Evaluation of the 2005 Changes in the AHS Income Questionnaire

U.S. Department of Housing and Urban Development
Office of Policy Development and Research

Evaluation of the 2005 Changes in the AHS Income Questionnaire.

Scott Susin, HHES

September 18, 2007

Background

After redesigning the American Housing Survey (AHS) in 1997, the Census Bureau and the Department of Housing and Urban Development compared the income data collected in that survey with those found in the Current Population Survey (CPS).[1] That study found that the AHS reported fewer households with non-wage income than the CPS and that AHS respondents tended to report self-employment income as wages. In addition, AHS data users requested that disability-related income sources be reported separately from other sources, to make it easier to count the number of households with disabled persons.

The 2005 American Housing Survey addressed these findings and requests by adopting a series of income questions similar to the questions used in the American Community Survey (ACS).[2] Prior to 2005, respondents were asked the wages and salaries of each person in the family, and all other sources of income were collected as a single amount for the family as a whole. The 2005 questions collected an amount for each person in the family from nine different types of income (such as wages and salaries or social security). In addition, for each non-relative (a person not related to the householder), the 2005 questionnaire was changed to ask about wages, self-employment, and other income separately. Prior to 2005, non-relatives were asked only to report their total income. (See Table 1 for a comparison of the 2003 and 2005 AHS income questions.) The AHS chose to adopt the ACS questions because they had already been tested and to facilitate comparisons between the two surveys.

2003 AHS Compared to 2005 AHS

The approach for obtaining household and family income prior to the 2005 AHS resulted in reported income that was generally lower than in other surveys. In turn, Census Bureau surveys usually measure less aggregate income than is reported in the National Income and Product Accounts (NIPA).[3] The new income questions were intended to more accurately collect income and were expected to increase the amount of income reported. Table 2 compares quantiles (percentiles) of household income in the 2005 AHS to the 2003 AHS.

---

[1] See Susin, Scott (2003) "Discrepancies Between Measured Income in the American Housing Survey (AHS) and the Current Population Survey (CPS): Final Report", March 27. http://www.census.gov/hhes/income/hudmemo8a.pdf

[2] The AHS adopted the income categories used in the 2003 ACS with one change. Disability-related payments were accorded a separate item in the AHS, while the ACS included these types of payments with retirement and survivor pensions.

[3] For a comparison of CPS and NIPA income, see Ruser, John, Adrienne Pilot, and Charles Nelson, (2004) "Alternative Measures of Household Income: BEA Personal Income, CPS Money Income, and Beyond," paper presented to the Federal Economic Statistics Advisory Committee (FESAC), December 14. http:www.bls.gov/bls/fesacp1061104.pdf

Household income rose the most in the middle of the distribution while rising less, or even falling, in the tails of the distribution. Median household income rose by 9 percent from 2003 to 2005 in nominal terms (not adjusted for inflation). The 9 percent increase was 3 percentage points faster than inflation, suggesting that reported income increased in the middle of the distribution.

The 75th, 90th, and 95th percentiles rose at a slower rate than the median. The 75th percentile of household income increased by 4 percent, while the 90th percentile remained about the same (falling by 1 percent).[4] The 95th percentile fell considerably and was 25 percent lower in 2005 than it had been in 2003. This probably indicates that fewer unrealistically large amounts are being reported due to the "range checks" implemented for the first time in 2005 that required respondents to confirm large income amounts.[5] As we will see in Table 4, these upper (90th and 95th) percentiles had been higher than the ACS in 2003, but the two data sets became much closer after the AHS questionnaire revisions.

Table 2 also shows that the 10th percentile and 25th percentiles of household income rose at a slower rate than the 9 percent increase in the median or even the 6 percent inflation rate. Household income at both percentiles remained about the same. It is not surprising that the AHS finds the lower percentiles of household income rising more slowly than inflation because the same trend is present in the ACS, as shown in table 4 below. The AHS findings are also consistent with the increase in poverty rates shown in Table 5.

The 5th percentile of household income fell by 12 percent from 2003 to 2005, a drop too large to be consistent with the trends seen in other surveys. Hence, the amount of income reported at the very low end seems to have fallen, contrary to the increase expected due to the new questionnaire.

Family non-wage income rose sharply in all quantiles. In 2003 and earlier, non-wage income showed substantial underreporting compared to other surveys while reported wage and salary income was much closer to other surveys. Hence the increase in non-wage income was consistent with the improvements expected to be produced by the new questionnaire.

It is worth keeping in mind that much of the increase in non-wage income was probably due to a shift in the way that respondents reported self-employment income. A 2003 study (cited in footnote 1) comparing AHS and CPS income data found strong indications that substantial amounts of self-employment income were erroneously reported as wage and salary income. The new 2005 AHS questionnaire seems to have shifted some self-employment income to the correct line.

---

[4] The 1 percent drop for the 90th percentile was not statistically significant.

[5] For example, those who reported earning more than $100,000 were asked to confirm the amount: "I have recorded that [Name] received $[Amount Reported] in wages, salary, tips, bonuses, or commissions DURING THE PAST 12 MONTHS. Is this correct?"

Table 3 compares the rates at which different types of income were reported in the 2003 and 2005 AHS surveys. Between these years, 3.7 percentage points fewer families reported earnings, which was offset by an increase of 3.6 percentage points reporting self-employment income.[6] As noted above, this probably indicates that households began reporting self-employment income in the correct category. The question referred to business income in 2003 but self-employment income in 2005, and it seems likely that many self-employed persons do not regard themselves as owning a business.

Table 3 also shows a large drop in reporting of interest income, dividend, and rental income, from 31.8 percent to 16.2 percent, probably due to a change in the questionnaire. In 2003, three separate questions were asked about these types of income, while in 2005, the questions were combined, and other types of income were added to the list as well, with the single question asking about interest, dividends, rental, estates, trusts, and royalties. Respondents might have incorrectly "keyed in" on the last few items in the list, and reported "no," perhaps meaning "no royalties or trusts," even when they had interest or dividend income.

In 2005, Social Security and pension income were split into two questions. From 2003 to 2005, the receipt rate for the combined income type rose from 26.9 to 28.9. Reporting of "other income" fell from 2003 to 2005, by 3.6 percentage points, possibly for similar reasons as for interest income, that is, because questions were combined. The list of "other" income types lengthened in 2005, to include alimony, which had previously been a separate question.

AHS Compared to Other Data Sets

Table 4 compares quantiles of household income in the AHS and the ACS.[7] With the exception of the 5th percentile, the AHS percentiles were closer to the ACS in 2005 than they had been in 2003, or remained about the same.

The median and higher quantiles were similar in both the ACS and the AHS in 2005. In 2003, the AHS/ACS ratio was 0.93, indicating that the AHS median had been 7 percent below the ACS figure. In 2005 the AHS/ACS ratio had risen to 0.97, indicating that AHS median was only 3 percent below the ACS, suggesting that more income was reported in the middle of the AHS distribution.

The AHS 75th percentile was almost the same as the corresponding ACS figure in both 2003 and 2005 (AHS/ACS ratios of 1.00 and 0.99, respectively). The AHS 90th and 95th percentiles had been higher than the ACS in 2003. After the 2005 questionnaire revisions, both these percentiles fell relative to the ACS in 2005, becoming quite similar to the other survey. The fall in the 95th percentile was particularly dramatic. In 2003,

[6] In this report, "earnings" refers to wage and salary income only and does not include self-employment income.
[7] Both surveys ask about income in the prior 12 months, but the ACS operates year-round while most AHS interviewing occurs during a 3-month period. Table 4 presents ACS income reported during the months corresponding to the height of the AHS interviewing period (Jun – August, 2003 and May – July, 2005).

the AHS 95th percentile was 38 percent above the ACS, while in 2005 it was within 3 percent. As noted above, this suggests that the "range checks" have eliminated erroneous high values.

Household income at the 10th and 25th percentiles was smaller in the AHS than in the ACS. The ratios (88 and 92 percent in 2005) were about the same in both years. Hence the change in questionnaire made little difference in this range of incomes.
The 5th percentile fell by 10 percentage points, from a 0.72 ratio in 2003 to a 0.62 ratio in 2005. As noted above, this suggests an increase in unreported income (unreported to the AHS interviewer). This could indicate a modest increase in unreported income by extremely low income households. However, it might also indicate a large increase in unreported income by a few higher income households. For example, some households that subsist mainly on interest and dividend income may fail to report these sources of income, and end up with reported incomes far below their true incomes.

Table 5 compares poverty rates in the AHS, ACS, and CPS. The AHS and ACS data report poverty during the 12 months prior to Summer 2003 and 2005 (see footnote 4). The CPS data are for calendar years 2002 and 2004.

