Allocated Values in Linked Files

Amy O'Hara[1]

Housing and Household Economic Statistics Division
U.S. Census Bureau
4600 Silver Hill Drive
Washington D.C. 20233
amy.b.ohara@census.gov

Introduction

Administrative records from other agencies are linked to U.S. Census Bureau survey data for statistical research. This paper discusses a link between the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) and Internal Revenue Service Individual Master File (IRS IMF). An evaluation of the pre- and post-tax income distributions in the CPS ASEC is possible when these data sources are linked. Tax information is not collected from survey respondents, but state and federal income tax liabilities are estimated based on income information that is collected. Investigating the quality of the tax estimates is important because they are publicly released and widely used in income distribution analyses. The income data upon which the tax estimates are based can also be evaluated using the IRS administrative records due to the inclusion of specific income types in both data sources. Differences in income concepts and reporting requirements cause the IRS IMF and CPS ASEC distributions to differ. Census Bureau processing also has an important impact on the survey data, especially regarding allocated values. When income data are missing in a survey response, Census Bureau imputes an amount using well established methods. This paper reviews differences in the income data from the two sources, issues with the linked file, and the potential effects imputations have in linked data analyses.

This paper specifically evaluates the use of allocated income in administrative record research. If cases with imputed income values are omitted from analyses, the usefulness of the linked data declines. If cases with imputed values are retained in analyses, incorrect conclusions may be made. Since CPS ASEC income is imputed to fill the distribution at the national level, these imputed values may be good in some cases and not in others. This paper investigates a calibration method to determine which imputed values are "good enough" and what happens to cell counts when these cases are retained.

Data

The CPS ASEC is conducted by the Census Bureau for the Bureau of Labor Statistics to collect detailed employment and income information. Approximately 100,000 households are in the sample, which covers the civilian, non-institutional population of the United States. The IRS IMF contains tax data and related information for individual taxpayers, organized into modules for each tax year. The Census Bureau is allowed access to this information for statistical research purposes only. The data pertain to the various forms and schedules in the 1040 Individual Income Tax set of returns.

The IRS IMF is preferred to published aggregates when evaluating modeled tax information in the CPS ASEC. IRS publishes preliminary tax data each spring on the previous year's tax returns. These aggregates are inferior to the administrative records because they reflect the population of tax filers rather than the civilian non-institutionalized population present in the CPS ASEC. Another source of tax information is the public use microdata file from the IRS Statistics of Income (SOI) Division, which is released annually with a three-year lag. The SOI public use file oversamples high income returns, establishing a basic inconsistency for comparisons with CPS ASEC. Using administrative records allows comparisons of tax fields based on responses from the same individuals. For instance, IRS data on filing status and the number of exemptions claimed can be compared to modeled estimates. These key tax variables impact income cutoffs for marginal tax rates and credit eligibility. The variables received from the IRS per U.S. Code regulations are listed in Table 1. Most of the income categories listed in Table 1 are collected in CPS ASEC. Inconsistencies arise when concepts vary between the data collections. Differences in definitions for marital status, children and income sources cause the greatest discrepancies when comparing CPS ASEC to tax return data. Appendix 1 specifies the concepts collected and highlights how the variables differ between the administrative records and survey data.

Table 1. Individual income tax return variables received from the IRS for statistical research[2]
Marital status
Number and classification of exemptions
Wage and salary income (continuous)
Dividend income (continuous)
Interest income (continuous)
Gross rent and royalty income (continuous)
Total money income (continuous)
Adjusted gross income (continuous)
Indicators for presence of Schedule A, C, D, E, F, and SE
Social security benefits (continuous)
Earned income used to compute Earned Income Tax Credit (EITC) (continuous)
Number of EITC eligible qualifying children

Using linked data

Linked file analyses are useful at the aggregate level as well as the microdata level. Comparing CPS ASEC estimated adjusted gross income (AGI) to the IRS AGI value for the records in the exactly matched data set is an example of an aggregate analysis. In the 2004 CPS ASEC - 2003 IRS IMF linked data (referring to income year 2003), CPS modeled AGI was 0.7% larger than IRS reported AGI. Weighted CPS ASEC modeled AGI falls far short of the aggregate amount published for all tax filers because the survey does not include the persons with extremely high income values present in the administrative data. Such benchmarking analyses are useful for evaluating the validity of the survey content and administrative record quality. Studies at the microdata level may look at the characteristics of filers who itemized, may investigate patterns of exemptions claimed, or may evaluate CPS ASEC based eligibility measures for different tax credits. The usefulness of aggregate and individual comparisons depends on a number of factors. This paper considers two factors: the strength of the probabilistic match, and item and unit non-response.

