Annals of Economic and Social Measurement, 2/2, 1973

EDITING CENSUS MICRODATA FILES FOR INCOME AND WEALTH

BY NELSON MCCLUNG[1]

Misreporting, which in general is underreporting, of income on Census surveys has the consequence for microsimulations run on survey files that we overestimate the number of families in poverty, underestimate the antipoverty effectiveness of existing programs, overestimate the budget cost and coverage of new programs and underestimate income taxes computed from file income. In regressions which introduce file total or component incomes as dependent or independent variables, parameter estimates are biased. The objective of this edit is to reduce these errors by adjusting reported incomes to yield weighted aggregates which are close to those estimated by other and presumably more reliable sources.

There are two characteristics of our procedure which may in some uses bias the adjusted data, as compared to a perfect CPS:[2] in general the procedure assumes that (1) misreporting of income from one source is independent of misreporting of income from another source and (2) receipt of income from one source is weakly independent of receipt of income from another or other sources. With respect to (1), we recode excess 1967 SEO government pensions as OASDI or veterans' benefits. But the real issue is whether interview units have different but consistent propensities to misreport. If they do, we do not have the information needed to take that fact into account. Giving everyone in some class an equal chance to shift position in the income distribution, we may on the average move the wrong people. With respect to (2), we recognize that receipts of large amounts of government and private employee pensions, for example, are implausible but do not recognize the strong interdependence, again for example, between interest and dividend incomes apart from an Adjusted Gross Income control.

1. METHODOLOGY

As we practice it, income editing proceeds in two steps: first, given an apparent discrepancy, we infer the evident discrepancy between a CPS or SEO file estimate and a reference estimate; second, we develop a rule for adjusting CPS or SEO respondent reported amounts of each type of income so that file distributions resemble reference distributions in as many dimensions as possible.

[1] Of the people working on The Urban Institute TRIM project who have made contributions to this file edit, the two to whom I am most indebted are Lou Koènig of The Urban Institute and Charlotte Barkerding of the Hendrickson Corp.

[2] For a discussion of how biases may be introduced into the artificial sample in the process of eliminating bias in certain control aggregates, see Benjamin Okner, "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File," comments and rejoinders, *Annals of Economic and Social Measurement*, Vol. 1, No. 3 (July 1972), pp. 325–362.

1.1. *Inferring Discrepancies*

Discrepancies between reference and CPS/SEO file estimates of income and of income recipients reflect differences in (1) concepts, (2) units of observation, (3) timing, (4) geographic coverage, (5) demographic coverage, (6) accuracy of data collection and edit procedures and (7) sampling error. Starting with the initial apparent discrepancy, we adjust the reference date to comparability with Census survey data. For each source of income, we consider making reconciliation adjustments for (1) mortality, (2) institutionalization, (3) foreign residence, (4) differences in reporting units, (5) recidivism and (6) income screens. Where we do not have end-of-year counts, reference numbers of recipients and amounts should be reduced by deaths between sometime in the survey income year and the date of the CPS or SEO survey. The institutionalized and foreign resident, like the dead, have a zero probability for inclusion in the CPS or SEO samples and should be cast out of the reference data. In general, we count CPS or SEO interview units but reference counts typically are in other units and the two counts should be reconciled by an adjustment for multiple recipients in interview units. Recidivism is an adjustment which needs be made only for Unemployment Compensation and Public Assistance because our reference counts are not of units on the rolls anytime during the year. Statistics of Income counts are low relative to CPS and SEO counts by the number of units not required to file returns who in fact do not. Both counts of units and of income are too low by illegal nonreporting and underreporting but we make no adjustment for that. After many months, the Commissioner of Internal Revenue still has not responded to our request for gross average Taxpayer Compliance Measurement Program audit results. In the procedure reported here, we make no adjustments for errors in Census Survey data other than misreporting or nonreporting of economic data. In a rare rich adjustment, we subtract from the SEO reference counts units and amounts which have AGI greater than or equal to $50,000 and double the sample weights of units in the CPS reporting AGI $\geq$ $50,000 and divided income. The 1971 CPS has, as the 1967 SEO does not, a reasonable representation of families with incomes over $50,000.