Household-level poverty in the AHS increased by 0.7 percentage points from 2003 to 2005. Family-level poverty increased by 0.4 percentage points.[8] In both the CPS and ACS, family-level poverty rates increased by 0.6 percent.

Since the AHS poverty rate went up by about the same amount as the rate in the other two surveys, the AHS seems to be capturing about the same amount of income from poor households as in the past.

Note that there is no contradiction between the the possibility that the AHS is capturing about the same amount of income from poor households (as suggested by Table 5) and the 5th percentile figure falling relative to the ACS (as shown in Table 4). The poverty rate includes a much larger fraction of households (14 percent) than does the 5th percentile. The poverty rate findings are quite similar to the results for the 10th percentile. For low-income households, the revised income questionnaire has not had much effect on the amount of income reported.

New Income Imputation Scheme

The hot deck allocation (or imputation) method is widely used at the U.S. Census Bureau and other statistical agencies. In this method, the analyst specifies an allocation matrix based on characteristics thought to predict the variable being allocated. For example, in an allocation matrix predicting earnings, one cell might consist of white renters, aged 18-25, with a high school education. When earnings are not reported, they are imputed from the reported earnings of the last observation processed (typically geographically close) that falls in the same allocation cell.

[8] The increase in AHS family poverty rates was not statistically significant.

The main drawback to hot deck imputation is that it considerably limits the number of variables that can be included as predictors. The size of the allocation matrix rapidly becomes very large as more predictors are added. Omitting *any* variable correlated with the imputed variable is undesirable, since this will bias the correlation between the omitted and imputed variable towards zero.[9] In regressions of earnings on education, for example, the coefficient on education will be biased, presumably towards zero, if education is omitted from the allocation matrix.

Of particular importance here is that there is a strong relationship between the different types of income. For example, a worker earning a salary is unlikely to also receive retirement income (and vice versa). In other words, it is important to predict the receipt of each type of income based on whether the other types are received.

Due to these considerations, the 2005 AHS uses a "regression-based hot deck" to impute income. This method allows many more variables to be included as predictors, but retains the hot deck's advantages of processing simplicity and ability to flexibly replicate the distribution of the data.

The method can be summarized as follows:

1. Estimate a regression predicting the variable to be imputed, using the sample of completely reported data (the donors).
2. Split the donor data into hot deck cells using the predicted values from the regression in (1). Choose cutpoints that put approximately 500 donor observations in each cell.
3. Apply the regression coefficients from (1) and the cutpoints in (2) to the cases with missing data, thus assigning each recipient case to a hot deck cell.
4. Impute the missing data by copying the value from the "last" donor case in the same hot deck cell as the recipient case. The data are sorted so that the last donor will typically be geographically close to the recipient.

In practice, the variables indicating receipt (yes/no) of each of the nine types of income are imputed first. Cases that answered one or more of the receipt questions, but not all of them, are imputed using a regression-based hot deck. Table 6 shows the regression variables, and Table 7 displays the particular variables in the regression equations for each of the nine receipt (yes/no) variables.

The income amount variables are imputed next. Because in many cases there is only one amount to be imputed, a separate set of regressions are used to create the hot deck for these cases. Table 8 displays the variables used in these regressions. Table 9 displays the variables in the regressions where there are two or more amounts to be imputed. Cases that declined to answer all the receipt questions have both receipts and amounts imputed using a traditional hot deck, as illustrated in table 10.

[9] See, for example, Little, R. J. A. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* 6, 287– 296.

Hot Deck Regressions

Tables 7-9 summarize a series of regressions with numerous explanatory variables and it may be instructive to go through a few columns. As an example, the first column of Table 9 describes the hot deck regressions for the amount of Social Security income. This regression was estimated in the sample of persons reporting all income information and is used to impute income to those reporting all the income information except for the Social Security amount. The "X" in the first row means that this regression includes an indicator for race. Table 6 shows that race includes 2 categories: white, non-Hispanic and minority. The next "X" in the Social Security column of Table 9 indicates that the regression includes a set of indicators for sex interacted with 5 Age categories. So there are 10 indicator variables, including one that is omitted from the regression. For instance, one of the indicators is for men aged 16-24. Skipping to the bottom panel, we see that there are few variables capturing other types of income. There is one indicator for the receipt of retirement income. The other income variables were excluded due to lack of statistical significance and the author's judgment. The amount variables are shown after the comma. No amount variables are included because Social Security is the first income amount to be allocated and the other income amounts still include missing values.

Turning to the second, earnings, column of Table 9, we see that the earnings regression includes 4 age categories, rather than 5. For example, the 65-69 and 70+ categories used in the Social Security regressions are combined into a single 65+ category for the earnings regression (see Table 6). Another difference is that the earnings regression includes an indicator for whether the person worked in the previous week that was not included in the social security regression. The bottom panel shows that many receipt indicators were included in the earnings regression: Social Security, interest/dividends/rental income, SSI, welfare, self-employment, and other. Social Security amounts also enter into the regression. No other income amount variables are candidates for inclusion, since they have not been imputed yet.

Correlation Structure of Imputed Income

An evaluation of the imputation procedures is available separately.[10] In brief, the new imputation procedure better captured the correlation structure of the data than the traditional hot deck. Specifically, the study compared the correlation among income types in the imputed data and the reported data in both the AHS and the ACS (which uses a traditional hot deck). In almost every instance, the AHS imputations better captured the correlation structure of the reported data, than did the ACS imputations.

[10] Susin, Scott (2006). "Imputation via Triangular Regression- Based Hot Deck," 2006 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

In addition, the coefficients in a standard "wage regression" estimated with the imputed data showed substantially less bias in the 2005 AHS compared to either the 2003 AHS or the ACS. For example, the relationship between education and earnings in the 2005 AHS imputed data was very similar to the relationship in the reported data. By this measure, the ACS and older AHS imputations were less successful in reproducing the relationship between earnings and education found in the reported data.

Imputed Income Compared to Reported Income

The AHS asks about income in two steps, first asking a yes/no question about whether a particular type of income was received, and then asking for the amount. Table 11 compares the rates at which various types of income were received for reported and imputed data. The reported and imputed data are fairly similar. Because response rates for the yes/no receipt questions are high – 91 percent of families answer all receipt questions – these differences almost entirely disappear when comparing the reported data to the full sample data.

Table 12 presents quantiles of individual income amounts disaggregated into wage and non-wage income. Only those who received these types of income are included. Again, imputed and reported data are compared. Cases that are edited for consistency were tabulated separately, on the bottom panel.[11] They are excluded from the analysis here, because the editing often came after an imputation, and we are examining the imputation system here, not the combined imputation and editing systems.

Individual earnings tend to be lower in the imputed data than in the reported data, although the difference is slight in the higher percentiles. Individual non-wage earnings tend to be higher in the imputed data. Combining the two, total income is lower in the imputed data for the median and smaller quantiles, but reported and imputed income are quite similar at the 75th percentile and above. The distributions of the reported and imputed data will differ if the background characteristics of reporters and nonreporters differ. So it is not a surprise that income distributions differ somewhat between the reported and imputed data. Still, it is useful to know that the imputation procedures have only a modest effect on the income data.

When comparing the full sample to the reported data, the differences in the income amounts narrow, but not by as much as the receipt rates narrowed, since only 72 percent of families answered all the amount questions. There is little difference between the full sample and the reported data for the median and higher quantiles. For the 5th and 10th percentiles, full sample income is only 83% and 88% of reported income, respectively. However, the dollar amounts involved are small. For the 5th and 10th percentiles, full sample income is $300 and $600 lower than reported income, respectively.

[11] The consistency edits are most commonly invoked when a person reports a total income amount, but does not report complete information for all types of income. The edits allocate the total amount to the 9 categories of income.

Conclusion

For those with median or higher incomes, the revised AHS income questions resulted in data much closer to the ACS and other Census Bureau surveys, such as the CPS. There was little improvement, but also little deterioration, for most of those in the bottom half of the income distribution. For those with the lowest incomes (the bottom 5 percent) there was actually a fall in the amount of income reported. The 2007 AHS survey will address this issue by splitting the questions that combined multiple types of income and where reported rates of receipt declined sharply from previous years' surveys.[12]

The new imputation scheme appears to have better captured the relationship among income types, and between income and its predictors. This is welcome news, since many analyses are concerned with just such relationships. There should also be an improvement in the quality and accuracy of AHS cross-tabulations published by the Census Bureau.

[12] These two questions are those on "interest, dividends, net rental income, royalty income, or income from estates and trusts," and on "any other source such as Veteran's Administration (VA) payments, unemployment compensation, child support or alimony." These were split into "interest from savings, money market funds, IRA's, CDs, or other interest bearing accounts," "dividends from stocks," "rental income," "alimony or child support," and "unemployment compensation, any veteran's payments not already mentioned or any other income."