Data from the 2004 and 2006 CPS ASEC linked to the 2003 and 2005 IRS IMF are used. CPS ASEC respondents have positive, negative, or zero income. IRS IMF filers have positive or negative income. Both data sources allow earned income to be differentiated from unearned income. This analysis uses cases with income present in both CPS ASEC and IRS IMF. More of the analysis in this paper uses the 2004 CPS ASEC-2003 IRS IMF matched data because the later year's data is a relatively recent acquisition. Some preliminary tabulations have been prepared using this file and more in depth analysis will occur as time allows.

The strength of the probabilistic match

The Data Integration Division at Census Bureau matches person records from the survey to tax returns using Social Security Numbers (SSN) in a secure environment. Prior to 2006, the Census Bureau asked respondents for their SSN during the survey administration. Any person refusing to provide their SSN was omitted from the pool of potential matches. Identities for persons who provided incorrect or incomplete SSNs were searched based on name, address and date of birth. After this verification process was completed, SSNs were dropped and Personal Identification Keys (PIK) were added. The cases with verified identities were then sorted and merged with the administrative record source. The resulting file of matched cases contained permitted IRS IMF variables and CPS ASEC identifiers. Neither SSN nor PIK were on the final linked file.

Nearly 50,000 adults refused to provide their SSN in the 2004 CPS ASEC. This greatly limited the potential pool for linking to IRS records and compromised the usefulness of the CPS ASEC weight used to make population inferences during analysis. An additional 5 percent of the adults were not matched due to missing name data on the survey, multiple matches found during the link, or lack of data in the administrative record sources used to verify identities. In total, 65 percent of adults were verified in the 2004 CPS ASEC.

The 2006 CPS ASEC stopped asking survey respondents for SSNs. Households selected for the survey were sent letters informing them that their data may be combined with other data sources if they did not object. Persons who contacted the

Census Bureau requesting their data not be linked were omitted from the pool of potential matches. The number of refusals plummeted compared to prior years. Relying solely on the name, address and date of birth to verify records, approximately 89.2 percent of all adults were verified in the 2006 CPS ASEC.

Using the linked cases at the microdata level involves two critical assumptions regarding verification: First, the identification process is assumed to be accurate, resulting in the same survey person as the tax filer. Second, the unverified cases –both the refusals and non-matches- are assumed to be missing completely at random. With these assumptions, the linked data set is a useful evaluative tool for assessing survey responses and modeled estimates.

Item and unit non-response

Using data collected in the CPS ASEC involves consideration of two types of non-response. Item non-response occurs when the participant completes the survey but refuses to answer certain questions. In the case of earnings, they may disclose their labor force attachment, industry and occupation, and other demographic characteristics, but refuse to say how much they earned. The Census Bureau deals with item non-response by allocating values from similar records that did provide complete information. This information is allocated using a hot deck imputation approach that is designed to replicate the earnings distribution at the national level but is not designed to simulate person-specific earnings.

A second form of non-response in the CPS ASEC is unit non-response. This occurs when a respondent participates in the CPS Basic survey but refuses to provide the income and labor force detail requested in the CPS ASEC. When a respondent fails to report enough information to qualify as a CPS ASEC interview, all supplement responses are allocated from a similar CPS ASEC interview household.

This paper focuses on income allocations, specifically earnings allocations. In CPS ASEC, total person level income is divided into earned income and other income. Earned income may come from wage and salary employment, self-employment or farm self-employment at the primary or other jobs. The main quantity of earned income is captured in ERN_VAL, which contains earnings from the longest job held in the previous year. An allocation flag (I_ERNVAL) exists to inform the data user that this earnings item has been imputed. Earnings from other jobs are separated based on type of employment. Wage, self or farm employment earnings from other jobs are presented WS_VAL, SE_VAL, and FRSE_VAL respectively. Each of these other job earnings variables has a corresponding allocation flag. Table 2 displays some descriptive statistics on earnings allocations for all adults in the 2004 CPS ASEC file. Income allocations are often discussed in terms of the number of dollars imputed. Table 2 indicates that 24.1 percent of adults with earnings had some imputed earnings in the 2004 CPS ASEC.

Table 2. Percentage of CPS ASEC Adults with Allocated Earnings

	2004 CPS ASEC
Not self employment earnings imputed	12.9
Self employed earnings imputed	0.9
Unit imputed (includes earnings)	10.8
Any imputed earnings	24.1

This paper is concerned with the count of records with allocated income, not improving on the dollar amounts imputed. Comparing survey data to administrative records at the person level should not prove useful if the survey data are imputed to the national distribution rather than microdata detail. However, omitting one-quarter of adults from linked file analyses because they have some or all of their income data imputed greatly limits the usefulness of administrative record research and evaluation.

