Tax Variable Imputation in the Current Population Survey

Amy O'Hara, U.S. Census Bureau

The subject of this paper is to describe a new methodology of imputing tax variables to the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS). The U.S. Census Bureau had produced Federal and State tax estimates each year since 1979 for the ASEC.[1] These tax estimates are used to compute after-tax income. Income from the ASEC is adjusted by modeled tax estimates and other market income concepts. The most recent report using the tax model estimates is *The Effects of Government Taxes and Transfers on Income and Poverty: 2004,*[2] which indicates how money income is affected when capital gain estimates are incorporated; how postsocial insurance income is affected by combined payroll, Federal and State tax liabilities; and the specific impact of the Earned Income Tax Credit (EITC) on market income.

Other Federal agencies and research organizations model taxes using the CPS as well. Most of those models start with IRS data and supplement with CPS data to incorporate information on nonfilers. The CPS tax model is unusual because it starts with persons and households and models filing status. The Census Bureau constructs tax units based initially on marital status. IRS rules are applied to determine which household occupants are permitted to be in a tax unit together. The model assigns single, married joint, head of household, or nonfiler status. Survey data are supplemented with public-use IRS data for the tax variables required to estimate Federal taxes. Exemptions are determined, income is calculated, and tax credits and rates are applied. The State tax models use the Federal tax income and credit amounts as inputs.

In 2004, the Census Bureau launched a new tax model that better simulates the individual income tax return. The new model estimates more variables and credits than the previous methodology and improves on the distributions of variables released in the public-use file. The first data year to use the new methodology was the March 2004 ASEC that contained information for Tax Year 2003. The new model produced the estimates used in the 2005 report mentioned above and is used for ASEC 2005 forward.

In the new tax model, payroll taxes are calculated for private sector employees but are imputed for some public sector employees who are not covered by the Federal Insurance Contributions Act (FICA). Several inputs to adjusted gross income (AGI) are imputed: capital gains, capital losses, IRA contributions, self-employed health insurance deductions, and self-employed savings deductions. Taxable income is computed by subtracting imputed itemized deductions or the standard deduction from AGI. Federal taxes, credits, and marginal tax rates are derived from taxable income. Many of the tax estimates are released on the person-level public-use CPS ASEC file.

Different approaches have been used to impute tax variables in the CPS tax model. From 1979 to 2002, the old Census Bureau tax model randomly assigned mean amounts for capital gains, capital losses, itemized deductions, and childcare expenses from IRS aggregate tables. This resulted in an uneven distribution but reasonable weighted aggregate amounts because they were pegged in the imputation process. For ASEC 2004 and 2005, an unconstrained statistical match assigned amounts for the variables listed in Table 1. Common variables between the CPS ASEC and IRS Statistics of Income (SOI) public-use file were aligned to determine the closest match between the data sets. This statistical match informed the entire imputation: all variables from the most similar IRS SOI record were donated to the CPS ASEC record.

Table 1: Tax Variables Imputed in CPS ASEC 2004/2005

Capital gains
Capital losses
IRA contributions
Self-employed health insurance deduction
Self-employed savings deduction (SEP, SIMPLE and qualified plans)
Itemized deductions
Child and dependent care expenses

For these 2 years of production, imputed values were produced that were erratic in range, distribution, and aggregate amounts. Still, the variation in imputed amounts across all records and the fact that the variables were tied to one another (i.e., capital gains and itemized deductions coming from the same donor record) were an improvement over the previous method. However, the statistical match approach was complicated by the 3-year lag between the most recent SOI microdata file and the survey data year.[3] Both the incidence and dollar amounts of each imputed tax variable had to be ratio-adjusted to account for the lag. This was most problematic for Tax Year 2003, which used

the SOI public-use file from Tax Year 2000 because of capital gains. When attempting to apply values from the 2000 SOI to the 2004 ASEC, the match was manipulated to counter high capital gains and low capital losses due to divergent market conditions between the 2 years. Tax Year 2005, which is currently in production, will use the statistical match approach.