1.2. *Rectification*

There are three very elementary decision rules for record adjustment. (1) If the number of CPS or SEO file recipients of a type of income agrees with the reference count but the amount of income reported by all or some subset of them is short, reported income on each record may be increased by ratios of aggregate reference to aggregate file reported income. (2) If the number of file recipients and their reported income both are less than reference counts, then additional units equal in number to those missing may be selected from among file units reporting zero receipts of income from that source and assigned amounts of income which make the weighted file counts equal to reference counts. (3) If the number of file recipients is less than reference but the counts of income agree, then additional file units may be selected from zero reporters and positive or negative amounts of income assigned to which make weighted file counts agree with reference. This is, of course, a conceptually complete list only on the assumption that SEO or

CPS counts are less than or equal to reference counts; with a few exceptions, that is the case.

We have generalized these basic rules somewhat, although in outlining the generalization it is convenient to continue the assumption that aggregate CPS or SEO counts of recipients or of income are less than or equal to reference. The procedure provides for (1) selecting interview units for imputation of type x income and (b) allocating units selected to type x income size classes. Let y be total income divided into $i = 1, \ldots, n$ size classes; x be income from a particular source divided into $j = 1, \ldots, m$ size classes; s_{ij} be SEO or CPS interview units in the i, j income size class; r_{ij} be the reference units in the i, j income size class. Then we may construct a table which distributes CPS or SEO interview units by joint income size classes, size of total income and size of income from a particular source:

	x_0	$\sum_j s_{ij}$	x_1	x_2	x_3	x_4
y_1	s_{10}	$\sum_j s_{1j}$	s_{11}	s_{12}	s_{13}	s_{14}
y_2	s_{20}	$\sum_j s_{2j}$	s_{21}	s_{22}	s_{23}	s_{24}
y_3	s_{30}	$\sum_j s_{3j}$	s_{31}	s_{32}	s_{33}	s_{34}
y_4	s_{40}	$\sum_j s_{4j}$	s_{41}	s_{42}	s_{43}	s_{44}

The interview units s_{i0} are those reporting zero receipts of type x income. The $\sum s_{ij}$ may extend over positive and negative income size classes and the y_i classes do not necessarily range over only positive total incomes. We distribute file units for comparison with reference units.

	$\sum_j r_{ij}$	x_1	x_2	x_3	x_4
y_1	$\sum_j r_{1j}$	r_{11}	r_{12}	r_{13}	r_{14}
y_2	$\sum_j r_{2j}$	r_{21}	r_{22}	r_{23}	r_{24}
y_3	$\sum_j r_{3j}$	r_{31}	r_{32}	r_{33}	r_{34}
y_4	$\sum_j r_{4j}$	r_{41}	r_{42}	r_{43}	r_{44}

The quantity $\sum_{i,j} s_{ij} / \sum_{i,j} r_{ij}$ is the CPS or SEO population of type x income recipients relative to the reference population. If $\sum_{i,j} s_{ij} / \sum_{i,j} r_{ij} < 1.0$, $\sum_j s_{ij} / \sum_j r_{ij} \gtreqless 1.0$. If either the file overall sum or a file total income size class sum exceeds the corresponding reference sum, application of the procedure is complicated somewhat. On the 1967 SEO file, $\sum_{i,j} s_{ij} / \sum_{i,j} r_{ij} > 1.0$ for only interest and government pensions; the latter is rather obviously the consequence of misreporting of income by type and the former a consequence primarily of an excessive multiple recipient adjustment. $\sum_j s_{ij} / \sum_j r_{ij} > 1.0$ for some classes of self-employment and rent income

recipients and that may be attributed to deficiencies in the Statistics of Income concepts. Apart from these exceptions, the differences $\sum_j r_{ij} - \sum_j s_{ij}$ are the numbers of interview units to be selected from the y_i income classes for imputation of type x income. The number of interview units in a total income class available for selection is $s_{i\emptyset}$; this is the number for which there is no record of type x income. The fraction of units to be selected is $(\sum_j r_{ij} - \sum_j s_{ij})/s_{i\emptyset} = p_i$.