Figure 1. Earnings Distributions by Imputation Status, 2004 CPS ASEC

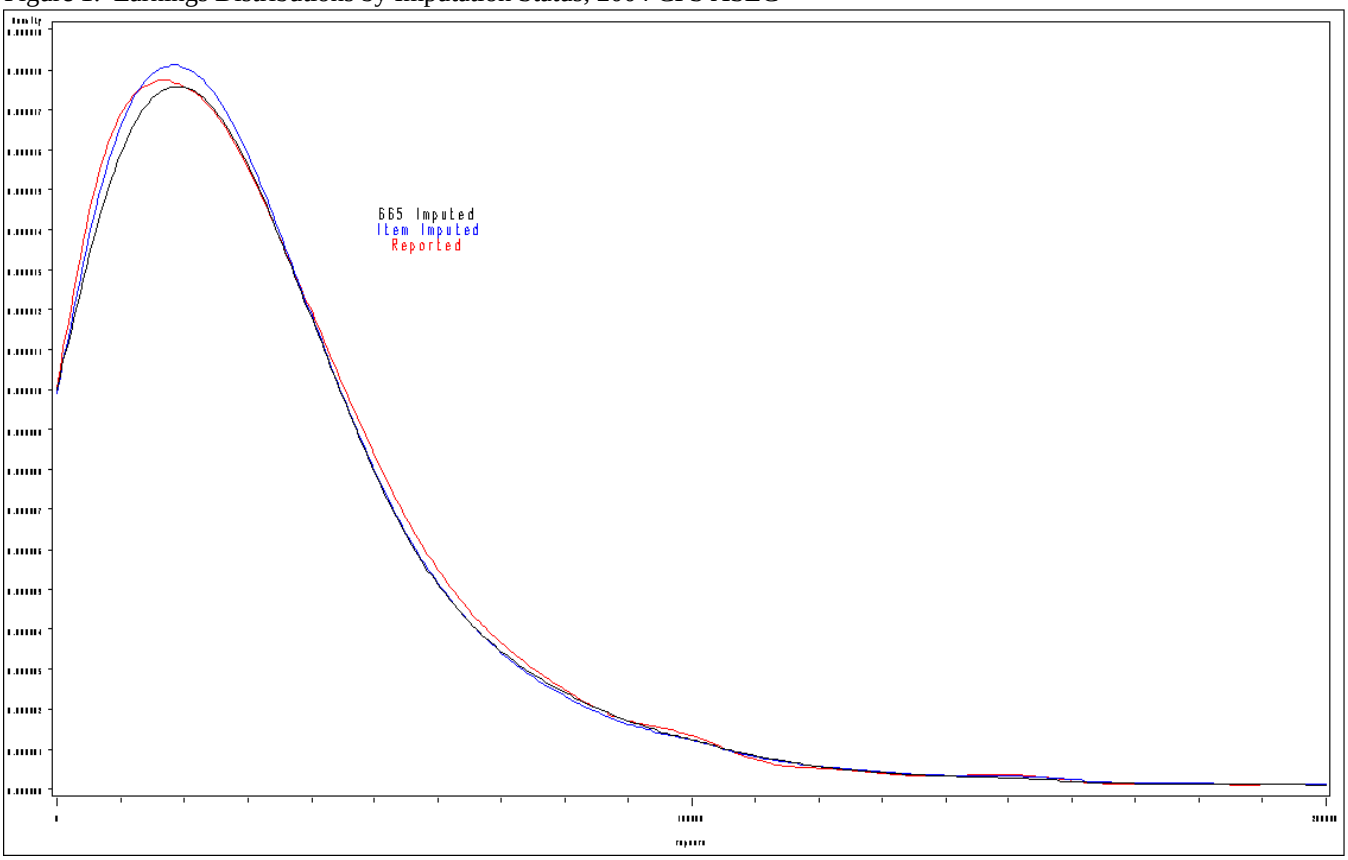


Figure 1 illustrates the earnings distributions for adults in the 2004 CPS ASEC by allocation status. Total person earnings are on the horizontal axis; the sample is restricted to positive values up to $200,000. The vertical axis indicates the frequency at which the value occurs in the data. The curves are very close, indicating that the national earnings allocations compare reasonably well at the person level. The allocated earnings cases look similar to the reported cases, as would be expected with a hot deck imputation process. At lower earnings values, up to approximately $30,000, more people have item imputed earnings than reported or fully imputed earnings (there is more area beneath the blue curve than the red or black curves). This is not a conclusive result but points to the need for more research on the impact of allocations at the low end of the income distribution.

The density functions in Figure 1 represent all 2004 CPS ASEC adults with earnings. Not all of these adults are in the universe of cases to link to the IRS IMF data. From the earlier discussion of the probabilistic match, we know that 65.0 percent of adults were verified. Not all adults who were verified filed a tax return. From Table 2 we know that 24.1 percent of all adults in the 2004 CPS ASEC had some imputed earnings. From Table 2 we know that 24.1 percent of all adults in the 2004 CPS ASEC had some imputed earnings. What is the interaction of imputations and verification? Table 3 displays the cross frequency for imputed and verified cases in the 2004 CPS ASEC that defines the linkable universe.