The limitations of these earlier methodologies have led to the development of a new imputation method. After evaluation, this new method will replace the current statistical match and functions as follows: A model-based approach is used to determine which records should have values assigned, and a Monte Carlo approach is used to assign amounts when indicated. The remainder of this paper describes the method and presents a comparison of its utility versus the earlier method.

Methodology

The model approach improves on the statistical match in two important areas. First, the method of assigning which records should receive a value is simplified by using logistic regression. While the strength of the statistical match relies on records common to both the CPS ASEC and SOI public-use file, the overlap of relevant variables is small because CPS ASEC contains no tax variables, and the SOI contains no demographic information. Additionally, income reported to the Census Bureau differs from income reported to the IRS.[4] Though the regression approach also relies on common variables, the improvement lies in incorporating all observations in the SOI and applying their normalized weights.

The second improvement is that amounts are assigned based on their IRS data distributions. The statistical match searched for the most similar SOI donor cases to apply values to the CPS base cases, and the match was run with replacement. Accordingly, the statistical match did not replicate the imputed variable distributions.

The Monte Carlo simulation of the missing tax variables incorporates means and standard deviations and controls for maximum values. Although the values presented in this paper derive from the SOI 2001 public-use file, future values may come from SOI data that are more recent than the public-use microdata file. Using the full SOI would nearly double the number of observations in the cells, improving the variance of the imputed values. The use of more recent data may make aging the data unnecessary.

For this analysis, the CPS ASEC 2005 internal research file is used. The ASEC records have been processed through the tax model to the point where filing status has been determined and exemptions have been counted. Only modeled filers are included. The SOI file is restricted to contain only nondependent, single, married joint, or head of household returns. Income-to-poverty

ratios (IPRs) based on the official poverty measure are constructed on both data sets. IPRs condition income amounts by family size; for this analysis, the total number of exemptions is used instead of the number of family members. Other indicator variables and transformations are created on both data sets. The model approach begins by partitioning both the CPS ASEC and SOI into self-employed and not self-employed filing units. Records with self-employment income are omitted; they will be processed separately in the future. The simulation of itemized deductions will be explained first, followed by capital gains.

Itemized Deductions

A logistic regression is run on the SOI data to determine the probability of having itemized deductions. Separate regressions are run for married and unmarried tax units, and weights are normalized. Two models were run because the incidence of itemizing appears to differ between married and unmarried filers. The unmarried group collapses single and head of household returns. The probability of itemizing deductions is modeled as a function of earned and unearned income variables, IPR, and whether the unit is in a State with no State income tax. Only SOI records with disclosed State values are included. Table 2 lists the weighted means of the explanatory variables used in the two regressions. The models both converge, and all explanatory variables are significant. The coefficients from the regressions are applied to the CPS data and transformed to compute the predicted probability of each CPS tax unit having itemized deductions. The adjusted R-squared value, predicted probability for CPS, and actual proportion of itemizers in the SOI data are presented in Table 2 for both the married and unmarried categories. Note that, in Table 2 and other tables, the estimates for CPS Tax Year 2004 are compared to SOI Tax Year 2001. This SOI public-use file is the most recent available for these experimental simulations.

The incidence of itemized deductions determined from the regression proceeds into the simulation stage. The numbers are not adjusted down to the SOI proportions, and the aggregates are not pegged to the SOI amounts in the following step. If this were done, income year CPS 2004 data would be pegged to 2001 SOI data.

To simulate the itemized deduction amounts, the predicted probability of itemizing for each CPS tax unit is compared to the SOI percentage of cases with itemized deductions. The percentage of married returns in SOI 2001 with itemized deductions is 53.54 percent. If the probability computed from the married regression is greater than or equal to this value, an itemized deduction amount is simulated.

