Sequential Regression Multivariate Imputation in the Current Population Survey Annual Social and Economic Supplement

Charles Hokayem[1], Trivellore Raghunathan[2], and Jonathan Rothbaum[3]
[1]Centre College, Department of Economics, 600 West Walnut Street, Danville, KY 40422
[2] University of Michigan, Survey Research Center, Institute for Social Research, 426 Thompson Street, Ann Arbor, Michigan 48106
[3]U.S. Census Bureau, Social, Economic, and Housing Statistics Division, 4600 Silver Hill Road, Washington, DC 20233

Abstract
The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) serves as the data source for official income, inequality, and poverty statistics in the United States. The Census Bureau has used a "hot deck" procedure to impute missing income values since 1962. This paper implements an alternative model-based methodology, sequential regression multivariate imputation (SRMI), to impute missing income values in the CPS ASEC. SRMI offers several potential advantages over the current hot deck method, including 1) greater flexibility to add additional covariates and 2) accounting for uncertainty in the imputation process. We implement a baseline SRMI with data from the 2011 CPS ASEC and then augment this with tax records on earnings from the Social Security Administration's Detailed Earnings Records (DER) file. We compare imputed income values from SRMI to those from the hot deck procedure along several dimensions including the median, variance, and poverty.

Key Words: CPS ASEC, SRMI, Multiple Imputation, Income Statistics

1. Introduction

The accurate measurement of the income distribution is vital to assessing economic growth, characterizing income inequality, and gauging the effectiveness of the federal safety net. The Current Population Survey Annual and Social Economic Supplement (CPS ASEC) serves as an important source of the income distribution for the United States. Like other surveys, the CPS ASEC suffers from growing nonresponse to income questions over time. For example, the share of all income that is imputed due to nonresponse was 34.7% in 2011. There is a concern that the increased nonresponse to income questions could deteriorate income data quality and distort statistics derived from income such as poverty and inequality. The CPS ASEC relies on a hot deck imputation procedure to address income nonresponse. The current procedure has been in place with few changes since 1989. It fills in missing data by matching observations with missing data to observations with complete data based on socioeconomic characteristics.

Considerable advances have been made in imputation methods since the initial CPS ASEC hot deck procedure was adopted in 1962. This paper implements one of these

methods, sequential regression multivariable imputation (SRMI), to impute missing income in the CPS ASEC. Unlike the hot deck procedure, SRMI is a model-based method that has a few key features. First, it allows for greater flexibility than the hot deck procedure and allows for the inclusion of additional covariate variables. Second, it accounts for uncertainty in the imputation process. Some Census surveys such as the Survey of Income and Program Participation have already adopted the SRMI method. We implement SRMI with data from the 2011 CPS ASEC matched to Social Security Detailed Earnings Records (DER) which contain earnings information derived from W-2 forms. We implement two versions of SRMI: (1) SRMI only using survey data as predictors and (2) SRMI that adds W-2 earnings as predictors. We compare imputed income values from each version of SRMI to imputed values from the hot deck procedure along several dimensions, including median income, variance, and poverty.

2. Background

2.1 CPS ASEC Hot Deck Procedure

The Census Bureau has used a hot deck procedure for imputing missing income since 1962.[1] The current system has been in place with few changes since 1989 (Welniak 1990). The CPS ASEC uses a variation of the cell hot deck procedure to impute missing income and earnings data in the monthly CPS.[2] The cell hot deck procedure assigns individuals with missing income values that come from individuals with similar characteristics. The hot deck procedure for the CPS ASEC income variables relies on a sequential match process. Here we describe the process for earnings imputation. The process is similar for other income sources. First, individuals with missing earnings data are divided into one of 12 allocation groups defined by the pattern of nonresponse. Welniak (1990) lists the 12 allocation groups and nonresponse patterns. Examples include a group that is only missing earnings from longest job or a group that is missing both longest job and earnings from longest job. Second, an observation in each allocation group is matched to another observation with complete data (the "donor") based on a set of socioeconomic variables, the match variables. If no match is found based on the set of match variables, then match variables are dropped and variable definitions are collapsed to be less restrictive at the next match level. This process of sequentially dropping variables and collapsing variable definitions is repeated until at least one match is found. When a match is found, the missing amount is substituted with the reported amount from a matched record. At the first match level, there are 16 match variables and over 620 billion cells, at the second level, there are 14 variables and 17 billion cells, at the third level, there are 12 variables and 3.8 million cells, and at the sixth and final match level, there are 4 variables and 96 cells.

The ASEC also uses a hot deck procedure for whole supplement nonresponse. In this context, imputation refers to an individual who responds to the monthly basic CPS but does not respond to the ASEC supplement and requires the entire supplement to be imputed. To be considered a donor for supplement imputations, an ASEC respondent has to meet the minimum requirement that at least one person in the household has answered one of the following questions: worked at a job or business in the last year; received

[1] The term hot deck comes from storing data with computer punch cards and refers to the deck of cards of available donors for a nonrespondent. The deck was "hot" as it was being used for processing.