Table 3. Verified Cases by Earnings Imputation Status for 2004 CPS ASEC

2004 CPS ASEC	Not verified		Verified		Total	
	Count	% of row	Count	% of row	Count	% of row
Earnings not imputed	35,982	29.7	85,192	70.3	121,174	100.0
Imputed earnings	19,859	51.7	18,585	48.3	38,444	100.0
Total	55,841	35.0	103,777	65.0	159,618	100.0

None of the 55,841 adults who were not verified are in the universe to match to an IRS record. They are, however, in Figure 1, which was not conditioned on the ability to link to the IRS file. Earnings were imputed for 17.9 percent (18,585/103,777) of the verified cases. Omitting all of these observations from linked file analyses would restrict the sample to 85,192 (103,777-18,585) which is just over half of the full adult count in the 2004 CPS ASEC. Recall that all 159,618 would be eligible for linked file analysis if there were no refusals to combine data with other agencies and all persons were verified in our probabilistic match.

Procedural changes when administering the 2006 CPS ASEC altered the proportion of cases that were verified. As mentioned earlier, 89.2 percent of all adults were verified. Table 4 displays the cross frequency on verification status and earnings imputations for the 2006 CPS ASEC. The increase of 36,385 adults in the verified column of Table 4 compared to the verified column in Table 3 should result in a larger number of records in the 2006 CPS ASEC – 2005 IRS IMF linked file. Due to time constraints, this has not been investigated. The proportion of cases with imputed earnings remains stable between the 2004 and 2006 CPS ASEC files, at 24.1 percent (38,444/159,618) and 21.6 percent (33,954/157,114) respectively. Between Table 3 and 4, the row for imputed earnings changes substantially. In Table 3, the count of cases with imputed earnings (38,444) was almost evenly split between verified and not verified. In Table 4, more than 80 percent of the imputed cases were verified. The impact of this shift cannot be described without further analysis, but the magnitude of the shift warrants mention. The count of verified cases where earnings were not imputed increased from 85,192 in the 2004 CPS ASEC to 112,636 in the 2006 CPS ASEC. Regardless of the treatment of cases with imputed earnings, this increase should generate more matches to the IRS IMF file. If a method of including some of the imputed cases can be determined, the usefulness of the linked file would increase.

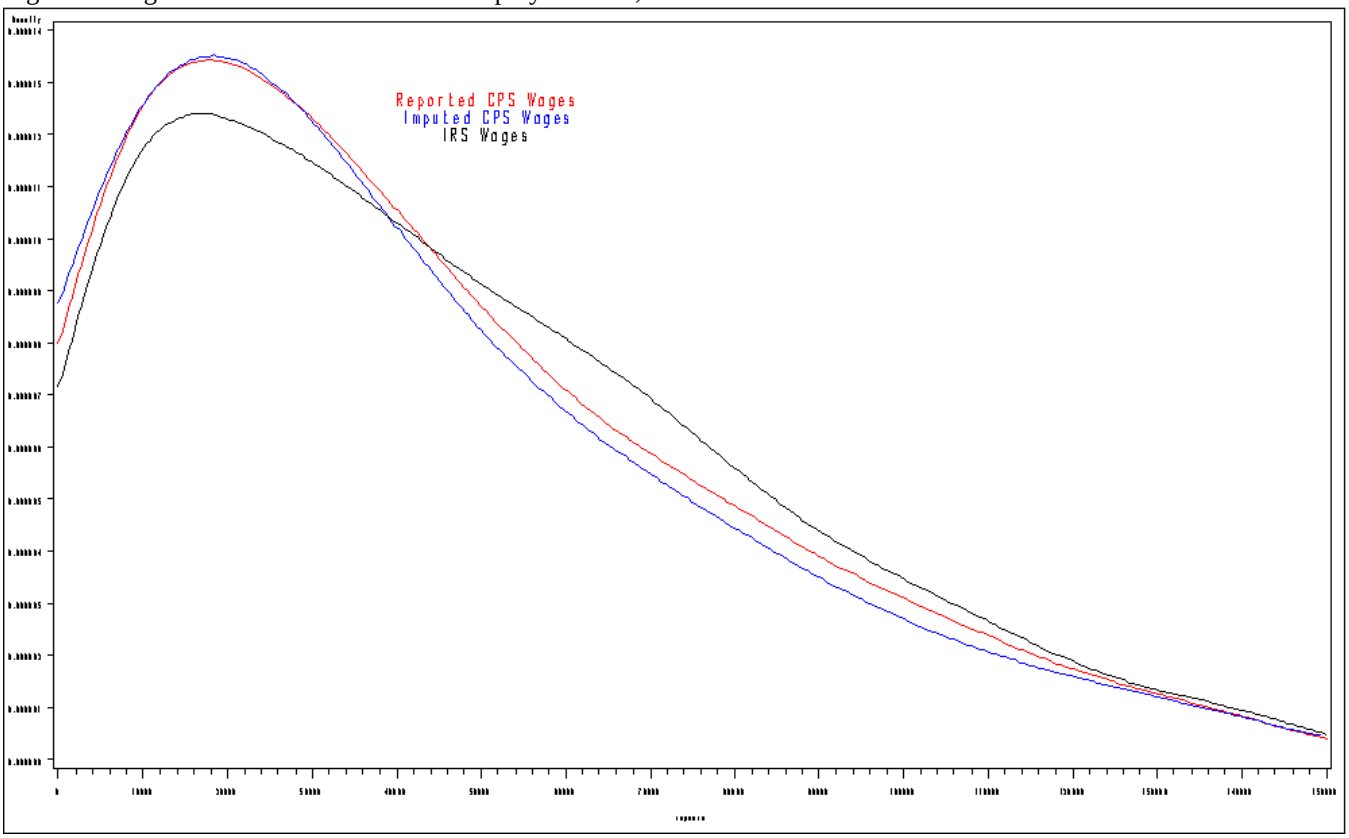
Table 4. Verified Cases by Earnings Imputation Status for 2006 CPS ASEC