Table 1: Question Text for Income Items in the AHS National Survey

2003 (person level)	**2005 (all items person-level)[a]**
In the past 12 months, how much did . . . earn in wages, salaries, tips, and commissions before deductions? (SAL)	During the past 12 months, did . . . receive any wages, salary, tips, bonuses or commissions? (QSAL) How much did...receive? (SALQ)
2003 (family level)	
In the past 12 months, did . . . have a business, farm or ranch? (QBUS)	Did . . . receive any self-employment income during the past 12 months? (QSELF) What was the amount? (SELFQ)
. . . receive Social Security/pensions? (QSS)	Did . . . receive any Social Security or Railroad Retirement benefits during the past 12 months? (QSS) What was the amount? (SSQ) Did . . . receive any retirement or survivor pensions during the past 12 months? (QRETIR) What was the amount? (RETIRQ)
. . . have interest from savings, money market funds, IRAs, CDs, other? (QINT) . . . have dividends from stocks? (QDIV) . . . receive rental income? (QRENT)	Did . . . receive any interest, dividends, net rental income, royalty income, or income from estates and trusts during the past 12 months? Report even small amounts credited to your account. (QIDRI) What was the amount received? (IDRIQ)
. . . receive SSI, public assistance or welfare payments such as [fill state program name]? (QWELF)	Did . . . receive any Supplemental Security Income (SSI) payments during the past 12 months? (QSSI) What was the amount? (SSIQ) Did . . . receive any public assistance or public welfare payments from the state or local welfare office during the past 12 months? (QWELF) What was the amount? (WELFQ)
. . . receive workers' compensation or other disability payment? (QWKCMP)	Did . . . receive any disability payments such as SSDI, workers' compensation, veterans' disability or other disability payments during the past 12 months? (QWKCMP) What was the amount? (WKCMPQ)
. . . receive alimony/child support? (QALIM) . . . receive unemployment compensation, veterans's payments not already mentioned or any other income? (QOTHER)	Did . . . receive income on a REGULAR basis from any other source such as Veterans' Administration (VA) payments, unemployment compensation, child support or alimony during the past 12 months? (QOTALM) What was the amount from all sources? (OTALMQ)
In the past 12 months, what was the total income from: [source 1, source2, etc.] ?	

[a] 2005 Questions are not presented in the same order as on the questionnaire.

Table 2: Quantiles of Income: 2003 vs. 2005 AHS

Ratio: Full Sample 2005/2003

	N	5th	10th	25th	Median	75th	90th	95th
Household Income		0.88	1.00	1.02	1.09	1.04	0.99	0.75
Family Income		0.91	0.99	1.01	1.05	1.04	0.98	0.73
Family Earnings		0.83	0.95	1.00	1.07	1.03	1.02	0.68
Family Non-Wage Income		6.30	4.42	2.04	1.31	1.30	1.12	1.08

Full Sample: 2003 AHS

	N	5th	10th	25th	Median	75th	90th	95th
Household Income	48,197	4,800	9,477	20,760	40,124	74,000	120,000	205,000
Family Income	47,066	6,600	10,000	20,568	40,000	72,460	118,800	207,416
Family Earnings	37,866	6,000	11,529	24,000	42,000	73,000	113,000	220,223
Family Non-Wage Income	27,086	50	226	2,500	9,900	22,332	49,000	74,000

Full Sample: 2005 AHS

	N	5th	10th	25th	Median	75th	90th	95th
Household Income	43,360	4,200	9,468	21,156	43,850	77,000	119,000	154,000
Family Income	42,592	6,000	9,900	20,700	42,000	75,000	116,000	151,500
Family Earnings	32,050	5,000	11,000	24,000	45,000	75,000	115,000	149,500
Family Non-Wage Income	23,585	315	1,000	5,100	13,000	29,000	55,000	80,000

NOTE: Sample excludes zero amounts.

Table 3: Family Income Receipt Rates, by Income Types. 2003 vs. 2005 AHS.

	Proportion			
	AHS 2005	AHS 2003	Difference	Note on change from 2003 to 2005.
Social Security and Pensions	0.289	0.269	0.020	SS and pensions were combined in 2003, split into 2 questions in 2005.
Wages and Salaries	0.747	0.785	-0.037	
Social Security	0.256			
Interest, Dividends, Rental Income	0.162	0.318	-0.156	These were 3 questions in 2003, combined into one in 2005.
Supplemental Security Income	0.038			
Welfare	0.022	0.047		Welfare and SSI were combined in 2003, split into 2 questions in 2005
Workers' Compensation	0.049	0.030	0.019	
Self-Employment	0.131	0.095	0.036	In 2003, question referred to business income. In 2005, it refers to self-employment.
Other Income	0.068	0.103	-0.036	Other income and alimony were 2 questions in 2003, combined into one in 2005.
Sample Size	43,360	48,197		

NOTE: Data is weighted. Receipt of income by family members only, excluding nonrelatives of the householder.

Table 4: Quantiles of Household Income. ACS vs. AHS, 2003 to 2005.

	N	Mean	Std. Dev.	5th	10th	25th	Median	75th	90th	95th
Ratio of 2005 AHS/ACS		0.97	1.18	0.62	0.88	0.92	0.97	0.99	0.99	0.97
Ratio of 2003 AHS/ACS		1.08	2.24	0.72	0.89	0.94	0.93	1.00	1.06	1.38
2005 AHS	43,360	59,477	85,541	4,200	9,468	21,156	43,850	77,000	119,000	154,000
2005 ACS[a]	461,220	61,170	72,620	6,768	10,768	23,000	45,000	78,000	120,100	159,490
2003 AHS	48,197	62,263	142,146	4,800	9,477	20,760	40,124	74,000	120,000	205,000
2003 ACS[b]	139,376	57,493	63,479	6,648	10,611	22,105	43,155	74,000	113,158	148,097

NOTES: a. 2005 ACS interviews during the main AHS interviewing period (May, June, and July).
b. 2003 ACS interviews during the main AHS interviewing period (June, July, and August).

Table 5: Poverty Rates Across Surveys

	Households (1000s)		Percent in Poverty		
	2003	2005	2003	2005	Difference
AHS Households	105,874	108,901	13.2	13.9	0.7
Family Households					
AHS[a]	71,639	73,409	10.9	11.3	0.4
CPS[b]	75,616	77,019	9.6	10.2	0.6
ACS[c]	73,530	75,606	9.8	10.4	0.6

NOTE: AHS, CPS, and ACS poverty estimates are not exactly comparable. The table reports poverty in family households for all three surveys, but the AHS poverty measure is household income as a percent of poverty level for the household members, while the other two surveys use family income as a percent of poverty level for the family members.

a. Income for the 12 months prior to the interview period (Summer 2003 and 2005).
b. Income for calendar years 2002 and 2004.
c. Income for the 12 months prior to the interviewing period. Data include only interviews during the main AHS interviewing period (June, July, and August in 2003; May, June, and July in 2005).

Table 6: Income Predictor Variables and Values Used for 2005 AHS Regression-Based Hot Deck

Income Predictor Variable	Possible Values
Worked	Yes/No
Sex	Male/Female
Race	White, Non-Hispanic/Minority
Age4	16-18/19-24/25-64/65+
Age5	16-24/25-61/62-64/65-69/70+
Tenure	Own/Rent
Housing Cost	lowest quartile/2nd quartile/3rd quartile/highest quartile
Relationship to Householder	Head/Spouse/Child/Parent/Sibling or Other
Family Type	Single Mother/Married/Other
Kids	0/1/2/3+
Education	< High School/High School/Some College/College or more
Married	Married/Divorced/Widowed/Never/Missing
Citizen	Yes/No
9 Income Receipt Indicators	Yes/No
9 Income Amounts	Logarithm of positive amounts, otherwise zero

Table 7: Variables Used in Receipt Allocation Regressions

	Social Security	Earnings	Retirement	Interest/ Dividends/ Rental Income	SSI	Welfare	Workers Comp.	Self-Employment	Other
Worked*Race*Sex	X	X	X	X	X	X	X	X	X
Sex*Age4		X			X	X	X	X	X
Sex*Age5	X		X	X					
Worked*Tenure*Housing Cost	X	X	X	X	X	X	X	X	X
Worked*Relationship to Householder*Family Type	X	X	X	X	X	X	X	X	X
Worked*Relationship to Householder*Kids	X	X	X	X	X	X	X	X	X
Worked*Sex*Education	X	X	X	X	X	X	X	X	X
Worked*Citizenship	X	X	X	X	X	X	X	X	X
Income Receipt Indicators									
Social Security		X	X	X	X		X	X	X
Earnings			X	X	X	X	X	X	X
Retirement				X	X			X	X
Interest/Dividends/Rental Income					X		X	X	X
Supplemental Security Income						X		X	X
Welfare									X
Workers Compensation								X	X
Self Employment									X
Other									
Sample Size	74,896	74,896	74,896	74,896	74,896	74,896	74,896	74,896	74,896