[2] Bollinger and Hirsch (2006) describe the cell hot deck procedure used in the monthly CPS.

federal or state unemployment compensation in the last year; received supplemental unemployment benefit in the last year; received union unemployment or strike benefit in the last year; or lived in the same house one year ago. This requirement implies that whole supplement donors do not have to answer all the ASEC questions and can have item imputations. Similar to the sequential hot deck procedure, the match process sequentially drops variables and makes them less restrictive until a donor is found. Whole supplement imputations account for about 13 percent of all ASEC supplement records.

Since donors come from observed data, the hot deck procedure offers the advantage that it imputes plausible values of missing income. It also preserves multivariate relationships. It does not require fitting a model, so it can potentially be less sensitive to model misspecification than an imputation method based on a parametric model (Andridge and Little, 2010). The procedure has its shortcomings. Earlier versions of the procedure omitted important determinants of income and earnings such as education and region of residence (Lillard et al, 1986). By using a single imputation, the current hot deck procedure does not account for imputation uncertainty so has the effect of understating standard errors. Due to the sparseness of the donor cells, donors can be used several times during the process.

The assessments of the CPS ASEC hot deck procedure are rather old. David et al (1986) use the March 1981 CPS file matched to IRS records to compare the procedure to regression methods that add residuals to predicted values of missing wage and salary. They find the hot deck procedure performs quite well, producing lower mean absolute error and mean relative error. Lillard et al (1986) examine the difference between average income of respondents and nonrespondents in the March 1980 CPS and suggest the procedure can severely underestimate income for certain occupations such as judges and lawyers.

2.2 Sequential Regression Multivariate Imputation (SRMI)

The sequential regression multivariate imputation (SRMI) is a pragmatic iterative approach to multiply impute the missing values in each variable using all other variables as predictors (Raghunathan et al., 2001). Various other names have been given to this approach such as Fully Conditional Specification or Flexible Conditional Models etc. Specifically, suppose that U is a collection of variables with no missing values and $Y_1, Y_2, \ldots, Y_p$ are the p variables with missing values. Though it is not necessary, suppose that the variables are ordered by number of missing values from lowest to the largest (the pattern of missing data, however, is arbitrary). An alternative approach is to order on the basis of dependence on other variables from "least dependent" to "most dependent". However, the ordering will have no effect, as the imputed values on any variable will eventually depend on all other variables.

In the first iteration, Y_1 is regressed on U and the missing values are imputed. An explicit regression model, a hot deck or predictive mean matching may be used to create imputed values. Let $Y_1^{(1)}$ denote the filled-in version of the variable Y_1. Now Y_2 is imputed using $\left(U, Y_1^{(1)}\right)$ as covariates. Let $Y_1^{(2)}$ denote the filled-in version of Y_2. This process continues until the missing values in Y_p are imputed using $\left(U, Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{p-1}^{(1)}\right)$ as predictors.

We cannot stop at iteration 1 because imputation of Y_1, for example, fails to exploit the observed information from $(Y_2, Y_3, \ldots, Y_p)$. The iteration $t = 2,3,\ldots$ proceed in the same manner except that all other variables (with some filled at the current and the rest in the previous iterations) are used in imputing each variable. Specifically, at iteration 2, Y_1 is re-imputed using $\left(U, Y_2^{(1)}, Y_3^{(1)}, \ldots, Y_p^{(1)}\right)$ as predictors; Y_2 is re-imputed using $\left(U, Y_1^{(2)}, Y_3^{(1)}, \ldots, Y_p^{(1)}\right)$ as predictors etc.

In general, at iteration t, Y_j is re-imputed using $\left(U, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)}\right)$ as predictors. The iterations are continued several times in order to fully use the predictive power of the rest of the variables when imputing each variable. Empirical analysis has shown that fewer than 20 and generally as few as 5 to 10 iterations are sufficient to condition the imputed values in any variable on all other variables (Ambler and Royston, 2007; van Buuren, 2007; He et al., 2009).

3. ASEC and DER Data Description

The data used for the analysis come from the internal Current Population Survey Annual Social and Economic Supplement (CPS ASEC) for survey year 2011 (reporting income for 2010). The Census internal CPS ASEC is matched to the Social Security Administration's Detailed Earnings Record (DER) file. The Detailed Earnings Record file is an extract of Social Security Administration's Master Earning File (MEF) and includes data on total earnings, including wages and salaries and income from self-employment subject to Federal Insurance Contributions Act (FICA) and/or Self-Employment Contributions Act (SECA) taxation. Since individuals do not make SECA contributions if they lose money in self-employment, only positive self-employment earnings are reported in the DER file (Nicholas and Wiseman 2009). The DER file contains all earnings reported on a worker's W-2 forms (and 1099 if self-employed). These earnings are not capped at the FICA contribution amounts and include earnings not covered by Old Age Survivor's Disability Insurance (OASDI) but subject to Medicare tax. The DER file also contains deferred wages such as contributions to 401(k), 403(b), 408(k), 457(b), 501(c), and HSA plans. The DER file is not a comprehensive source of gross compensation. Abowd and Stinson (2013) describe parts of gross compensation that may not appear in the DER file such as pre-tax health insurance premiums and education benefits. It also cannot measure off-the-books earnings (Roemer 2002; Hokayem, Bollinger, and Ziliak, Forthcoming).