2006 CPS ASEC	Not verified		Verified		Total	
	Count	% of row	Count	% of row	Count	% of
Earnings not imputed	10,524	8.5	112,636	91.5	123,160	100.0
Imputed earnings	6,428	18.9	27,526	81.1	33,954	100.0
Total	16,952	10.8	140,162	89.2	157,114	100.0

The 18,585 verified adults with imputed earnings in the 2004 CPS ASEC weight to 25.0 million filing units using the Census Bureau tax model. In the 2006 CPS ASEC, the 27,562 verified adults with imputed earnings weight to 42.4 million filing units. There were 130.6 million individual income tax returns files in 2003.[3] If every verified 2004 CPS ASEC adult were linked to an IRS IMF record, income comparisons would be suspect for the 19.1 percent with partial or total imputed earnings. Not every CPS ASEC verified case matches to an IRS IMF record. Of the 103,777 verified 2004 CPS ASEC persons, 58,225 matched to a 2003 IRS IMF tax return. Twenty-one percent (12,158) of these linked cases have imputed earnings.

How do earnings in the CPS ASEC linked cases – both reported and allocated – compare to the amounts reported to the IRS? Figure 2 displays the density functions for the cases in the linked file. The three curves show reported CPS ASEC wages, imputed CPS ASEC wages (now grouping cases with item or fully imputed earnings together), and IRS IMF wages for persons who were not self-employed[4] in their main job and whose total income was greater than zero. Figure 2 is based on a truncated distribution, earnings below zero and above $150,000 are omitted from the density function estimation, and only cases where the modeled filing status aligned with the actual filing status are included. Below $40,000, there is more area under both curves based on CPS ASEC data compared to the IRS data, indicating that more people have incomes in that range in the survey than the administrative records. In the $50,000 to $120,000 range, more cases are represented in the IRS data. Separate analysis shows that imputed earnings exceed IRS IMF earnings past $150,000. The curves of earnings distributions exhibit interactions absent in tabular data comparisons and point to areas for continued analysis.

Figure 2. Wage Distributions for Not Self-Employed Filers, 2004 CPS ASEC- 2003IRS IMF Linked File