Table 8: Variables in Amount Regressions for Those With Only 1 Amount to be Allocated

	Social Security	Earnings	Retirement	Interest/ Dividends / Rental Income	SSI	Welfare	Workers Comp.	Self-Emp.	Other
Race	X	X					X		
Sex									X
Age4					X	X			
Sex*Age4		X							
Sex*Age5	X		X	X					
Tenure			X						
Tenure*Housing Cost	X	X		X				X	
Married	X		X	X				X	X
Relationship to Householder					X				
Relationship to Householder*Family Type	X	X	X				X	X	X
Relationship to Householder*Kids		X		X		X			X
Education								X	X
Sex*Education	X	X	X						
Worked		X		X				X	X
Citizen	X		X						
<u>Income Receipt Indicators, Income Amounts</u>									
Social Security	_,_	X,X	X,X	X,X	X,X	X,X	X,X	_,_	_,X
Earnings	X,X	_,_	X,X	X,_	_,_	_,_	X,X	X,X	_,_
Retirement	X,X	X,X	_,_	X,_	X,X	_,X	_,_	_,_	_,_
Interest/Dividends/Rental Income	X,X	_,_	X,X	_,_	_,_	_,_	_,_	X,X	_,_
Supplemental Security Income	X,X	_,_	X,X	_,_	_,_	X,X	X,X	X,X	_,_
Welfare	X,X	_,_	_,_	_,_	X,X	_,_	X,X	_,_	_,_
Workers Compensation	X,X	X,X	_,_	_,_	X,X	X,X	_,_	_,_	_,X
Self-Employment	_,_	X,X	_,_	X,X	X,X	_,_	X,X	_,_	_,_
Other	X,X	_,_	_,_	_,_	_,_	_,_	_,_	_,_	_,_
Sample Size	10,063	36,520	4,793	5,584	1,413	717	1,706	4,691	2,536

Note: All amounts were entered as natural logarithms by recoding negative values to zero and then taking the natural log of the amount plus one.

Table 9: Variables in Amount Regressions for Those With 2 or More Amounts to be Allocated

	Social Security	Earnings	Retirement	Interest/ Dividends / Rental Income	SSI	Welfare	Workers Comp.	Self-Emp.	Other
Race	X	X					X		
Sex									X
Age4					X	X			
Sex*Age4		X							
Sex*Age5	X		X	X					
Tenure			X						
Tenure*Housing Cost	X	X		X				X	
Married	X		X	X				X	X
Relationship to Householder					X				
Relationship to Householder*Family Type	X	X	X				X	X	X
Relationship to Householder*Kids		X		X		X			X
Education								X	X
Sex*Education	X	X	X						
Worked		X		X				X	X
Citizen	X		X						
Income Receipt Indicators, Income Amounts									
Social Security	_,_	X,X	X,X	X,X	X,X	X,X	X,X	_,_	X,_
Earnings	_,_	_,_	X,X	X,_	_,_	_,_	X,X	_,X	_,_
Retirement	X,_	_,_	_,_	X,_	X,X	_,X	_,_	_,_	_,_
Interest/Dividends/Rental Income	_,_	X,_	_,_	_,_	_,_	_,_	_,_	X,X	_,X
Supplemental Security Income	_,_	X,_	X,_	_,_	_,_	X,X	X,X	X,_	_,_
Welfare	_,_	X,_	_,_	_,_	_,_	_,_	X,X	_,_	X,_
Workers Compensation	_,_	_,_	_,_	_,_	X,_	_,_	_,_	_,_	X,X
Self-Employment	_,_	X,_	_,_	_,_	_,_	_,_	_,_	_,_	_,_
Other	_,_	X,_	_,_	_,_	_,_	_,_	_,_	_,_	_,_
Sample Size	10,063	36,520	4,793	5,584	1,413	717	1,706	4,691	2,536

Note: All amounts were entered as natural logarithms by recoding negative values to zero and then taking the natural log of the amount plus one.

Table 10: Allocation Matrix for Family Non-reporters: Cell Sample Sizes for Complete Reporters

	Head/Spouse						Other Family Member					
	Age < 25 Male	Age < 25 Female	Age 25-64 Male	Age 25-64 Female	Age > 64 Male	Age > 64 Female	Age < 25 Male	Age < 25 Female	Age 25-64 Male	Age 25-64 Female	Age > 64 Male	Age > 64 Female
Worked, Lowest Productivity Quintile	351	657	159	1,545					234	275		
Worked, 2nd Productivity Quintile	277		1,388	4,499	303	318					combined with	
Worked, 3rd Productivity Quintile			2,861	3,747			1,516	1,351	637	336	Head/Spouse	
Worked, 4th Productivity Quintile	140	180	4,276	2,558	316	138						
Worked, Highest Productivity Quintile			6,102	696								
No Work, Lowest Productivity Quintile				2,014	150	1,976						
No Work, 2nd Productivity Quintile			668	2,897	685	1,640						
No Work, 3rd Productivity Quintile	180	546	810	1,817			2,423	2,331	677	732	146	397
No Work, 4th Productivity Quintile			788	1,104	2,171	791						
No Work, Highest Productivity Quintile			780	230								

Note: The rows indicate whether a person worked last week and their quintile of predicted wages. Productivity Quintiles are based on predicted earnings from a regression of the logarithm of annual earnings on Sex*Race, a quartic in age, Tenure*Housing Cost, Relationship to Householder*Family Type, Relationship to Householder*Kids, and Sex*Education.

Table 11: Proportion of Individuals Receiving Income Types, by Imputation Status. 2005 AHS

	Full Sample	Responded	Imputed	Difference	Response Rate
Social Security and Retirement	0.181	0.181	0.179	-0.002	0.95
Wages and Salaries	0.621	0.622	0.593	-0.029	0.96
Social Security	0.084	0.084	0.090	0.006	0.95
Interest, Dividends, Rental Income	0.101	0.101	0.100	-0.001	0.93
Supplemental Security Income	0.022	0.022	0.023	0.001	0.95
Welfare	0.012	0.012	0.010	-0.002	0.95
Workers' Compensation	0.028	0.027	0.028	0.001	0.95
Self-Employment	0.079	0.081	0.056	-0.025	0.95
Other Income	0.038	0.039	0.026	-0.012	0.95

Notes: Data are weighted. Sample Size is 80,615 individuals 16 years or more years of age. Nonrelatives are excluded.

Table 12: Quantiles of Individual Income for Persons with Non-Zero Amounts, by Edit/Imputation Status. 2005 AHS

Ratio: Imputed/Responded

	N	5th	10th	25th	Median	75th	90th	95th
Earnings		0.75	0.72	0.80	0.90	0.96	0.96	0.98
Non-Wage Income		1.02	1.22	1.19	1.11	1.08	1.13	1.10
Total Income		0.70	0.73	0.88	0.94	1.00	0.98	1.02

Ratio: Full Sample/Responded

	N	5th	10th	25th	Median	75th	90th	95th
Earnings		1.00	1.00	0.97	1.00	1.00	1.00	1.00
Non-Wage Income		1.02	1.07	1.08	1.01	1.01	1.03	1.01
Total Income		0.83	0.88	0.97	1.00	1.00	0.99	1.01

Full Sample (Responded & Imputed)

	N	5th	10th	25th	Median	75th	90th	95th
Earnings	47,870	2,000	5,000	14,560	30,000	50,000	75,000	100,000
Non-Wage Income	27,694	300	769	4,000	9,600	20,167	41,000	60,804
Total Income	65,170	1,500	4,200	11,000	25,000	45,000	71,000	96,800

Responded

	N	5th	10th	25th	Median	75th	90th	95th
Earnings	36,296	2,000	5,000	15,000	30,000	50,000	75,000	100,000
Non-Wage Income	20,443	295	718	3,700	9,464	20,000	40,000	60,000
Total Income	48,430	1,800	4,800	11,390	25,000	45,000	71,500	96,000

Imputed

	N	5th	10th	25th	Median	75th	90th	95th
Earnings	11,574	1,500	3,600	12,000	27,000	48,000	72,000	98,000
Non-Wage Income	7,251	300	875	4,400	10,488	21,600	45,000	65,740
Total Income	16,740	1,267	3,500	10,000	23,500	45,000	70,200	98,000

Addendum: Edited for Consistency

	N	5th	10th	25th	Median	75th	90th	95th
Earnings	1,742	2	1,500	8,061	22,000	42,915	77,200	108,845
Non-Wage Income	2,845	1	95	5,000	12,000	27,400	51,000	84,000
Total Income	4,545	2	1,000	8,400	21,000	44,500	80,000	115,000

Notes: Data are weighted. Sample consists of family members 16 years or more years of age. Nonrelatives are excluded.

