.60	1.93	6.30	4.51	9.90	10.55	7.82	7.39
1,000 to < 2,000	42.48	6.21	1.35	13.94	3.04	9.43	7.65	15.91
2,000 or more	57.87	.00	1.78	3.95	1.78	6.90	10.06	17.66

While the panel sample will then reproduce the population totals and provide more precise estimates of characteristics that are related to the post-stratifying variables, estimates of characteristics that are *not* correlated with the post-stratifying variables may still be biased significantly. Supplementing the panel sample with additional units during the out-years is an alternative strategy that provides a means to achieve true cross-sectional representativeness, in theory, but this tactic is often not very practical.

The top panel of Table 4 reports SOI cross-sectional sample estimates of the entire filing population (total tax returns) by AGI for the years 1985 through 1991. The middle panel reports SOCA Panel estimates of the total tax returns filed by the survivors of the 1985 filing popuiation. The bottom panel reports the percentage difference between the SOCA estimate for each year and AGI class and the SOI cross-sectional estimate of the entire filing population.

By 1991 the number of returns filed by survivors of the 1985 filing population has fallen to under 82.5 percent of the base year population. Mortality will account for about a 1 percent loss per year, and late filing will subtract a few percent by 1991. The rest of the decline is due to nonfiling--primarily by persons whose incomes dropped below the filing limits.

Over time there are striking differences between the two sets of distributions by AGI class. While the population with current year AGI between $25,000 and $50,000 remains essentially constant in size, the income classes above that level grow to between two and three times their 1985 sizes. By contrast, all of the income classes below $25,000 exhibit losses in size between 1985 and 1991, with the greatest loss occurring in the lowest positive income class, which drops to less than one-third its 1985 size.

During the same period the total filing population grew by about 13 percent so that by 1991 the returns represented by the SOCA Panel accounted for barely 73 percent of all returns. The shortfall was heavily concentrated among low income returns. The SOCA estimate of returns in the lowest positive income class falls short of the full population estimate by nearly 70 percent.

The pattern of differences between the SOCA population estimates and the cross-sectional sample provides a strong argument for the consideration of periodic sample supplementation if the SOCA Panel is to be used on a regular basis to develop inferences about the entire filing population. A weakness of sample supplementation, generally, is the difficulty of identifying prospective sample members who have entered the population of interest since the year the panel was selected. Unless it is feasible to screen prospective new observations so as to eliminate those who would have been eligible for selection in the base year, most of the units selected in a supplemental sample will belong to population segments that are already represented. Table 4 shows that in some of the low income strata more than half of the 1991 population lies outside the SOCA universe. If a supplemental sample of returns from these strata were added to the SOCA Panel, therefore, more than half of the new observations would consist of persons who did not file in 1985--a very successful hit rate for a low cost selection scheme.

ACKNOWLEDGMENTS

This research was performed under contract to the Statistics of Income (SOI) Division of the Internal Revenue Service. The author is grateful for this support and would like to thank Susan Hostetter in particular for her encouragement of methodological research on estimation with panel data. At MPR, Daisy Ewell was responsible for all programming and research support, and the author gratefully acknowledges her contributions.

Table 4. Comparison of SOCA Panel and SOI Cross-Sectional Sample Estimates of Filing Population, 1985-1991

	Tax year						
AGI ($1,000s)	1985	1986	1987	1988	1989	1990	1991
Cross-sectional sample estimate of total filing population (thousands of returns)							
Total returns	101,660	103,045	106,996	109,708	112,136	113,717	114,730
No AGI ($0 or less)	1,035	957	842	835	823	905	926
> 0 to < 5	15,714	15,988	16,974	17,050	16,769	16,478	16,069
5 to < 10	16,492	15,910	15,698	15,402	15,007	14,953	15,229
10 to < 25	34,527	34,218	34,291	34,754	35,374	35,038	35,063
25 to < 50	25,795	26,507	26,962	27,739	28,306	28,958	29,037
50 to < 100	6,892	7,975	10,175	11,425	12,981	14,220	14,962
100 to < 200	909	1,116	1,514	1,778	2,090	2,330	2,598
200 to < 500	238	291	430	548	613	644	676
500 to < 1,000	41	52	75	115	116	130	118
1,000 or more	17	32	35	62	58	61	52
SOCA Panel estimate of survivors of the 1985 filing population (thousands of returns)							
Total returns	101,660	95,645	91,999	91,251	88,617	87,500	83,869
No AGI ($0 or less)	817	813	600	499	646	546	604
> 0 to < 5	15,532	10,246	8,882	7,835	5,470	5,640	4,840
5 to < 10	16,070	14,410	12,330	10,171	8,697	8,604	7,352
10 to < 25	35,230	34,274	31,087	32,312	31,441	28,828	27,454
25 to < 50	25,948	26,557	26,604	26,743	26,963	27,136	25,335
50 to < 100	6,855	7,829	10,511	11,215	12,684	13,685	15,066
100 to < 200	915	1,180	1,464	1,734	1,945	2,347	2,469
200 to < 500	233	253	406	549	588	510	561
500 to < 1,000	42	56	78	122	131	152	130
1,000 or more	17	28	35	70	52	50	56
Percentage deviation: (SOCA less cross-section)/cross-section							
Total returns	.0%	−7.2%	−14.0%	−16.8%	−21.0%	−23.1%	−26.9%
No AGI ($0 or less)	−21.1	−15.0	−28.7	−40.2	−21.5	−39.7	−34.8
> 0 to < 5	−1.2	−35.9	−47.7	−54.0	−67.4	−65.8	−69.9
5 to < 10	−2.6	−9.4	−21.5	−34.0	−42.0	−42.5	−51.7
10 to < 25	2.0	.2	−9.3	−7.0	−11.1	−17.7	−21.7
25 to < 50	.6	.2	−1.3	−3.6	−4.7	−6.3	−12.7
50 to < 100	−.5	−1.8	3.3	−1.8	−2.3	−3.8	.7
100 to < 200	.7	5.7	−3.3	−2.5	−6.9	.7	−5.0
200 to < 500	−2.1	−13.1	−5.6	.2	−4.1	−20.8	−17.9
500 to < 1,000	2.4	7.7	4.0	6.1	12.9	16.9	10.2
1,000 or more	.0	−12.5	.0	12.9	−10.3	−18.0	7.7

REFERENCES

Czajka, John L. and Allen L. Schirm. "Selection and Maintenance of a Highly Stratified Panel Sample." *Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys.* Ottawa, Canada: Statistics Canada, 1993.

Internal Revenue Service. *Statistics of Income--1985: Individual Income Tax Returns.* Washington, DC: U.S. Government Printing Office, 1988.

Kalton, Graham and J. Michael Brick. "Weighting Schemes for Cross-Sectional Analyses of Household Panel Surveys." *Proceedings of the Survey Research Methods Section.* Alexandria, VA: American Statistical Association, 1994.