Table 2. Itemized Deductions Regression, Weighted Means and Results

	Married	Married	Not Married	Not Married
	2004 ASEC	2001 SOI	2004 ASEC	2001 SOI
Number of observations	32,037	24,082	37,500	24,165
Total income (a)	8.18	6.39	3.28	2.84
Income tax free state indicator (b)	0.17	0.17	0.17	0.17
Presence of interest or dividends	0.2	0.33	0.12	0.17
Presence of retirement income ©	0.64	0.73	0.4	0.42
Presence of rent or royalty income	0.08	0.11	0.04	0.04
IPR	5.36	4.24	3.16	2.7

	Married	Not Married
Adjusted R2	0.79	0.72
ASEC TY04 predicted probability, wtd.	60.52%	24.41%
SOI TY01 incidence of itemizing, wtd.	53.54%	22.38%

(a) Total income is the sum of wages, interest, dividends, alimony, pensions and IRA distributions, Social Security, rental income, royalty income, and unemployment compensation.
(b) Seven states have no income tax: AK, FL, NV, SD, TX, WA and WY
(c) Retirement income is the sum of pensions, annuities and Social Security income.

SOI cases with itemized deduction amounts are partitioned by three variables as defined in Table 3. The mean itemized deduction amount and standard deviation for each of the 32 partitions are calculated. All CPS ASEC cases are partitioned in the same manner. If the predicted probability from the regression equals or exceeds the SOI percentage, a Monte Carlo simulation determines the amount of itemized deductions to be applied. A normal distribution is modeled. A dollar amount for itemized deductions is randomly selected from the distribution of each partition, controlling for mean and variance. The simulated values are constrained to be greater than zero[5] and less than the 99th percentile value from the SOI data for that partition.

Table 3. Partitions for Itemized Deductions Simulation

Filing status	Married
	Not married
State	State with no income tax
	State with income tax
Income percentile	10th percentile and under
	Over 10th to 25th percentile
	Over 25th to 50th percentile
	Over 50th to 75th percentile
	Over 75th to 90th percentile
	Over 90th to 95th percentile
	Over 95th to 99th percentile
	Over 99th percentile

Table 4 contains the simulation results for itemized deductions. The results are encouraging. Despite the 3-year lag between the data sources, the amounts assigned in the simulation follow similar trends to SOI-published aggregates. Because income is not adjusted in either data source, the impact of the coefficients from the regression may be magnified, resulting in a larger number of observations being assigned itemized deductions. This is a positive feature of the model since it accounts for growth in income and thus itemized deduction amounts over the 3 years. SOI does not disclose State values for high-income returns. Only SOI observations with disclosed States are included in this exercise, lowering the distribution of values being assigned and simultaneously reducing outliers.

Table 4. Itemized Deduction Simulation Results
Weighted number of observations and aggregate dollars in thousands

	ASEC TY04	SOI TY01
Number of obs.	47,493	43,499
Aggregate dollars	815,489,344	858,979,275
Mean	17,171	19,747
90th percentile	29,484	32,260
10th percentile	7,882	5,624

Capital Gains

This same method is applied to determine the probability of having capital gains. Capital gains are difficult to impute to the CPS due to limited understanding of when gains are realized. Literature analyzing wealth and investment typically pertains to acquisitions and views investments as stock amounts. To impute capital gains is to capture the act of converting a stock to a flow. Many factors contribute to the decision to sell off investments. Perhaps behavioral and financial factors could explain the decision, but such variables are absent in the CPS ASEC data. As SOI's are is the only available microdata with capital gain data by tax unit, the regression approach is being tested for assigning capital gain incidence to the CPS ASEC. Note that capital losses will be predicted and simulated separately; this section only discusses capital gains.