Imputation via Triangular Regression-Based Hot Deck*†

Scott Susin
U.S. Census Bureau‡

Abstract

In principle, hot deck imputation methods preserve means and variances, and can also preserve covariances with other variables included in the allocation matrix. In practice, dimensionality problems arise quickly as predictive variables are added and allocation matrix cells become small, undermining the hot deck's theoretical advantages. Predictive-mean nearest-neighbor imputation avoids dimensionality problems, but can reduce the variance. A combination method is described: using the predicted values from a set of sequential, triangular regressions to form hot deck matrices. Triangularity allows the inclusion of predictive variables that are themselves subject to non-response. The method enables the rapid development of allocation schemes, eliminates dimensionality problems, and aids in predictor selection. The implementation of this method in American Housing Survey income data is described and evaluated.

Keywords: Imputation; Allocation; Predictive Mean; Sequential Regressions; Hot Deck

1 Introduction

This paper describes the income imputation system developed for the 2005 American Housing Survey (AHS), called a triangular regression-based hot deck. The method breaks little new theoretical ground. In fact, it could easily be improved on, albeit at some cost in additional computation and complexity. Nonetheless, the method represents a good compromise between current U.S. Census Bureau methods (the hot deck) and more sophisticated imputation schemes (such as chained regressions and multiple imputation).

The method has many desirable properties. It eliminates dimensionality problems, requires only weak assumptions about the distribution of the data, allows for the flexible imposition of logical constraints, and is robust to model misspecification. Compared to hot deck methods, the imputation scheme described reduces the amount of work required to develop an allocation scheme and provides the analyst guidance in creating the scheme. In practice, it is also better than a hot deck at imputing data that reproduces the relationships among variables (covariances) present in the reported data.

2 Hot Deck Methods

The hot deck allocation (or imputation) method is widely used at the U.S. Census Bureau and other statistical agencies. In this method, the analyst specifies an allocation matrix based on characteristics thought to predict the variable being allocated. For example, in an allocation matrix predicting earnings, one cell might consist of white renters, aged 18-25, with a high school education. When earnings are not reported, they are imputed from the reported earnings of the last observation processed (typically geographically close) that falls in the same allocation cell.

The hot deck method has several advantages. First, is its processing simplicity: the hot deck requires a single pass through the data, and can be easily implemented in statistical packages such as SAS that process a single observation at a time rather than holding a complete data set in memory. Of course, with the rapid development of computing technology, this processing simplicity has become less and less important, but still matters for very large datasets, such as the decennial census. Second, the hot deck preserves the distribution of the data. Recipient cases will have the same mean and variance as the donors. Importantly, the hot deck imposes no distributional assumptions on the data. Hence, it will also preserve other features of the data, such as heavy tails or heaping (when data are reported in rounded numbers).[1]

In principle, hot decks can also preserve the relationship between the allocated variable and other variables. In practice, however, dimensionality problems arise quickly, sharply limiting the number of variables that can be used. The American Community Survey (ACS) earnings allocation matrix, for example, uses 6 variables, with 2-20 categories in each, generating 3000 cells in the matrix.[2] Too many cells are undesirable, because small cells increase the usage of starting values, and increase the chance that a single donor will be used multiple times. The necessity for collapsing smaller cells, a process that requires considerable time and effort, is an important drawback of the hot deck method.

Dimensionality problems imply that many important variables are routinely omitted from allocation matrices. Economists have criticized the Census Bureau for not including education, a variable with strong theoretical foundations, in the Current Population Survey (CPS) earnings allocation matrix

*This paper is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

†Citation: Susin, Scott (2006). "Imputation via Triangular Regression-Based Hot Deck," 2006 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association: pp-pp.

‡U.S. Census Bureau, Housing and Household Economic Statistics Division, 4700 Silver Hill Road, Washington, DC 20233-8500; 301-763-5880; scott.j.susin@census.gov.

[1]The variance of the mean and other sample statistics calculated from the imputed data set will be understated, however (Rubin 1987)

[2]Occupation/class of worker (20 Categories) × Weeks worked (5) × Hours (3) × Age (5) × Sex (2)

(Lillard, Smith & Welch 1986). Education is not included in the ACS allocation matrix either, and it is easy to see why, since adding a four category education variable would expand the matrix to 12,000 cells! Many other variables strongly correlated with earnings are not included due to this dimensionality problem (for example, public assistance payments).

Omitting *any* variable correlated with the imputed variable is undesirable, since this will bias the correlation between the omitted and imputed variable towards zero. (see, e.g., Little (1988)). In regressions of earnings on education, for example, the coefficient on education will be biased, presumably towards zero, if education is omitted from the allocation matrix. Since the aim of statistical agencies is to produce general purpose allocations suitable for numerous different analyses, this is a strong argument for including as many variables as possible as allocation predictors, which is generally not possible with hot deck methods.

At the same time as it restricts the number of variables that can be used in the hot deck, the hot deck provides little guidance to the analyst creating the imputation matrix as to which variables should be used. Should education replace weeks worked in the earnings allocation matrix? Analysts typically rely on theoretical knowledge, knowledge of the literature, and intuition to make the choice. In principal, one could impute using several alternative matrices and examine summary statistics to compare them, or run auxiliary regressions to make such choices. However, this adds a substantial amount of additional work to what is already a cumbersome procedure.

3 Alternative Imputation Methods

Many alternatives to the hot deck have been proposed by statisticians. For a recent overview of various imputation methods, see Durrant (2005). The goal here is to find a method that is flexible (in the sense that it can be combined with a priori constraints on the data); simple to implement and justify to nonstatisticians; and that avoids placing parametric constraints on the data. Of course, the method should also reproduce the distribution of the reported data.

This study does not address methods for making valid inferences from statistics estimated from the final data set after imputation. Although it is possible to impute multiple values to each recipient case, in order to calculate standard errors that take into account the uncertainty due to imputation, that is not the focus here.

Multiple imputation, as proposed by Rubin (1987), entails much more than simply imputing several different values to each recipient. It also requires specifying a particular joint distribution of the data, typically multivariate normality. Methods which allow the distribution of the allocated variables to be determined by the data (nonparametrically), such as the hot deck, are often much more appealing. In the case of income data, multivariate normality is a particularly unattractive assumption, since the various components of income are clearly not distributed normally. Many income components, such as interest income, are constrained to be nonnegative, have a spike in the distribution at zero, and have heavy upper tails.[3] It must be noted, however, that multiple imputation does have the most fully developed body of theory.

Another alternative is predictive mean matching (Little 1988). This involves regressing the variable to be imputed on a vector of predictors (in the sample of donors with complete data). Next, predicted values are calculated for both the donors (with complete data) and recipients (with incomplete data). The donor with the closest predicted value to a particular recipient is chosen, and that donor's observed value (not the predicted value) is imputed to the recipient. This method is quite attractive, since it avoids the dimensionality problems discussed above and, by avoiding distributional assumptions, is likely to be robust to mispecification (Chen & Shao 2000). It also reproduces the distribution of the complete data. The drawback is that one case can easily be used as a donor multiple times. As as result, estimators calculated from the final data set may be inefficient (Durrant & Skinner 2006).

Finally, a number of authors have proposed using a series of equations to model the conditional distribution of each variable, in order to avoid multiple imputation's requirement of specifying a joint model of all the predictors and imputed variables. That is, instead of assuming, say, joint normality, one can specify a series of equations, mixing OLS regression with logit models, or anything else. This adds considerable flexibility to model nonnormal distributions as well as to impose logical consistency constraints, such as the fact that only homeowners can have a mortgage (Raghunathan, Lepkowski, van Hoewyk & Solenberger 2001, Buuren, Boshuizen & Knock 1999, Buuren, Brand, Groothuis-Oudshoorn & Rubin 2005). This study describes and tests a simple version of this type of imputation via sequential regression, combined with a version of the predictive mean matching approach.

4 Allocation Methodology

The problem here is impute a series of nine income variables (denoted $Y_i, i = 1 \ldots 9$): earnings, social security income, etc. Each of these variables may have missing data. The missing component of the data vectors are denoted Y^{miss}, and the reported components are Y^{obs}. These variables are to be imputed jointly, with the goal of preserving the covariance structure of the income components, as well as the relationship between Y_i and a set of predictor variables ($X_j, j = 1 \ldots J$) which do not have missing data. That is, the X_js have already been imputed using some simpler method.