The Census Bureau's Center for Administrative Records Research and Applications (CARRA) matches the DER file to the CPS ASEC. Since the CPS does not currently ask respondents for a Social Security Number, CARRA uses its own record linkage software system, the Person Validation System, to assign a Social Security Number.[3] This assignment relies on a probabilistic matching model based on name, address, date of birth, and gender (NORC 2011). The Social Security Number is then converted to a Protected Identification Key. The Social Security Number from the DER file received from SSA is also converted to a Protected Identification Key. The CPS ASEC and DER files are matched based on the Protected Identification Key and do not contain the Social

[3] Respondents are automatically matched to the DER unless they notify Census otherwise through the website or a mail-in form; an "opt-out" consent option.

Security Number. The 2011 ASEC-DER match rate is 89.7 percent. We collapse the file into one earnings observation per worker by aggregating total compensation (Box 1 of W-2), SSA covered self-employment earnings (SEI-FICA), Medicare covered self-employment earnings (SEI-MEDICARE), and deferred contributions across all employers. We define DER earnings as the sum of total compensation and deferred contributions plus the maximum of SSA covered self-employment income or Medicare covered self-employment: DER Earnings = Box 1 of W-2 + Deferred Contributions + max(SEI-FICA,SEI-MEDICARE).

4. Implementing SRMI for CPS ASEC Income

The 2011 CPS ASEC sample includes 96,958 addresses and 204,983 individuals. We impute missing income for individuals aged 15 and older (156,849 individuals). We impute income for 20 categories: wage and salary earnings, self-employment earnings, unemployment compensation, workers' compensation, Social Security, Supplemental Security Income, public assistance, veterans' benefits, survivors' benefits, disability benefits, retirement income, interest income, dividend income, rental income, education assistance, child support income, alimony income, financial assistance, and other income.

As discussed in Section 0, there are two reasons that income information could be missing in the CPS ASEC, item non-response and supplement non-response.[4] In Table 1, we show the non-response rates for each income type imputed in this paper. For earnings from the longest job, only 0.1% of individuals did not respond to the recipiency question, but 12.7% did not respond to the value question. However, because 12.9% of individuals were supplement non-respondents, 25.7% of individuals had their earnings from the longest job imputed. Non-response rates are highest for interest income (16.5%), earnings from longest job (12.7%), dividend income (6.6%), and Social Security (4.4%). In total, 34.7% of total income in the CPS ASEC is imputed due to item and supplement nonresponse.

SRMI modelling for each binary variable was implemented using a logistic specification. For each continuous variable, such as income, ordinary least squares (OLS) was used. However, the distribution of income is rarely conditionally normally distributed. We use the empirical normal transformation as it both ensures normality in all cases and is not affected by the presence of negative values, which is possible for some income types.

In addition to income recipiency and value, we also model other labor force related variables, such as weeks worked last year, hours worked per week, and occupation. While these variables are present for most respondents, they are missing for the 12.9% of observations that are supplement non-respondents. Imputation of occupation group presents a particular challenge. It is not feasible to model the probability of working in one of the over 500 4-digit occupation categories. Instead, we divide occupation into 11

[4] In addition, CPS households can be classified as Type A, B, or C non-interview households. Type A non-interview households are those that the field representative determines as eligible for CPS response, but from which no useable data were collected. No imputation is done for Type A non-interviews. Type B and C non-interview households are those that are not eligible for CPS interview. For example, if the housing unit was converted to a permanent business, condemned or demolished, it is classified as a Type C non-interview. If no eligible individuals occupy the housing unit, but the unit is still intended for occupancy, it is classified as a Type B non-interview.

categories.[5] We separated these 11 occupation groups into a series of binary categories connected by the tree structure fully described in our working paper (Hokayem, Raghunathan and Rothbaum 2015). In the imputation process, each individual with a missing occupation progresses through the occupation tree using logistic models until they are assigned an occupation category.

The most significant challenge to applying SRMI to the CPS ASEC income variables is selecting the models for each imputed variable. In order to avoid omitted variable bias in the imputation model, we would like to include as many potential predictors as possible. However, if we include too many variables, we run the risk of overfitting the model.

Our list of potential predictors include the recipiency and value variables for each income type, gender, relationship to householder, education dummies, marital status, cohabiting partner status, spouse/partner earnings, number of children in household (under 18 and under 6), urban/rural status, small or large metropolitan area, Census region, public housing, energy assistance benefits, Supplemental Nutrition Assistance Program benefits, health insurance status and type (Medicaid, Medicare, VA, private, etc.), renter/homeowner, unemployment, school enrollment, citizenship, race dummies (separate dummy for each race which are not mutually exclusive), age (including dummies for various ages such as 62, 65, and 70 or greater), weeks worked last year (with dummies for 40 and 50 or more), hours worked per week (with dummies for 40 and 60 or more), occupation categories. We also included a large set of interaction terms in our list of predictors including major income types (earnings, Social Security, spouse earnings), education, weeks and hours worked, race and age. In the imputation using the DER file, we include total W-2 wage and self-employment earnings, number of W-2 jobs, and spouse DER information to the list of predictors and interaction terms. In all, over 3,000 potential predictors and interaction terms can be included in our SRMI models.[6]