Capital gains are most prevalent among high-income filing units. The IRS disclosure proofs high-income returns on the SOI public-use file in various ways, including concealment of the State of residence. Nearly 40 percent of the returns with concealed States have capital gains, compared to less than 10

percent of returns with disclosed States. Due to the gap in incidence between the groups, for the regression, the SOI data are split by the presence/absence of a State code. The sample is further divided by filing status. To preserve cell sizes, single and head of household returns are again combined into an unmarried category, and married joint returns are labeled married. Using these divisions, four regressions are used to determine the odds ratios for capital gains. Table 5 lists the weighted means of the explanatory variables used in the capital gain regressions. The four models converge, and all explanatory variables are significant. As in the itemized deduction models, the coefficients from the regressions are applied to the CPS data and transformed to compute the predicted probability of each CPS tax unit having capital gains. The adjusted R-squared value, predicted probability for CPS, and actual proportion of capital gain recipients in the SOI data are presented in Table 5 for the four regression groups.

Table 5. Capital Gains Regression, Weighted Means and Results

	Married State disclosed	Married State disclosed	Not married State disclosed	Not married State disclosed
	ASEC 2004	SOI 2001	ASEC 2004	SOI 2001
Number of observations	30,786	24,082	37,337	24,165
Total income	0.07	0.06	0.03	0.03
Presence of earned income	0.93	0.88	0.92	0.88
Presence of retirement income	0.2	0.33	0.12	0.17
Interest or dividends > $1000 indicator	0.23	0.26	0.11	0.14
Income ≤ 150% of poverty indicator	\	\	0.29	0.37
Presence of rent or royalty income	0.08	0.11	\	\

	Married State withheld	Married State withheld	Not married State withheld	Not married State withheld
	ASEC 2004	SOI 2001	ASEC 2004	SOI 2001
Number of observations	1,241	40,561	163	8,453
Total income	0.34	0.47	0.41	0.45
Presence of earned income	0.97	0.93	0.98	0.76
Presence of retirement income	0.13	0.32	0.07	0.35
Interest or dividends > $1000 indicator	0.62	0.8	0.58	0.81
Income ≤ 150% of poverty indicator	\	\	\	\
Presence of rent or royalty income	0.22	0.28	0.21	0.28

	Married State withheld	Married State disclosed	Not married State withheld	Not married State disclosed
Adjusted R2	0.2855	0.3931	0.3278	0.4812
ASEC TY04 predicted probability, wtd.	31.66%	10.49%	34.18%	5.27%
SOI TY01 incidence of capital gains, wtd.	45.71%	12.63%	37.77%	6.23%

For the simulation stage, the married/unmarried and State disclosed/withheld categories from the regression are further divided by income amounts. Eight income cuts by percentile amounts are applied to the four groups, as shown in Table 6. Means and standard deviations calculated from these 32 partitions are used to simulate a capital gain amount for cases where the predicted probability of having capital gains meets or exceeds the proportion of SOI cases with capital gains. Once again, the simulated values are constrained to be

Table 6. Partitions for Capital Gains Simulation

Filing status	Married
	Not married
State	State disclosed
	State withheld
Income percentile	10th percentile and under
	Over 10th to 25th percentile
	Over 25th to 50th percentile
	Over 50th to 75th percentile
	Over 75th to 90th percentile
	Over 90th to 95th percentile
	Over 95th to 99th percentile
	Over 99th percentile

greater than zero and less than the 99th percentile value for that partition. For capital gains, a topcode of $2 million is applied. This choice was arbitrary but necessary to avoid extreme values in the ASEC that would inflate the aggregate because the ASEC weights are larger than those in SOI's. To avoid assigning capital gains to cases with income below the poverty line (IPR less than 1), their predicted probability from the regression is divided by four. More research is needed on high- and low-income persons and households in the ASEC to determine better parameters for these two restrictions.