The imputation method is a simple version of the chained equations method. Because the income components have a substantial fraction of zeros, and are mostly constrained to be nonnegative (with the exception of self-employment or business income), the imputation proceeds in two steps. First, we impute indicators for the receipt of each type of income (denoted $D_i, i = 1 \ldots 9$). Next, amounts (Y_i^{miss} are imputed for the cases with $D_i^{miss} = 1$. An example of a receipt variables is employment, with earnings being the corresponding amount

[3] Schafer & Olsen (1999) discuss alternatives to multiple imputation for this case.

variable.

The method can be summarized as follows:

1. Estimate a set of nine regressions predicting D in the sample of completely reported data. These regressions are sequential and triangular (see below).[4]
2. Split the observed data into hot deck cells using the predicted values from each regression in (1). Choose cutpoints that put approximately 500 observations in each cell, creating one hotdeck per income type.[5]
3. Apply the regression coefficients from (1) and the cutpoints in (2) to the cases with missing data, thus assigning each piece of missing data in each case to a hot deck cell.
4. Impute missing data using the nine hot decks in the usual manner.

In order to preserve the covariance matrix of the income amounts, it would be desirable to use other income variables as predictors. However, since all the income variables contain missing values, there is a problem of circular dependence: social security cannot be imputed until earnings are computed, and vice versa. To overcome this problem, we specify a set of sequential, triangular equations:

$$
\begin{aligned}
D_1^{obs} &= f(X^{obs}) \\
D_2^{obs} &= f(X^{obs}, D_1^{obs}) \\
D_3^{obs} &= f(X^{obs}, D_1^{obs}, D_2^{obs}) \\
&\dots \\
D_9^{obs} &= f(X^{obs}, D_1^{obs}, D_2^{obs}, \dots, D_8^{obs}).
\end{aligned}
$$

Table 1a shows the variables in X for the receipt equations. These equations are estimated using OLS regression in the set of cases with completely reported income data. Receipt of the first income type is imputed using only the X variables. Receipt of the second income type is imputed using X plus the first income type, and so on. The equations are ordered by the R^2 from the regression of D_i on the X variables.

Table 1a: Explanatory Variables in Typical Receipt Regression Model
Working (2) X Sex (2) X Race (2)
Sex (2) * Age (5)
Working (2) X Tenure (2) X Housing Cost (4)
Working (2) X Family Type (3) X Relat. to hhdr (5)
Working (2) X Kids (4) X Relat. to hhdr (5)
Working (2) X Sex (2) X Education (4)
Working (2) X Citizenship (2)
Receipt indicators (0 to 8 dummies)

[4]Although logits or probits would be more efficient, OLS regression is used here. OLS is still consistent and I prefer its functional form assumptions (the additivity of indicator coefficients) in this context.

[5]Putting 500 donors in each cell is basically an arbitrary number, chosen to limit the reuse of donor cases.

Then, for each income type, the reported data is split into a series of hot deck cells based on the predicted values from each regression, with the cutpoints chosen to put about 500 observations in each cell. For example, the 500 cases with the highest predicted earnings go into the first hot deck cell. The next 500 cases go into the second cell, and so on. The process is repeated for the other income types, each of which has its own hot deck.

Finally, imputation proceeds in the usual hot deck manner. The same coefficients (from the regression on the observed data) are applied to the nonreporters (cases with missing data). The missing values are filled from the donor case that is within the matching hot deck cell and is most recent in sort order (typically a case close in physical proximity).

This method of triangular equations is recommended by Raghunathan et al. (2001) (who call them "sequential" equations). An alternative would be to impute a set of starting values that don't condition on the other income types (perhaps using only X), and then iterate a number of times though $D_i = f(X, D_{-1})$, where D_{-1} are all the income receipt indicators other than D_i. We don't use this more sophisticated alternative because it would require multiple passes through the data, or the use of specialize software.

Having imputed the receipt indicators, the amount variables are imputed next. The allocation proceeds in a similar manner, except that we can now condition on the receipt variables, and the regressions are confined to those with positive amounts ($Y_i^{obs} > 0$). Specifically, a set of equations are estimated as follows:

$$
\begin{aligned}
Y_1^{obs} &= f(X^{obs}, D^{obs}) \\
Y_2^{obs} &= f(X^{obs}, Y_1^{obs}, D^{obs}) \\
Y_3^{obs} &= f(X^{obs}, Y_1^{obs}, Y_2^{obs}, D^{obs}) \\
&\dots \\
Y_9^{obs} &= f(X^{obs}, Y_1^{obs}, Y_2^{obs}, \dots, Y_8^{obs}, D^{obs}).
\end{aligned}
$$

As with the receipt variables, these equations are used to create a set of nine hot decks. The X variables for the amount equations are shown in Table 1b.

Table 1b: Explanatory Variables in Typical Amount Regression Model
Marital Status (4)
Race (2)
Sex (2) X Age (5)
Tenure (2) X Housing cost (4)
Family Type (3) X Relationship to householder (5)
Kids (4) X Relationship to housholder (5)
Sex (2) X Education (4)
Working (2)
Income Receipt Indicators (8 dummies)
Income Amounts (0 to 8 amounts)

The previous set of equations is used only for those with for those with two or more missing amounts. Since it is common for only one amount to be missing, these cases are treated

separately. For them, a different set of nine equations are estimated:

$$Y_i^{obs} = f(X^{obs}, D^{obs}, Y_{-1}^{obs}), Y_i^{obs} > 0$$

and these cases get their own set of nine hot decks.[6]

Cases with no reported income data (neither amounts nor receipts) are treated separately. For them, we aim to preserve the covariance matrix not by conditioning on the other income variables (since there aren't any variables to condition on), but by jointly allocating all the income variables, so that a recipient receives all its data from a single donor. For these cases, a traditional hot deck is created, since there is no obvious way to combine a set of nine separate equations and hot decks into a single matrix.

5 Comparing Reported and Imputed Data

Testing the imputations of the receipt variables is straightforward, as is testing the imputations of amount variables conditional on receipt. In both cases, donors are simply compared to recipients. However, for some purposes, we would also like to compare the amount variables without conditioning on receipt, that is, filling in zeros for those who do not receive a particular type of income. In particular, this kind of information is needed to compare the covariance matrices of the imputed and reported data. In the case of amounts-with-zeros, the right comparison is not obvious, so this section discusses all three comparisons in detail.

Consider a simple example with only one income type, which will be called "wages" for concreteness. Wage information is collected using two questions: "Are you employed?" and "What are your wages?" (asked of the employed). There are three possible patterns of nonresponse: (1) Answer employment and answer wages (if employed), (2) Answer employment, refuse wages, and (3) Refuse employment. In the last case, those who do not answer the employment question are not asked the wage question. Notice that the employment rate for group 2 is 100 percent, since only the employed get the opportunity to refuse the wage question.

Employment is imputed using as donors all who answered the employment question (groups 1 and 2) as donors for group 3 who refused to answer the question. Let the employment rate for the donors (groups 1 and 2 combined) be E. Assuming that there are no systematic differences between the donors and recipients, the expected value of group 3's employment rate is also E. Hence, to test whether the allocations are working, the donors should be compared to the recipients, within each cell of the imputation matrix.

Wages are imputed only for the employed (others have zero wages). In this case the donors are the employed members of group 1, who answered both questions. The recipients are the employed members of groups 2 and 3. Let average wages for the employed members of group 1 (the donors) be W. Then, again assuming no systematic relationship between wages and the reporting of wages (no systematic differences between donors and recipients), the expected value of wages in groups 2 and 3 combined is also W. So, again, the allocations can be tested by comparing the donors to the recipients.

In order to test the employment and wage imputations jointly, we can fill in zero wages for the non-employed and compare average wages in groups 1 and 2 (those answering the employment question) to average wages in group 3 (those not answering employment and having both employment and wages imputed. These two quantities should be equal. The employment rate of groups 1 and 2 is E, as assumed above, which implies that the employment rate of group 3 is also E. Assuming that there are no systematic differences among the employed members of the three groups, the wages of the employed are all equal in expectation to W (the average wages in group 1, conditional on employment). Including the nonemployed, expected wages in group 1 and 2 combined are EW, and the expected wages of group 3 are the same.

Note that groups 1 and 2 cannot be considered "donors," since those in group 2 had wages imputed. All members of group 3 are recipients, but the recipients in group 2 are combined with the donors in group 1 for the sake of this comparison. Although this comparison may seem a little odd, the discussion in the previous paragraph has shown that it is a valid one (the average wages of groups 1 and 2 combined should equal the average wages of group 3).