We could choose the $p_i s_{i\emptyset}$ units by purely random selection; to do so, however, is to disregard information which we have. We stratify the $s_{i\emptyset}$ by k attributes which we know are associated with receipt of type x income. For simplicity of notation k ranges over kinds of attributes and values of each; that is, it is a two-dimensional index. If nonreporters are not known to differ from reporters, we select the $p_i s_{i\emptyset}$ units from the k classes of $s_{i\emptyset}$ so as to preserve the frequencies of occurrence of k attributes among CPS/SEO recipients of type x income; otherwise, we discriminate in selection so as to obtain the reference distribution.

As each file interview unit from an i, k class is selected for imputation of type x income, it is assigned an amount of type x income. The process of assigning amounts we call j classing because we in effect allocate fractions of the $s_{i\emptyset}$ to cells in the j classification. Allocation of interview unit income to persons within a unit is done by a TRIM[3] procedure designed for this purpose. We could allocate selected $s_{i\emptyset}$ interview units to j classes randomly but that would be inefficient. The natural way to do the job is to allocate units as they are selected to j classes with probabilities that are proportionate to the sizes of the initial differences between the CPS/SEO and reference cell numbers. If it were to happen that $s_{ij}/r_{ij} > 1.0$, we could first allocate the units in excess cells to deficit cells before computing the j classing probabilities. Actually, we have not found it necessary to do this.

Implementing the procedure, we compute the s_{ijk} and r_{ijk} as initial information insofar as it is possible to do so. For no type of income can we supply a full spread of relevant r_{ijk}. For labor and property incomes we have r_{ij} from the *Statistics of Income* and r_{ik}, where k is age over and under 65. For most sources of grant income we have r_i and, separately, r_k in a few dimensions. Robert Pugh of the Social Security Administration is extending a method originated by Deming[4] for filling interior cells knowing only rim totals. But we take income distributions as they come to us. Nevertheless, the so-far-as-possible r_{ijk} are developed in absolute values and the computer program in reassigning a unit from $s_{i\emptyset}$ to an s_{ijk} cell adds the reassigned unit to the preexisting units and compares the new s_{ijk} cell count to the r_{ijk} count in order to determine whether $r_{ijk} - s_{ijk}$ has gone to zero or not. If it has, the unit is assigned to the nearest cell with a vacancy. By this means, the stochastic assignment is constrained to a right outcome.

The rectification procedure outlined works well enough, aside from data limitations, for labor and property incomes and for non means tested grant income. But for means tested tax and grant transfers, there is a better way. We compute taxes and grants using filing unit income and other characteristics that define eligibility. Where these computations yield counts of units and transfers which are higher than reference estimates, we reduce the computed results using con-

[3] For a description of TRIM, see McClung, Moeller and Siguel, *Transfer Income Program Evaluation*, Urban Institute Paper 950–3.

[4] W. Edwards Deming, *Statistical Adjustment of Data*, New York: Dover Publications, 1964.

strained random participation probabilities. Filing units reporting, for example, receipt of a grant under some program are given a participation probability of 1.0; others for whom net positive grants are computed are assigned participation probabilities less than 1.0 by whatever is required to bring computed and reference estimates into agreement.

2. Applications

We make adjustments to SEO and CPS records for the following elements of income: (1) wage (including salary), (2) self-employment, non farm, (3) self-employment, farm, (4) rent, (5) interest, (6) dividend, (7) Old Age, Survivors and Disability Insurance and Railroad Retirement, (8) government employee pension, (9) private employee pension, (10) Unemployment Insurance, (11) Workmen's Compensation, (12) Veterans' Compensation, (13) Veterans' Pension, (14) Aid to Families with Dependent Children, (15) Old Age Assistance, (16) Aid to the Permanently and Totally Disabled, (17) Aid to the Blind, (18) General Assistance, (19) realized capital gain, (20) Federal Income Tax, (21) Federal Insurance Contributions Act tax. Not all elements are identified and some not present on file records. For the SEO, we allocate Veterans' Disability Benefits to Compensation and Pensions, Public Assistance to the five components, impute capital gain income and compute Federal Income and Federal Insurance Contribution Act taxes using TRIM procedures which we have developed for doing those things. For the CPS, we also allocate the five types of so-called unearned income to the component sources listed above using a TRIM procedure for that.