We chose to implement two stages of model selection regressions to prune the list of possible predictors to a more manageable one for each variable. In the first model-selection stage, we would like to reduce the number of variables that are candidates for the SRMI prediction models in the second stage. To do this, we limit the number of potential interactions by stepwise selection of all possible predictors on the sample of observed responses. This yields a smaller set of potential predictors. However, this set can still be very large. For example, in the model for wage earnings from primary job with the DER administrative data, there were 685 predictors selected. This pruned list of

[5] The 11 categories are 1) Management, business, and financial occupations (0010-0950), 2) Professional and related occupations (1000-3540), 3) Service occupations (3600-4650), 4) Sales and related occupations (4700-4960), 5) Office and administrative support occupations (5000-5930), 6) Farming, fishing, and forestry occupations (6000-6130), 7) Construction and extraction occupations (6200-6940), 8) Installation, maintenance, and repair occupations (7000-7620), 9) Production occupations (7700-8960), 10) Transportation and material moving occupations (9000-9750), and 11) Armed Forces (9840).

[6] In part, the large number of variables is due to the conversion of categorical variables into separate dummies. For example, there are seven marital statuses so the categorical marital status variable (A_MARITL) is converted into seven dummy variables, with each interacted with all the other possible interaction terms. This yields a large number of possible predictors from just the marital status variable.

model variables is used during each iteration of the SRMI (discussed below), where another stepwise model selection process is implemented.

As a first step in the imputation, we first transform all continuous variable to a normal distribution using the empirical normal transformation used in Woodcock and Benedetto (2009). We then create any interaction terms that are potential modeling variables. Next, we use stepwise model selection for each separate variable to be imputed to prune list of potential interaction term predictors as discussed above.

We then impute the missing values with SRMI. In each iteration of the SRMI, for each imputed variable, we stratify by race and gender. For each race-gender stratum, we select the list of predictors to include using stepwise selection on the pruned list. This is the second model-selection stage. We impute the missing values within each stratum using logistic and OLS regressions for binary variables and continuous respectively. The predictions are generated by taking the expected probability or value and sampling from the appropriate error distribution. For continuous variables with defined bounds, we ensure that the predicted values are within the acceptable bounds of the variable.[7]

An important part of the SRMI step is that prior to modelling and imputation of each variable, an Approximate Bayesian Bootstrap of the original sample is taken. This allows us to approximate the uncertainty in the model selection process and the uncertainty in the parameter values in the imputation model itself (the logistic or OLS regression in step.

We have created two multiple imputation data sets: 1) SRMI – *without* the use of administrative earnings data as predictors and 2) DER SRMI – *with* the use of administrative earnings data as predictors. In the second case, we are only using the administrative data to improve predictions about what the missing survey responses would have been. This allows us to analyze whether the responses are missing at random conditional on the survey responses only by testing how the addition of administrative data impacts the imputation diagnostics and results.

5. Results

In order to evaluate the amount of information our models add to the predictions, we also document the R^2 values for each regression. To give some examples of the improved fit that including the administrative earnings data makes possible, the pseudo-R^2 for whether an individual had earnings is 0.38 in the SRMI model and 0.57 in the DER SRMI, a difference of 0.19. For the value of wages from the longest job, the SRMI R^2 is 0.71 compared to 0.87 for the DER SRMI (0.16 difference). For nearly all recipiency and value models, the DER data improves the prediction.

We examine the extent to which wage earnings in each of the imputation methods, including the hot deck, matches the administrative earnings. In Figure 1, we show box plots of the imputed wage earnings for individuals with positive earnings in the DER by DER earnings decile. Not surprisingly, the DER SRMI seems to impute wage earnings closer to the DER ones than the SRMI or hot deck.

[7] For example, wage earnings must be between 0 and 1,099,000 in the CPS ASEC.

In order to evaluate the impact of 1) using SRMI imputation in place of the hot deck and 2) using administrative data in the SRMI separately, we replicated tables from the Census Bureau's annual Income, Poverty, and Health Insurance Coverage Report (DeNavas-Walt et al., 2011). We compare the median income estimates from the SRMI and DER SRMI to the hot deck in Table 2. In Table 3, we compare estimates of poverty between the hot decked sample and the two SRMI samples. For the hot decked sample, we calculate each statistic from the single implicate in the internal CPS ASEC file with replicate weights that were used for the calculation of the 2010 report.[8] For the SRMI estimates and estimates of differences between the SRMI and hot deck, we use replicates weights to calculate the standard errors for each SRMI implicate and combine the estimates to get the multiple imputation standard errors.

As seen in Table 2, the point estimate for household median income is lower in both the SRMI and DER SRMI than the hot deck, but statistically significantly for only the DER SRMI.[9] For nearly all subgroups, the DER SRMI has a statistically significantly lower median income. Median household income in the SRMI is lower in the SRMI for married couples, Blacks, and 25-34 year-olds. Although the standard errors are wider for both the SRMI and DER SRMI compared to the hot deck for nearly all groups, the differences are primarily due to within implicate variance. Although the standard errors for median income of all households are 75% greater in the SRMI and 51% greater in the DER SRMI respectively than the hot deck (not shown in Table 2) The imputation uncertainty increases the standard error by only 26% in the SRMI and 11% in the DER SRMI (not shown in Table 2).