We might naively consider comparing group 1 (complete responders) to groups 2 and 3 (all with either wages or employment allocated). Although this seems like a natural comparison, these two groups do not have the same expected wages. Expected wages in groups 2 and 3 combined are the weighted average of group 2 with 100 percent employment and group 3 with E employment, or $QW + (1 - Q)EW$, for some $Q \in [0, 1]$. Let the employment rate in group 1 be E_1. We know that $E_1 \leq E$. That is, group 1 must have an employment rate no greater than the rate for groups 1 and 2 combined, since $E_2 = 1$ (the employment rate for group 2 is 100 percent). Clearly, $E_1W < QW + (1 - Q)EW$, since $E_1W < W$ and $E_1W < EW$. So the naive comparison is invalid. By including group 2 (with 100 percent employment) among the recipients, we inflate that groups average wages relative to the the complete responders.

6 Results

6.1 Means and Standard Deviations

The method described above is used on the American Housing Survey (AHS), starting with the 2005 survey. The AHS is a nationally representative survey which includes a battery of income questions quite similar to the U. S. Census Bureau's American Community Survey (ACS).[7]

Table 2 compares the means and standard deviations of the reported data and the imputed data in the AHS, labeled re-

[6] In principle a separate set of nine equations and hot decks could be created for each pattern of missing data. However, this would require 1023 regressions and hot decks. This is one reason for using a set of triangular equations.

[7] For a description of these data sets see http://www.census.gov/hhes/www/housing/ahs/ahs.html and http://www.census.gov/acs/www/.

Table 2: Reported and Imputed Means and Standard Deviations

	Reporters: Unadjusted				Nonreporters[a]: Unadjusted			
	Mean	Std Dev	Std Error	N	Mean	Std Dev	Std Error	N
SS	9,311	5,729	59	9,461	9,487	6,357	112	3,202
Wages	37,797	43,033	229	35,351	37,281	49,517	501	9,756
Retirement	15,632	18,877	279	4,566	14,370	15,232	411	1,375
Interest, dividends, rental	11,468	41,754	572	5,330	13,352	45,769	1,019	2,019
SSI	5,064	4,265	118	1,306	4,268	3,760	243	240
Welfare	2,786	3,030	117	670	2,099	2,107	175	145
Workers' Comp.	8,629	9,323	233	1,604	9,741	11,720	655	320
Self-emp.	31,747	62,176	1,037	3,593	31,232	70,202	2,005	1,226
Other	6,754	12,348	276	2,000	5,704	6,923	407	290

NOTE: [a]. Nonreporters had the amount indicated in the table imputed but also reported at least one receipt (complete nonreporters are excluded). Zeros and edited data are excluded.

Table 3: Regressions of Log Annual Earnings. Coefficients (SEs in parentheses)

	2004 ACS		2003 AHS		2005 AHS	
	Reported	Imputed	Reported	Imputed	Reported	Imputed
Intercept	7.18	7.56	7.24	7.97	7.21	7.28
	(0.01)	(0.02)	(0.04)	(0.07)	(0.03)	(0.06)
Years of Education	0.146	0.109	0.148	0.084	0.148	0.134
	(0.001)	(0.001)	(0.002)	(0.005)	(0.002)	(0.004)
Experience	0.107	0.103	0.097	0.086	0.106	0.104
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Experience squared X 1000	-1.860	-1.630	-1.680	-1.380	-1.880	-1.770
	(0.007)	(0.016)	(0.024)	(0.044)	(0.027)	(0.047)
Female	-0.483	-0.430	-0.493	-0.392	-0.469	-0.470
	(0.003)	(0.008)	(0.012)	(0.025)	(0.011)	(0.022)
Black	-0.122	-0.106	-0.096	-0.141	-0.052	-0.064
	(0.005)	(0.010)	(0.019)	(0.038)	(0.019)	(0.031)
Hispanic	-0.033	-0.032	0.031	-0.161	0.052	0.017
	(0.005)	(0.011)	(0.019)	(0.041)	(0.017)	(0.035)
N	534,320	89,732	42,771	16,664	36,296	11,574

Table 4: Selected Correlations Between Income Components. Correlation Coefficients with the Largest Differences Between Reported and Nonreported

	Wages/ Social Security	Wages/ Retirement	Wages/ Self-employment	Wages/ Interest	Social Security/ Retirement
AHS					
Reporters	-0.47	-0.26	-0.15	-0.06	0.43
Nonreporters	-0.56	-0.31	-0.07	-0.11	0.44
Difference	-0.09	-0.05	0.08	-0.05	0.01
ACS					
Reporters	-0.47	-0.29	-0.21	-0.10	0.44
Nonreporters	-0.37	-0.16	-0.03	0.00	0.36
Difference	0.10	0.13	0.18	0.10	-0.08

porters and nonreporters. Reporters are those who answered all the income amount questions. Nonreporters did not respond to at least one amount question. Those who answered no income questions at all (neither receipt nor amount) are excluded, because this group was imputed using a different method (a traditional hot deck).

In general the means and standard deviations in the imputed data are fairly close to the data from reporters. Four of the differences in means are statistically significant at the 5 percent level, but none are larger than $2,000 (interest: $11,468 vs. $13,352) or 25 percent (welfare: $2,786 vs. $2,099) different. Of course, we would expect the means and SDs to be different if the characteristics of reporters and nonreporters differ systematically. Hence, the top panel suggests that any systematic difference in response rates is relatively minor.

6.2 Wage Regression

An important goal of an imputation system is to preserve the relationship between the allocated variables and other variables in the data set. As a check of this, Table 3 estimates basic earnings regressions using the 2003 AHS, the 2004 ACS, and the 2005 AHS. Comparing the regressions on the reported data, many of the coefficients are remarkably close in the three data sets: the constant, education, experience (age - education - 6), experience squared, and female. Only the coefficients on Black and Hispanic differ appreciably.

The main comparison is between the regressions in the reported and imputed samples. The focus is on the education coefficient, the "return to education," which is the target of much attention from labor economists. In the 2004 ACS, the coefficient on education is 0.146 in the reported data but only 0.109 in the imputed data. The difference is statistically significant, and substantively large. This result is not surprising, since education is not included in the ACS allocation matrix. As noted above, it is impractical to include more than a limited number of variables in a traditional hot deck such as the ACS uses. The 2003 AHS, which also uses a hot deck that lacks an education variable, shows an even bigger difference. The return to education falls from 0.148 in the reported data to 0.084 in the imputed data. In the 2005 AHS, which uses the simple chained equations method discussed above, the coefficients are 0.148 (reported) and 0.134 (imputed), the closest of the three, although still showing some signs of bias.

The female coefficient shows a similar pattern, which is a little surprising since all three allocation methods use a sex variable (although the 2003 AHS combined female and minority into a single variable). The 2004 ACS imputed female dummy is 0.05 lower than the dummy in the reported regression, 0.10 lower in the 2003 AHS, but in the 2005 AHS the coefficient is almost exactly the same in both the imputed and reported data. Again we see that the triangular equations methods comes closest to reproducing the relationships in the original data, because it takes into account many more variables in determining the imputations.

6.3 Covariance Matrix

In both data sets, the covariance matrices for reporters and nonreporters are fairly similar.[8] Also noteworthy is the similarity of the correlation matrices of the reporters across data sets, with the exception of correlations with the interest variable.[9] In almost every case, the difference between reporters and nonreporters is smaller in the AHS than in the ACS.

Reporters are defined as those who answered all the receipt questions. They are compared with nonreporters, who had at least one receipt imputed. As discussed above, this is the correct comparison, rather than a naive comparison of donors versus recipients. In addition, cases with any edited data and complete non-reporters are excluded, as discussed previously.

Table 4 summarizes the correlation matrices, displaying the correlations for the five pairs of variables with the largest differences between the reporters and nonreporters (in either data

[8] The complete correlation matrices are available from the author or from www.huduser.org.

[9] The fraction reporting the receipt of interest, dividends or rental income is higher in the ACS than in the AHS, perhaps driven by differences in the interview modes (predominantly mailed questionnaires in the ACS and personal interviews in the AHS).

set). The AHS imputation system always comes closer to reproducing the correlations than the ACS. The improvement is smallest for Social Security versus wages (a difference of 0.08 in the AHS and -0.11 in the ACS). It is largest for retirement income versus salary (0.06 in the AHS; -0.13 in the ACS) and retirement income versus Social Security (-0.01 in the AHS; 0.08 in the ACS).