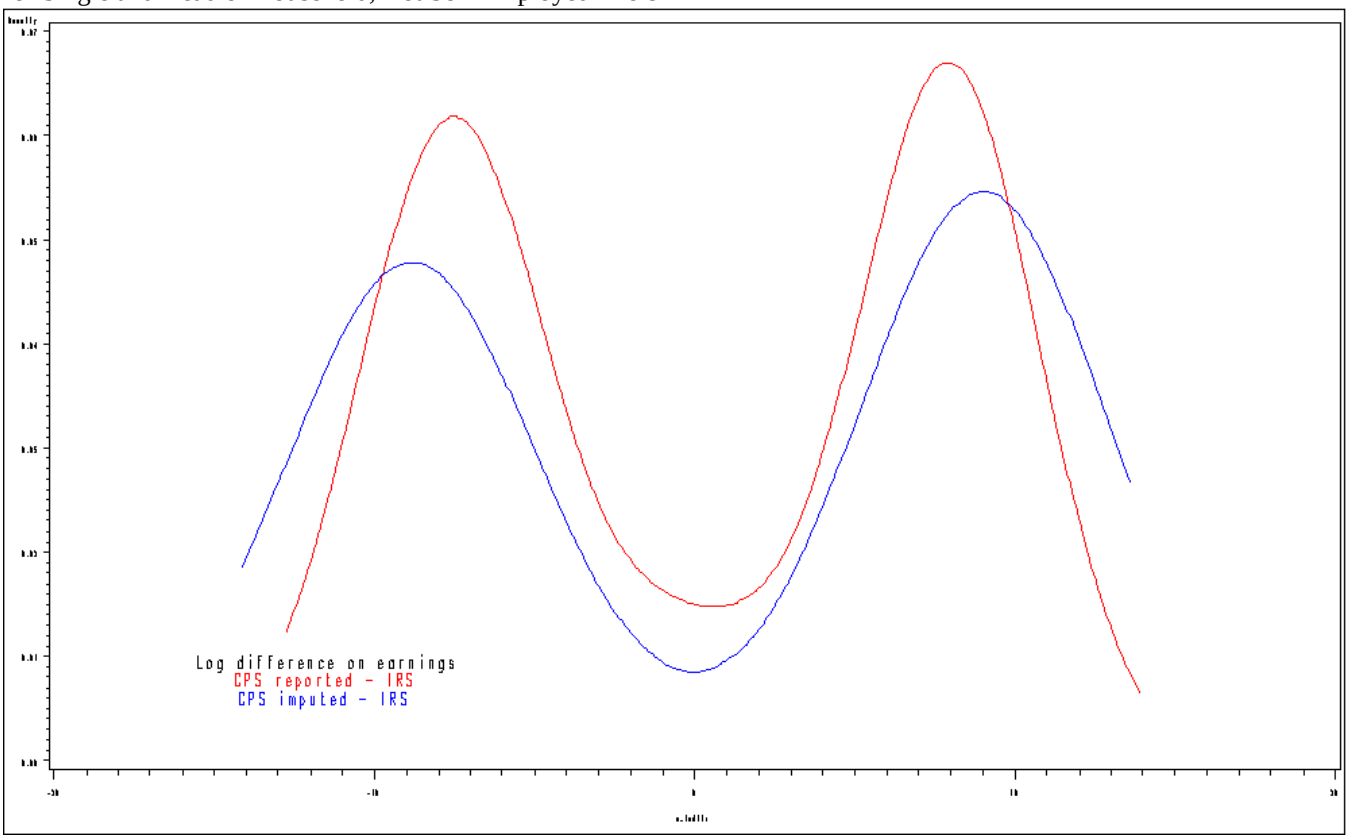
Good enough

Twenty percent of the cases used in the density functions would be omitted if all CPS ASEC cases with imputed earnings were dropped. But Figure 2 shows that imputed earnings resemble reported earnings. Are any of the imputed cases "good enough" to include? To develop a tolerance for keeping observations with imputed earnings in a linked analysis, a sample of records with comparable income concepts was created. Single and head of household tax returns in the 2004 CPS ASEC – 2003 IRS IMF file were retained. No married filing jointly returns were permitted, whether reported to the IRS or modeled by the Census Bureau. This limited the chance of missing an earner or erroneously including an earner. Observations with self-employed earnings were also omitted to prevent the inclusion of erroneously classified income in the comparison. Income concept differences between the data sources are not addressed in this exercise. As noted in Appendix 1, the IRS collects taxable earnings while the CPS ASEC collects gross earnings. Therefore, payroll taxes, income tax withholding, and pretax deductions for insurance premiums or thrift savings plans would reduce IRS earnings.[5] This sample of 24,815 linked observations was used to attempt to identify a tolerance by which imputed wages may be included to preserve sample. Of the 24,815 single and head of household, not self employed persons, 5,806 or 23.4 percent had some or all earnings imputed. Earnings were reported to the CPS ASEC in the remaining 19,009 cases.

The natural log of the difference between the CPS ASEC reported wage and the IRS IMF reported wage was calculated for the sample. The log was computed on the absolute value of the difference in the wage amounts, and the resulting log scale was signed according to whether the CPS ASEC amount exceeded the IRS IMF amount (positive) or fell below the IRS IMF amount (negative). The signed log values for the reported CPS ASEC earnings ranged from –12.8 to 13.9, displaying symmetry around zero. In less than 5 percent of the cases, the reported wages matched exactly; the $0 difference was replaced with a $1 difference to keep these observations in the analysis with a log value of zero. The distribution of the log differences was also constructed for the imputed cases, where less than one percent of the imputed values exactly matched the IRS IMF value.

The distributions of these log differences are shown in Figure 3. This strange looking graph results from the decision to show the sign of the log difference to indicate where the CPS ASEC amounts were larger than the IRS IMF amounts (to the right side of zero). The higher, narrowed peaks in the positive domains (where CPS ASEC earnings exceeded IRS IMF earnings) were expected due to the gross versus taxable difference in data collected.