Wage income is adjusted using a Case 1 rule; that is, we merely increase reported amounts by the ratio of aggregate reference to aggregate SEO or CPS amounts. The adjustment to nonfarm self-employment income entails a search for additional units. Because our reference data for farm self-employment income are so hopeless, we adjust farm income on the SEO to a USDA control using rates of return. Having neither assets nor gross receipts on the CPS file, the adjustments to CPS records are more imaginative. The problem is translation of tax return income into economic income and we must do this from relationships found in the SEO data. We do not adjust rent income, since agreement between file and reference estimates is reasonably good. SEO interest and divided income we adjust using a Case 1 rule; CPS adjustments require a search. For the SEO, adjustments to OASDI and RRR income as to Veterans' Disability income are made by, first, recoding some government employee pension income as OASDI or Veterans' Disability income and then searching for likely additional recipients. In the CPS file we do not have an excess of units reporting government pensions. For Unemployment and Workmen's Compensation income we use a Case 1 rule.

The private employee pension income adjustment sends us looking for additional recipients and we present this application as an example of the general procedure. For the 1967 SEO we require 754,000 interview units from a population of units with male heads age ≥ 55 not reporting government employee pensions. Given this obviously approximate specification, we construct a vector T_i of weighted SEO interview units in FMI class i and PPEN class $j = 0$, that is the numbers of units reporting a zero amount of private employee pension income. We then

construct an $n \times 2$ matrix P_{ik}, where P_{ik} is the weighted number of SEO interview units in FMI class i reporting PPEN and $k = 1$ if age of head is under 65 and $k = 2$ if age of head is greater than or equal to 65. We next compute ratios $P_{ik}/\sum_i \sum_j P_{ij}$ and use these ratios to prorate the 754,000 discrepants to cells in an $n \times 2D_{ik}$ matrix. Ratios $R_{ik} = D_{ik}/T_{ik}$ are the probabilities that units in a T_{ik} cell have of being selected for membership in the P_i rows. Once selected, a unit is assigned to a j cell with a probability $P_{ij}/\sum_j P_{ij}$ and, assigned, is given the mean PPEN for that cell.

The appendix table shows for each source of income, the initial reference counts, the reconciliation adjustment, the adjusted reference counts, the 1967 SEO counts and the evident discrepancy. In notes we indicate the record adjustment rules. A more detailed description of the SEO and CPS income reconciliation and record adjustments is available in an Urban Institute Working Paper (WP 505–3). In that paper we also explain our adjustments to SEO assets and liabilities and the procedures for imputing assets and liabilities to CPS records. The appendix table does not show reference data for computed elements of income or imputed realized capital gains. These elements of income either are not on the SEO and CPS records or do not enter into the adjustments to income. Thus, a comparison of file and reference data is not of much interest and even brief explanations of the adjustment processes would add several pages to this paper. Further, the adjustments shown are the crucial ones, for it is upon their correctness that the accuracy of the computed adjustments depend.

The Urban Institute

Appendix

UNITS IN 10^3; AMOUNTS IN 10^6

	Reference	Reconciliation Adjustment	Adjusted Reference	1967 SEO	Evident Discrepancy
Wage					
Units	62,361	−12,928	49,433	49,460	−27
Amounts	381,067	−9,348	371,719	353,854	17,865

Source: Reference is *Statistics of Income, Individual Tax Returns, 1966*, (SOII (1966)), Table 10; reconciliation adjustments are for multiple recipients, mortality, institutionalization, foreign residence, nonfilers and rare rich.

Record adjustment: No change in recipient units; amounts multiplied by 1.020.

Nonfarm self-employment					
Units	8,092	−469	7,623	6,502	1,121
Amounts	38,109	−5,508	32,601	40,382	−7,781

Source: SOII (1966) Tables 15, 18, 7 (Cols. 46–49); reconciliation adjustments are for multiple recipients and rare rich.

Record adjustment: Final 1,121 units reporting real estate assets but not rent and assign them mean profits or losses averaging zero.