7 Conclusion

This paper has described an income imputation system which uses a triangular sequence of regression equations to define a series of hot decks. Although this method is less efficient than some more sophisticated alternatives described in the literature (because only a single pass through the data is made, rather than iterating) it is simpler to implement. We showed that this imputation method is capable of reproducing the distribution of the reported data fairly closely, if not perfectly. More importantly, the method was able to reproduce the correlations of the imputed variables with each other, and with other variables in the data set.

A number of comparisons suggest that this method represents an improvement over methods currently used at the U.S. Census Bureau. First, the paper showed that there is substantial bias when a wage regression, a model commonly used by labor economists, is estimated in wage data imputed using conventional Census Bureau methods.[10] When this regression is estimated in the 2003 AHS and 2004 ACS, both of which use a conventional hot deck, the return to education and the female wage penalty are biased towards zero. This is unsurprising: the dimensionality problems inherent in traditional hot deck methods sharply limited the number of variables that can be included in the hot deck. Neither data set included education in its hot deck, and the 2003 AHS did not include a sex variable directly. Regression-based hot decks, however, place no such limits on the number of variables that can be included. When implemented in the 2005 AHS, this methods was able to almost eliminate the bias.

Both methods do fairly well in reproducing the correlations among the various income variables. The ACS's traditional hot deck method did much better than might be expected, given that no explicit attempt is made to reproduce the correlation structure. Apparently the variables included in the hot deck are strong enough predictors that the correlation matrix for nonreporters is fairly similar to that in the reported data. However, for almost all correlation coefficients, the regression-based hot deck used in the 2005 AHS came closer to reproducing the correlation matrix of the reporters.

It should be noted that imputing income is harder in the AHS than in the ACS. Since the AHS has a sample size less than a tenth that of the ACS, the hot deck is constrained to have fewer cells, limiting the ability to find close matches for nonreporters. In addition, the AHS has fewer variables available for predicting income than the ACS. In particular the AHS has few labor force variables, such as occupation. The ability of the hot deck based on a sequence of triangular regressions to do "more with less" represents an impressive achievement.

[10] Including cases with an imputed dependent variable is a poor practice which will generally result in biased estimates, but it is a common practice.

References

Buuren, S. V., Boshuizen, H. C. & Knock, D. L. (1999), 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medicine* **18**, 681–696.

Buuren, S. V., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2005), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* . forthcoming.

Chen, J. & Shao, J. (2000), 'Nearest neighbor imputation for survey data', *Journal of Official Statistics* **16**(2), 113–131.

Durrant, G. B. (2005), 'Imputation methods for handling item-nonresponse in the social sciences: A methodological review.', *National Center for Research Methods Working Paper 002* . http://www.ncrm.ac.uk/publications/.

Durrant, G. B. & Skinner, C. (2006), 'Using missing data methods to correct for measurement error in a distribution function', *Survey Methodology* . forthcoming.

Lillard, L., Smith, J. P. & Welch, F. (1986), 'What do we really know about wages? the importance of nonreporting and census imputation', *Journal of Political Economy* **94**(3), 489–506.

Little, R. J. A. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* **6**, 287–296.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & Solenberger, P. W. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey Methodology* **27**, 85–95.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Schafer, J. L. & Olsen, M. K. (1999), 'Modeling and imputation of semicontinuous survey variables', *Proceedings of the Federal Committee on Statistical Methodology Research Conference* . http://www.fcsm.gov/99papers/.

Supplementary Table 1: 2005 AHS Correlation Matrix of Income Variables: Reporters vs. Nonreporters

No Receipts Imputed

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Workers' Comp.	Self-emp.	Other
SS	1.00	-0.47	0.43	0.18	0.02	-0.02	0.01	-0.05	-0.02
Wages	-0.47	1.00	-0.26	-0.06	-0.16	-0.07	-0.11	-0.15	0.02
Retirement	0.43	-0.26	1.00	0.18	-0.02	-0.03	0.03	-0.03	-0.01
Interest, dividends, rental	0.18	-0.06	0.18	1.00	-0.04	-0.03	-0.01	0.09	0.00
SSI	0.02	-0.16	-0.02	-0.04	1.00	0.11	0.04	-0.03	0.00
Welfare	-0.02	-0.07	-0.03	-0.03	0.11	1.00	0.01	-0.01	0.03
Workers' Comp.	0.01	-0.11	0.03	-0.01	0.04	0.01	1.00	-0.02	0.03
Self-emp.	-0.05	-0.15	-0.03	0.09	-0.03	-0.01	-0.02	1.00	0.00
Other	-0.02	0.02	-0.01	0.00	0.00	0.03	0.03	0.00	1.00

Some Receipts Imputed

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Workers' Comp.	Self-emp.	Other
SS	1.00	-0.56	0.44	0.19	0.03	-0.01	0.04	-0.07	-0.03
Wages	-0.56	1.00	-0.31	-0.11	-0.15	-0.04	-0.10	-0.03	-0.02
Retirement	0.44	-0.31	1.00	0.14	-0.02	-0.02	0.01	-0.04	-0.01
Interest, dividends, rental	0.19	-0.11	0.14	1.00	-0.05	-0.03	-0.01	0.07	-0.02
SSI	0.03	-0.15	-0.02	-0.05	1.00	0.08	0.10	-0.04	-0.02
Welfare	-0.01	-0.04	-0.02	-0.03	0.08	1.00	0.08	-0.03	0.04
Workers' Comp.	0.04	-0.10	0.01	-0.01	0.10	0.08	1.00	-0.03	0.03
Self-emp.	-0.07	-0.03	-0.04	0.07	-0.04	-0.03	-0.03	1.00	0.02
Other	-0.03	-0.02	-0.01	-0.02	-0.02	0.04	0.03	0.02	1.00

Difference, for cells with correlation > 0.05

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Workers' Comp.	Self-emp.	Other
SS		0.08	-0.01	-0.01					
Wages	0.08		0.06	0.05	-0.02	-0.03	-0.01	-0.12	
Retirement	-0.01	0.06		0.03					
Interest, dividends, rental	-0.01	0.05	0.03					0.02	
SSI		-0.02				0.04			
Welfare		-0.03			0.04				
Workers' Comp.		-0.01							
Self-emp.		-0.12		0.02					
Other									

NOTE: Cases with any edited income data, and complete non-reporters are excluded. All variables are measured in logarithms, with zero and negative values recoded to $1.

Supplementary Table 2: 2004 ACS Correlation Matrix of Income Variables: Any Imputation vs. None

No Receipts Imputed

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Self-emp.	Other
SS	1.00	-0.47	0.44	0.26	0.03	-0.02	-0.03	0.00
Wages	-0.47	1.00	-0.29	-0.10	-0.15	-0.06	-0.21	-0.01
Retirement	0.44	-0.29	1.00	0.23	0.00	-0.02	-0.03	0.02
Interest, dividends, rental	0.26	-0.10	0.23	1.00	-0.03	-0.03	0.09	-0.01
SSI	0.03	-0.15	0.00	-0.03	1.00	0.07	-0.02	0.00
Welfare	-0.02	-0.06	-0.02	-0.03	0.07	1.00	-0.01	0.03
Self-emp.	-0.03	-0.21	-0.03	0.09	-0.02	-0.01	1.00	-0.01
Other	0.00	-0.01	0.02	-0.01	0.00	0.03	-0.01	1.00

Any Receipts Imputed

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Self-emp.	Other
SS	1.00	-0.37	0.36	0.23	-0.06	-0.05	-0.08	0.02
Wages	-0.37	1.00	-0.16	0.00	-0.13	-0.04	-0.03	0.02
Retirement	0.36	-0.16	1.00	0.19	-0.06	-0.04	-0.04	0.03
Interest, dividends, rental	0.23	0.00	0.19	1.00	-0.07	-0.04	0.05	0.02
SSI	-0.06	-0.13	-0.06	-0.07	1.00	0.12	-0.04	-0.01
Welfare	-0.05	-0.04	-0.04	-0.04	0.12	1.00	-0.02	0.01
Self-emp.	-0.08	-0.03	-0.04	0.05	-0.04	-0.02	1.00	-0.01
Other	0.02	0.02	0.03	0.02	-0.01	0.01	-0.01	1.00

Difference, for cells with correlation > 0.05

	SS	Wages	Retirement	Interest, dividends, rental	SSI	Welfare	Self-emp.	Other
SS		-0.11	0.08	0.03				
Wages	-0.11		-0.13	-0.10	-0.02	-0.02	-0.19	
Retirement	0.08	-0.13		0.03				
Interest, dividends, rental	0.03	-0.10	0.03				0.03	
SSI		-0.02				-0.05		
Welfare		-0.02			-0.05			
Self-emp.		-0.19		0.03				
Other								

NOTE: Cases with any edited income data, and complete non-reporters are excluded. All variables are measured in logarithms, with zero and negative values recoded to $1.