As seen in Table 3, SRMI poverty estimates are higher than the hot deck for unrelated individuals (0.9%) and Blacks (0.8%). The results differ somewhat for the DER SRMI and the hot deck estimates of poverty. Most importantly, the overall poverty estimate is 0.4% higher in the DER SRMI than in the hot deck. With model-based imputation using administrative data, the estimated number of individuals in poverty is over 1.1 million greater than using the existing hot deck procedure. The DER SRMI also estimates statistically significantly more poverty for unrelated indivudals (0.9%), Whites (White alone, 0.3%), Hispanics (1.1%), males and females (0.4% for both), individuals aged 18-64 (0.4%), the foreign born (0.7%) and non-citizens (0.8%).

For poverty, the standard errors are 51% wider due to imputation uncertainty in the SRMI and 33% wider in the DER SRMI than the within imputation standard error estimate (not shown in Table 3). However, because the two SRMI models better predict income than the hot deck, the overall standard errors are 7% narrower in the SRMI and 16% narrower in the DER SRMI (not shown in Table 3). In other words, even with the added variance introduced by accounting for the imputation uncertainty, both SRMI models have more precise estimates of poverty than the hot deck.

We show a modified QQ plot to compare the final distribution of household income in the hot deck, SRMI, and DER SRMI in Figure 2. In this figure, we calculate the average household income at each percentile. We then plot the difference between each SRMI

[8] The weights used in this paper are balanced to 2000 Census controls and correspond to the one in the 2010 report.
[9] This table uses the Census' median income interpolation technique and is therefore comparable to the Table 1 in the 2011 Income and Poverty Report (De Navas-Walt et al., 2011).

impute and the hot deck at each percentile up to the 95th.[10] For example at the unweighted median, the SRMI estimate for median household income is nearly $300 lower than the hot deck and the DER SRMI estimate is over $800 lower.[11] This includes all imputed and observed income values in one implicate for each imputation technique. At every percentile below the 90th, the point estimates for the SRMI and DER SRMI are lower than the hot deck. Below the 80th percentile, household income is lower in the DER SRMI than the SRMI as well.

6. Conclusion

This paper implements an alternative model-based methodology, sequential regression multiple imputation, to impute missing income values in the 2011 CPS ASEC. The Census Bureau currently employs the hot deck procedure to impute missing income values. Unlike the hot deck procedure, sequential regression multiple imputation adds greater flexibility by accommodating additional covariates in the analysis and accounting for uncertainty in the imputation process. We implement a baseline model solely using data from the 2011 CPS ASEC and then add to this data W-2 earnings information from the Social Security Detailed Earnings Records (DER).

While this initial work compared median income and poverty, future work should consider other outcomes as well. Given the importance of measuring inequality, future work will produce common inequality measures such as the Gini coefficient and various percentile ratios. Since the CPS ASEC is often the workhorse data set among labor economists, future work will also provide estimates of the standard Mincer wage equation to gauge the impact on the return to education.

References

Abowd, John and Martha Stinson. 2013. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data." *Review of Economics and Statistics*, 95(5), pp 1451-1467.

Ambler G, Omar RZ, Royston P. 2007. "A comparison of imputation techniques for handling missing data predictor values in a risk model with a binary outcome." *Statistical Methods in Medical Research* 16:277–298.

Andridge, Rebecca, and Roderick Little. 2010. "A Review of Hot-Deck Imputation for Survey Nonresponse." *International Statistical Review*, 78(1), pp 40-64.

Bollinger, Christopher and Barry Hirsch. 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics*, 24(3), pp. 483-519.

van Buuren S. 2007. "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical Methods in Medical Research* 16: 219–242.

David, Martin, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest. 1986. "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association* 81: 29-41.

[10] Above the 95th percentile both the SRMI and DER SRMI greatly exceed the hot deck to such an extent that the differences below the 95th percentile are not visible given the change to the scale of the y-axis. For example, at the top percentile, each exceeds the hot deck by over $250,000.

[11] These comparisons are to illustrate how the figure is drawn, and we make no statements about the statistical significance of these differences.

DeNavas-Walt, Carmen, Bernadette D. Proctor, and Jessica C. Smith, U.S. Census Bureau, "Current Population Reports, P60-239, Income, Poverty, and Health Insurance Coverage in the United States: 2010", U.S. Government Printing Office, Washington, DC, 2011.

He, Yulei, Alan Zaslvsky, David Harrington, Paul Catalano, and Mary Beth Landrum. 2009. "Multiple imputation in a large-scale complex survey: a practical guide." *Statistical Methods in Medical Research* 19 (6): 653-670.

Hokayem, Charles, Christopher R. Bollinger, and James P. Ziliak. Forthcoming. "The Role of CPS Nonresponse in the Measurement of Poverty," *Journal of the American Statistical Association*.

Hokayem, Charles, Trivellore Raghunathan, and Jonathan Rothbaum. 2015. "Sequential Regression Multivariate Imputation in the Current Population Survey Annual Social and Economic Supplement." SEHSD Working Paper Number 2015-17.

Lillard, L., J.P. Smith, and F. Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy* 94 (3): 489-506.

Raghunathan, Trivellore, James Lepkowski, John Van Hoewyk, Peter Solenberger. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27 (1): 85-96.

Nicholas, Joyce and Michael Wiseman. 2009. "Elderly Poverty and Supplemental Security Income" *Social Security Bulletin*, 69(1), pp. 45-73.