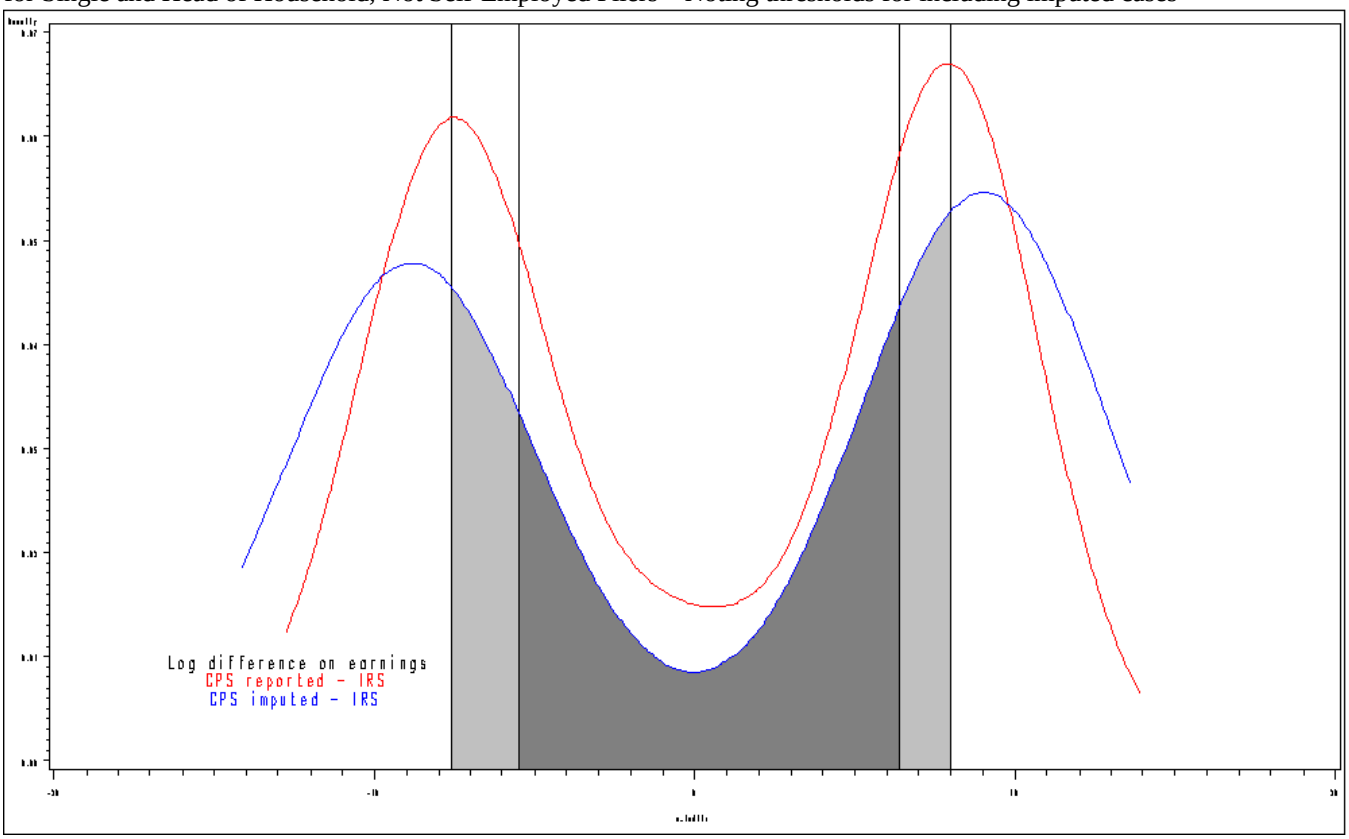
Figure 3. Distributions of Log Differences between CPS ASEC Wages and IRS IMF Wages in Linked Tax Year 2003 File for Single and Head of Household, Not Self-Employed Filers



The area under the trough, at and near a log difference of zero, represents the cases with the least difference in earnings between the data sources. The points on the log distribution that define the center quintile of the reported income distribution (red curve) are negative 5.5 and positive 6.4. These refer to income differentials where IRS exceeds CPS by $245 (exp 5.5) and where CPS exceeds IRS by $602 (exp 6.4), and all values in between.

Figure 4 displays the boundary of this center quintile and shades the area under the imputed curve in dark gray. When cases in this range are retained, 267 imputed cases would be included in a linked file analysis, being “good enough” based on this one metric. These 267 cases reflect 338,850 persons when weighted, and make up 4.6 percent of the imputed cases (5,806) in this exercise. Increasing the range to include cases to the center of the medians for both peaks (-7.6 and 8.0 respectively) nets 1,221 cases. This broader range is also displayed in Figure 4 adding light gray shading to the dark gray center to represent the larger group of cases. Using the median thresholds increases the income differentials (IRS exceeds CPS by $1,998 at exp 7.6 and CPS exceeds IRS by $2,981 at exp 8.0). The 1,221 cases comprise 21.0 percent of the 5,806 imputed cases and weight to 1,629,000 persons.