Initial simulation results produced many large values due to the large variance around the SOI means. Particularly for the high-income group where the State was withheld, the standard deviations generated a wide distribution from which the imputed amounts are generated. To rein in the distributions, the standard deviations are reduced. Standard deviations for the high-income group are divided by four, and the standard deviations for all remaining cases are divided by two. Table 7 shows the impact of this restriction on the high-income SOI cases. Before the adjustment, ten of the sixteen partitions had standard deviations over one million. After the adjustment, only three partitions have standard deviations that large. The results are not as dramatic for the other sixteen partitions where State is disclosed. The largest reduction from halving the standard deviations in the State-disclosed partitions occurs for married returns at or below the 10th percentile of income, resulting in a reduction from $179,200 to $89,600. Though not as striking as the reductions for the high-income group, around a mean of $25,258 (for that particular partition), the impact is still great.

Table 7. Standard Deviation Adjustment for 16 High-income Partitions, Wtd. Dollars

Filing status	Income Percentile	Mean	Std. Deviation	Std. Deviation/4
Not married	10th and under	537,574	4,358,235	1,089,559
State withheld	To 25th	241,526	845,805	211,451
	To 50th	84,617	877,429	219,357
	To 75th	114,386	866,923	216,731
	To 90th	220,899	1,031,169	257,792
	To 95th	470,545	1,145,453	286,363
	To 99th	639,964	1,043,259	260,815
	Over 99th	3,727,227	5,064,941	1,266,235
Married	10th and under	331,399	2,537,063	634,266
State withheld	To 25th	46,336	538,526	134,631
	To 50th	54,441	557,185	139,296
	To 75th	90,015	685,805	171,451
	To 90th	209,215	1,008,930	252,232
	To 95th	388,587	1,182,115	295,529
	To 99th	775,545	1,366,466	341,617
	Over 99th	3,130,135	4,173,965	1,043,491

The results of the capital gain simulation are presented in Table 8. It is challenging to compare the CPS ASEC results to the SOI. The aggregates should not match. The underlying data differ in terms of sample selection, weighting factors, and income reporting. Table 8 shows the initial simulation results, followed by the results after reducing standard deviation amounts. These columns may be carefully compared to values in the last two columns using the State-restricted and full SOI, respectively. The ASEC results fall between the State-restricted and full SOI samples. Again, the SOI data have not been aged or otherwise adjusted to account for the 3-year lag between the samples. While these findings appear promising, further analysis is needed to determine a more appropriate benchmark for the ASEC results.

Table 8. Capital Gains Simulation Results
Weighted number of observations and aggregate dollars, in thousands

	ASEC TY04 Pre-σ adj.	ASEC TY04 Post-σ adj.	SOI TY01 State>0	SOI TY01 All Records
Number of obs.	9,721	9,721	11,229	12,239
Agg. dollars	400,556,727	196,169,936	76,420,031	321,862,140
Mean	41,205	20,180	6,806	26,298
90th percentile	30,625	17,575	17,190	26,400
10th percentile	2,478	1,837	49	52

Evaluation

Analyzing the CPS ASEC tax estimates is challenging because no current-year data are available from the IRS before the estimates are released. The data are evaluated against the previous year's ASEC amounts for consistency and against the previous year's published IRS aggregates. This experimental data exercise uses last year's ASEC data (Survey Year 2005, Tax Year 2004), so that these results can be compared to the statistical match approach. Table 9 compares estimates of the two imputed variables using the two approaches to the SOI 2001 data.[6] Results for itemized deductions appear reasonable. In the simulation, itemized deductions are imputed to more single returns than in the previous approach. Results for the other two filing categories are more consistent. Looking at the capital gain results in the lower panel, the incidence of capital gains is high for single and married returns, but the amount simulated moderates the impact.