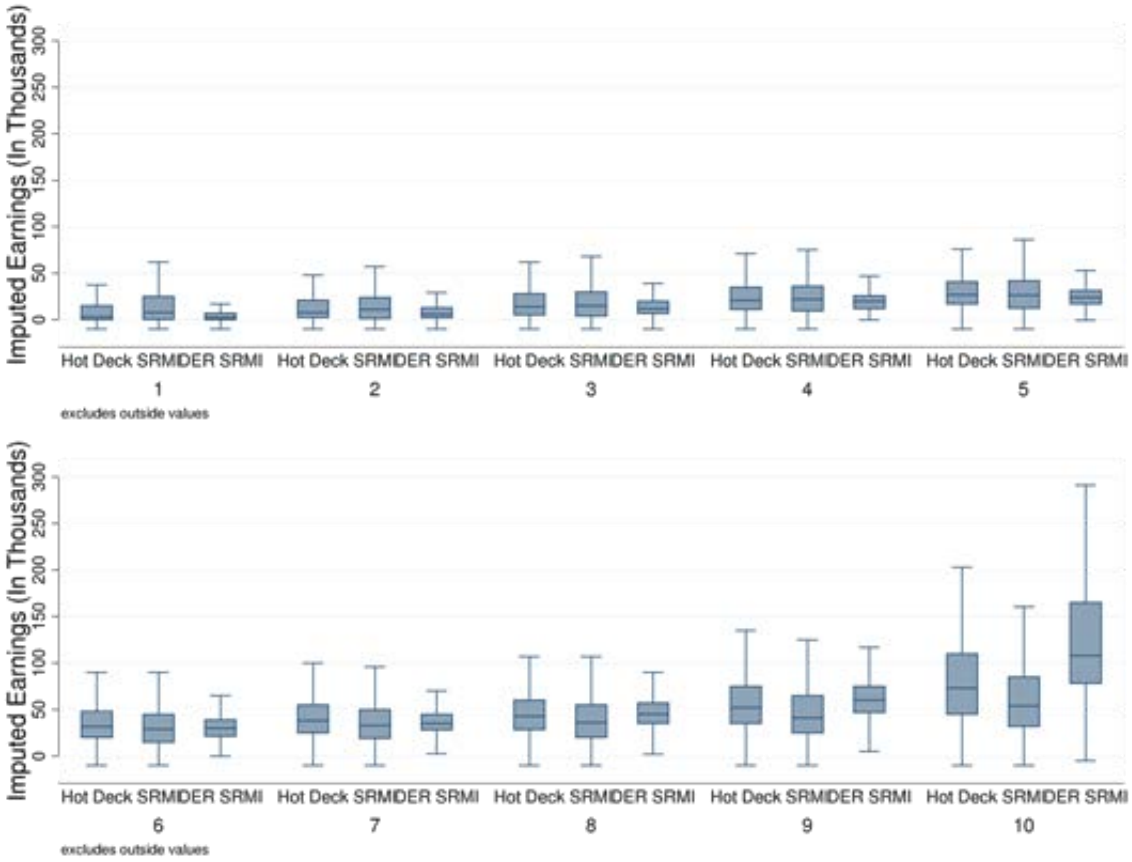
NORC at the University of Chicago. 2011. "Assessment of the US Census Bureau's Person Identification Validation System." Final Report presented to the US Census Bureau.

Roemer, Mark. 2002. "Using Administrative Earnings Records to Assess Wage Data Quality in the Current Population Survey and the Survey of Income and Program Participation." Longitudinal Employer-Household Dynamics Program Technical Paper No. TP-2002-22, US Census Bureau.

Welniak, Edward J. 1990. "Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty." US Census Bureau, Washington, DC, 1990.