Figure 4. Distributions of Log Differences between CPS ASEC Wages and IRS IMF Wages in Linked Tax Year 2003 File for Single and Head of Household, Not Self-Employed Filers – Noting thresholds for including imputed cases



Conclusion and Extensions

The inverse log of the median amounts shown in Figure 4 range from approximately $2,000 to $3,000. These amounts may or may not be acceptable, depending on the location along the earnings distribution. Further research will investigate methods to scale the wages, so that a $2,000 reporting difference is viewed differently for a person with $10,000 versus a person with $200,000. Due to the nature of tax data, the same difference may impact single, married joint, or head of household filers differently. Quantile regression may be useful to disentangle some causes and effects for smaller log differences and allow inferences on the use of imputed wages on survey data alone.

Repeating (and improving) this exercise using the 2006 CPS ASEC – 2005 IRS IMF should prove interesting, given the large increase in verified cases. This match will also reflect tax modeling improvements made by Census Bureau. The distributions of the earnings differentials between the data sources should differ given the improvements in filing status assignment and increased number of cases.

Including cases with imputed income in the linked data set will be useful in tax model evaluation studies. Cases with reasonable income comparability should allow a better understanding of when people choose to itemize their deductions or participate in credits. An increased number of cases will also be useful when seeking patterns in capital gain or loss claimants. Once a method of retaining imputed income cases is refined, the impact on these tax issues will be determined.

References

Roemer, Marc I. 2000. "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990-1996. http://www.census.gov/hhes/www/income/assess1.pdf.

Roemer, Marc I. 2002. "Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation," <http://www.census.gov/hhes/www/income/asa2002.pdf>.

Webster Jr., Bruce H. 2007. "Evaluation of Median Income and Earnings Estimates: A Comparison of the American Community Survey and the Current Population Survey." Paper presented at the American Statistical Association meeting.

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[2] These items of return information are provided by the IRS to the U.S. Census Bureau under the authority of 26 U.S.C. Section 6103(j)(1).

[3] Individual Income Tax Returns, Preliminary Data, 2003. <http://www.irs.gov/pub/irs-soi/05inplim.pdf>.

[4] Self employed persons are omitted from this analysis. More research is needed on how self employed persons classify their income. In the CPS ASEC, it is collected as earnings but it may be reported to the IRS as salary or business income, as net or gross.

[5] Roemer (2002) evaluated CPS ASEC income against Detailed Earnings Records which allowed a closer alignment of earning concepts.

Appendix 1. Comparison of Concepts Collected in CPS ASEC and IRS IMF

Concept	CPS ASEC	IRS IMF
Marital Status	1 Married (civilian spouse present) 2 Married (Armed Forces spouse present) 3 Married (spouse absent) 4 Widowed 5 Divorced 6 Separated 7 Never married	1 Single 2 Married filing jointly 3 Married filing separately 4 Head of household 5 Qualifying widow/er

Child	A child is a person under age 15; their relationship to the reference person is collected as follows in variable A_EXPRRP: 5 Natural/adopted child 7 Grandchild 9 Brother/sister 10 Other relative 11 Foster child 12 Nonrelative (with own relatives in household) 13 Partner/roommate 14 Nonrelative (with no relatives in household)	A "qualifying child" must meet the following criteria: 1 Relationship: the taxpayer's child or stepchild by blood or adoption, foster child, sibling or stepsibling, or a descendant of one of these. 2 Residence: lives with the taxpayer for more than half the tax year. 3 Age: under age 19 at the end of the tax year, or under the age of 24 if a full-time student for at least five months of the year, or be permanently and totally disabled at any time during the year. 4 Support: did not provide more than one-half of his/her own support for the year. The number of child exemptions and number of children claimed as qualifying for the Earned Income Tax Credit (EITC) are transmitted.
Wage and salary income	Gross earnings (before payroll taxes, withholding, pretax deductions for savings, health insurance, etc.)	Taxable earnings are collected.
Self employment income	Net earnings from business or farm after expenses	Gross income and expenses separately, net amount from Schedule C or F; filing requirements are based on gross income. A flag is transmitted for the presence of Schedule SE; no continuous self employment income variable is transmitted.
Interest income	Total interest income received	Taxable interest received
Dividend income	Total dividend income received	Dividends are specified as ordinary, qualified, and capital gain distributions.
Retirement income	Up to two values are collected, designated by the following sources: 1 Company/union pension 2 Federal government retirement 3 US military retirement 4 State/local government retirement 5 US Railroad Retirement 6 Annuities 7 IRA, KEOGH, or 401K 8 Other sources or don't know	Taxable distributions from 1 IRAs 2 Pensions and annuities

Rental income	Net rental income	Gross and net rental income collected. A continuous gross value for combined rental income and royalties is transmitted.