Table 9. Simulation Approach vs. Statistical Match
Weighted number of observations and aggregate dollars, in thousands

		ASEC TY04 Simulation	ASEC TY04 Statistical Match	SOI TY01
Itemized Deductions				
Single	Count	17,610	14,041	11,360
	Aggregate	268,615,490	163,578,175	170,315,736
Married joint	Count	26,045	22,799	27,719
	Aggregate	487,924,372	446,377,249	622,819,358
Head of household	Count	3,838	3,317	3,418
	Aggregate	58,949,482	48,003,891	50,841,624
Total	Count	47,493	40,156	42,498
	Aggregate	815,489,344	657,959,315	843,976,718
Capital Gains				
Single	Count	7,093	4,782	4,505
	Aggregate	64,602,862	43,251,112	65,203,332
Married joint	Count	10,095	8,464	7,147
	Aggregate	209,462,936	156,502,425	240,845,774
Head of household	Count	885	427	448
	Aggregate	8,804,638	5,533,924	7,447,505
Total	Count	18,072	13,673	12,099
	Aggregate	282,870,436	205,287,461	313,496,611

Conclusion and Future Work

The nonstatistical match approach to imputing tax variables seems promising; the results appear more stable than the previous methods. Once the best regressors are determined, they can be applied annually. The statistical match had to be manipulated each year to address outliers, and the data were aged forward to align income. If this new approach is adopted, estimates from the after-tax income file will become more consistent.[7]

Improving the imputed variables will directly impact the alternative definitions of income reported by the U.S. Census Bureau. Looking at the three alternative definitions reported last year, Market and Postsocial Insurance Income definitions include capital gains and losses, while Disposable Income also includes payroll taxes, Federal income taxes, and State income taxes. The imputation process for capital gains and losses not only has an embedded impact in the tax calculations, but the amounts are also viewed as "income" in these alternative definitions. It is important to simulate a reasonable distribution of capital gains and losses while recognizing that the SOI and ASEC have different samples. This exercise has proceeded assuming that a low-income person in CPS is not always equivalent to a low-income tax filer for a variety of reasons. Not all low-income persons file a tax return; many do not meet the filing requirement threshold, but some file to apply for credits or to recapture withholding. Also, some SOI cases appear to be low-income when their capital gains are excluded. More research is needed to understand the differences in low-income cases between the data sources.

Future work includes a more precise evaluation of the approach using the linked ASEC-Individual Master File (IMF) data set. These commingled data allow a comparison of the actual administrative data from the tax return with the modeled information from the CPS ASEC. The U.S. Census Bureau is only permitted to receive certain income fields and does not receive amounts for capital gains or itemized deductions. A flag indicating whether a Schedule A or D was included with the return is included. These data will allow an analysis of the regression portion of the new methodology that determines the probability of receiving an imputed value. Different specifications of the regression will be tested to improve goodness of fit. The linked data will also be used to test extensions of the method. Modeling the joint distribution of certain variables is desirable in the future; tests on the linked data should indicate the applicability of such an approach.

Endnotes

[1] Previously called the March Supplement to the CPS.

[2] Report located at http://www.census.gov/hhes/www/poverty/effect2004/effectofgovtandt2004.pdf. For income year 2003, two reports were released. *Alternative Income Estimates in the United States: 2003* and *Alternative Poverty Estimates in the United States: 2003*. These are located at http://www.census.gov/prod/www/abs/income.html.

[3] SOI 2000 was used for CPS ASEC 2004; SOI 2001 was used for CPS ASEC 2005.

[4] Marc Roemer used Detailed Earning Records to evaluate CPS wage data in *Using Administrative Earnings RecordsTto Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation* (2002). Research is currently under way comparing CPS ASEC income to IRS reported income using a linked data file.

[5] A positive value is assigned because the logistic regression indicates the tax unit should receive an amount.

[6] Note that SOI 2001 was used for both the released statistical match-based imputations and the experimental simulation approach presented in this paper. For the statistical match, the values were aged and constrained to IRS published aggregates.

[7] The lag between the ASEC and public-use SOI files has resulted in inconsistent imputed values, particularly when viewed as a time series. The new approach should ease these erratic values and stabilize the time series.