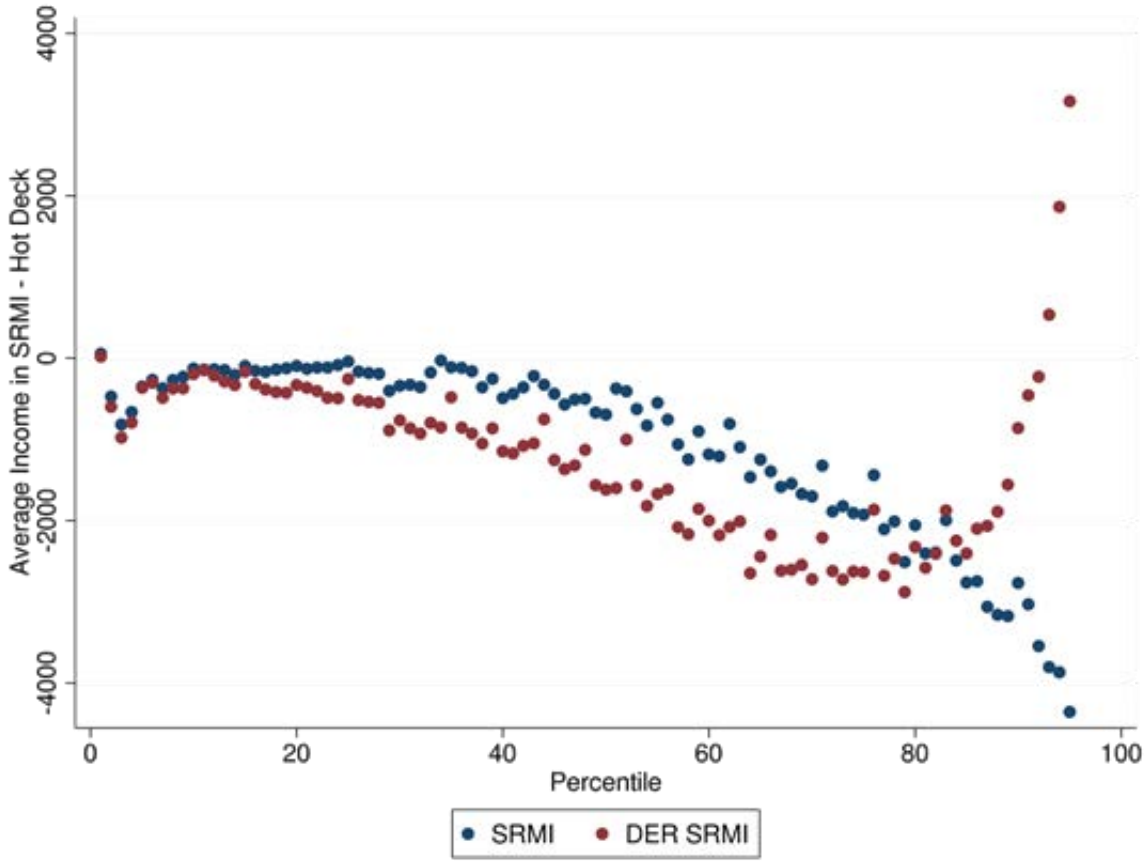
Woodcock, Simon D. and Gary Benedetto. 2009. "Distribution-preserving statistical disclosure limitation", *Computational Statistics & Data Analysis*, Volume 53, Issue 12.

Figure 1: Imputed Earnings by DER Earnings Decile



This figure shows box plots of imputed earnings for individuals with positive DER wage earnings by DER earnings decile in the 1) Hot Deck, 2) SRMI without administrative data (SRMI), and 3) SRMI with administrative data (DER SRMI) respectively.

Figure 2: Difference between Household Income in Hot Deck and SRMI Imputations by Percentile



This figure shows at each percentile the difference between average household income in the Hot Deck and 1) the SRMI without administrative data (SRMI) and 2) the SRMI with administrative data (DER SRMI) respectively.

Table 1: Non-Response Rates by Income Type

Income Type	Weighted Non-Response Rate		Share of Income Imputed
	Recipiency (Yes/No)	Value	
Wage and Self-Employment Earnings			
Primary Job	0.08%	12.71%	20.69%
Other wage earnings	0.03%	0.78%	15.21%
Other farm self-employment earnings	0.04%	0.28%	38.46%
Other non-farm self-employment earnings	0.03%	0.38%	16.93%
Unemployment Compensation	1.69%	0.08%	15.59%
Social Security	2.11%	4.38%	23.93%
Supplement Security Income	1.86%	0.38%	16.31%
Public Assistance	2.84%	0.12%	15.99%
Veterans' Benefits	2.38%	0.25%	22.81%
Survivors' Benefits	2.73%	0.25%	19.48%
Disability Benefits	0.57%	0.16%	24.36%
Retirement Income	3.17%	1.84%	24.28%
Interest Income	6.44%	16.50%	59.67%
Dividend Income	6.21%	6.54%	53.20%
Rental Income	4.77%	1.05%	18.90%
Education Assistance	3.07%	0.67%	21.05%
Child Support Income	3.22%	0.31%	16.28%
Alimony Income	3.18%	0.04%	21.47%
Financial Assistance	3.34%	0.26%	28.73%
Other Income		0.10%	8.18%
Supplement Non-Response			
All Income Recipiency/Value Information Missing		12.94%	12.87%
Any Income Type Missing	22.74%	44.19%	34.69%

This table show the imputation rate in the 2011 CPS ASEC by income type using individual weights for individuals age 15 and older. In the first column, we show the non-response rate for income recipiency (for example, did you receive Social Security Income?). In the second column, we show non-response rates for income values (for example, how much did you receive in Social Security income?). The third column, shows the share of total income that is imputed for each income type. For Supplement non-response and any income type missing, the share is imputed income as a share of total income.

Table 2: Median Income by Selected Characteristics: 2011 Hot Deck, SRMI, and DER SRMI