Farm self-employment					
Units	3,009	−127	2,882	2,776	106
Amounts	13,263	+36	13,299	7,536	5,763

Source: Recipient units from SOII (1966) Table 17; amount is the USDA estimate of realized net income reported in *Statistical Abstract, 1970*, Table 929, reduced by rental value of farm dwellings and increased by rent paid to farm landlords. Reconciliation adjustments are for multiple recipients and rare rich.

Record adjustment: No change in recipients; SEO farm sales value expanded to the Flow of Funds estimate is multiplied by separate rates of return for primary and non-primary farmers which yield the adjusted USDA aggregate.

	Reference	Reconciliation Adjustment	Adjusted Reference	1967 SEO	Evident Discrepancy
Rent					
Units	6,763	−344	6,406	6,360	46
Amounts	3,320	−426	2,894	5,478	−2,584

Sourse: SOII (1966) Table 7 (Cols. 34–41). Reconciliation adjustments are for multiple recipients and rare rich.

Record adjustment: None.

Interest					
Units	28,316	3,907	24,409	29,475	−5,066
Amounts	13,225	−1,034	12,191	7,433	4,758

Source: SOII (1966) Table 14; reconciliation adjustments are for multiple recipients and rare rich.

Record adjustment: No change in recipients; amounts multiplied by 1.640.

Dividend					
Units	11,632	−1,087	10,545	9,689	856
Amounts	16,057	−5,510	10,547	7,088	3,459

Source: SOII (1966) Table 11 (Cols. 7 and 8) and Table 7 (Cols. 42–45). Reconciliation adjustments are for multiple recipients and rare rich.

Record adjustment: No change in recipients; amounts multiplied by 1.488.

Old Age, Survivors and Disability and Railroad Retirement grants					
Units	23,366	−7,943	15,423	14,094	1,329
Amounts	21,006	−2,435	18,571	17,401	1,170

Source: *Social Security Bulletin, Statistical Supplement, (1966)* Table 87 and Storey, *Public Income Transfer Programs*, Joint Economic Committee, 1972, Table 8; reconciliation adjustments are for multiple recipients, mortality, institutionalization and foreign residence.

Record adjustment: Two-thirds of the excess records reporting Government Pensions are recoded OASDI; mean OASDI benefits by age, marital status and current wage income are imputed to enough nonreporters to bring the SEO aggregate into agreement with the adjusted reference; remaining discrepants are considered benefits drawn under multiple account numbers.

Government pension					
Units	2,329	−283	2,046	2,618	−572
Amounts	5,685	−648	5,037	5,252	−215

Source: SS BULL SS (1967) Table 9 adjusted to end of year and for dual civilian and military pensions.

Record adjustment: Excess records are recoded two-thirds OASDI and one-third Veterans' Disability.

Private pension					
Units	3,110	−368	2,742	1,988	754
Amounts	4,190	−445	3,745	2,665	1,080

Source: Kolodrubetz, SS BULL (Apr. 1972)

Record adjustment: Random search for 754 units with male heads age ≥ 55 not reporting government pensions; selected units are assigned mean private pension amounts for age and AGI classes.

Unemployment Compensation					
Units	4,455	−1,325	3,130	2,867	263
Amounts	2,547	—	—	1,146	1,401

Source: *Statistical Abstract (1968)* Table 429 and *Handbook of Labor Statistics (1969)* Table 1; reconciliation adjustments are for multiple recipients and recidivism.

Record adjustment: No change in recipients; amounts multiplied by 2.223.

Workmen's Compensation					
Units	—	—	—	2,028	—
Amounts	1,293	—	—	1,037	256

Source: SS BULL SS (1967) Table 9.

Record adjustment: Amounts multiplied by 1.246.

	Reference	Reconciliation Adjustment	Adjusted Reference	1967 SEO	Evident Discrepancy
Veterans' cash grants					
Units	5,193	−528	4,023	3,360	663
Amounts	4,373	−506	3,857	3,088	769

Source: Veterans' Administration *Annual Report* (1969).

Record adjustment: Discrepants left after government pension recode are sought among SEO units subject to SEO counts of living and decreased units and pension and compensation units agreeing with reference.