Characteristic	Hot Deck		SRMI		DER SRMI		Percentage Difference (HD-SRMI)/ SRMI		Percentage Difference (HD-DER SRMI)/ DER SRMI	
	Number (Thousands)	Estimate	Number (Thousands)	Estimate	Number (Thousands)	Estimate	Estimate		Estimate	
All Households	119,927	49,276	118,682	48,740	118,682	48,059	1.10	*	2.53	
Family households	79,539	61,395	78,613	61,153	78,613	60,452	0.40	*	1.56	
Married-couple families	58,656	72,495	58,036	71,449	58,036	70,783	* 1.47	*	2.42	
Female householder, no husband present	15,235	31,970	15,019	32,669	15,019	32,151	-2.13		-0.56	
Male householder, no wife present	5,648	49,813	5,559	48,503	5,559	47,526	2.71	*	4.82	
Nonfamily households	40,388	29,578	40,069	29,331	40,069	29,023	0.85		1.92	
Female householder	21,420	25,365	21,234	25,256	21,234	25,037	0.43		1.31	
Male householder	18,968	35,486	18,835	34,664	18,835	34,185	2.39	*	3.82	
Race and Hispanic Origin										
White	96,306	51,709	96,144	51,330	96,144	50,742	0.74	*	1.91	
White, not Hispanic	83,314	54,460	83,471	53,790	83,471	52,864	1.25	*	3.02	
Black	15,265	32,124	15,065	31,419	15,065	31,431	* 2.24		2.21	
Asian	5,212	64,259	4,747	62,566	4,747	61,319	2.72	*	4.80	
Hispanic (any race)	14,435	37,631	13,665	37,218	13,665	36,807	1.11	*	2.24	
Age										
Under 65 years	94,190	55,112	93,320	54,228	93,320	53,403	1.64	*	3.20	
15 to 24 years	6,231	28,224	6,140	28,132	6,140	27,937	0.34		1.04	
25 to 34 years	19,487	49,877	19,572	47,915	19,572	47,693	* 4.10	*	4.58	
35 to 44 years	21,458	61,418	21,250	60,726	21,250	60,204	1.14		2.02	
45 to 54 years	24,767	62,341	24,530	62,282	24,530	61,538	0.10		1.31	
55 to 64 years	22,246	56,474	21,828	55,722	21,828	54,819	1.36	*	3.02	
65 years and older	25,737	31,461	25,362	31,297	25,362	31,101	0.53		1.16	
Nativity										
Native born	103,232	50,154	102,647	49,573	102,647	49,020	1.18	*	2.32	
Foreign born	16,695	43,967	16,036	43,698	16,036	42,259	0.64	*	4.04	
Naturalized citizen	8,568	52,945	8,277	51,995	8,277	51,472	1.84	*	2.87	
Not a citizen	8,127	36,413	7,758	36,692	7,758	35,674	-0.76	*	2.07	

This table shows the imputation results without administrative data (SRMI) and with administrative data (DER SRMI) in the model. Income in 2010 dollars. Households and people as of March of the following year. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see ftp://ftp2.census.gov/programs-surveys/cps/techdocs/cpsmar11.pdf. Standard errors calculated using replicate weights. Multiple imputation formulas to account for imputation uncertainty used for SRMI and DER SRMI only.

*Significant different from zero at the 90 percent confidence level. Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement.

Table 3: Poverty by Selected Characteristics: Comparing Hot Deck, SRMI and DER SRMI Imputation

Characteristic	Hot Deck		SRMI		DER SRMI		Difference (HD-SRMI)/ SRMI		Difference (HD-DER SRMI)/ DER SRMI	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent
PEOPLE										
Total	46,343	15.1	46,687	15.3	47,481	15.5	343	0.1	* 1,138	* 0.4
Family Status										
In families	33,120	13.2	32,961	13.2	33,758	13.5	-159	-0.1	638	0.3
Householder	9,400	11.8	9,401	11.8	9,623	12.1	1	0.0	223	0.3
Related children under 18	15,598	21.5	15,558	21.4	15,790	21.8	-39	0.0	192	0.3
Related children under 6	6,037	25.3	6,013	25.2	6,139	25.7	-24	-0.1	102	0.4
In unrelated subfamilies	774	46.1	756	45.0	773	46.0	-17	-1.0	0	-0.1
Reference person	283	43.2	275	42.1	281	43.0	-8	-1.2	-2	-0.2
Children under 18	469	50.3	458	49.1	467	50.0	-11	-1.1	-2	-0.3
Unrelated individual	12,449	23.0	12,969	23.9	12,949	23.9	* 520	* 0.9	* 500	* 0.9
Race and Hispanic Origin										
White alone	31,083	13.0	31,063	12.9	31,850	13.3	-20	0.0	* 767	* 0.3
White alone, not Hispanic	19,251	9.9	18,973	9.7	19,549	10.0	-278	-0.1	298	0.2
Black alone	10,746	27.4	11,045	28.1	10,929	27.8	* 299	* 0.8	183	0.5
Asian alone	1,899	12.2	1,897	12.2	1,948	12.5	-2	0.0	49	0.3
Hispanic (of any race)	13,522	26.5	13,897	27.3	14,083	27.6	375	0.7	* 561	* 1.1
Sex										
Male	20,893	14.0	21,122	14.1	21,485	14.3	230	0.2	* 593	* 0.4
Female	25,451	16.3	25,564	16.3	25,996	16.6	114	0.1	* 545	* 0.4
Age										
Under 18 years	16,286	22.1	16,227	22.0	16,472	22.3	-59	-0.1	186	0.3
18 to 64 years	26,499	13.8	26,848	14.0	27,327	14.2	349	0.2	* 828	* 0.4
65 years and over	3,558	9.0	3,611	9.1	3,681	9.3	53	0.1	123	0.3
Nativity										
Native	38,485	14.4	38,746	14.5	39,361	14.8	261	0.1	876	0.3
Foreign born	7,858	19.9	7,941	20.2	8,120	20.6	83	0.2	* 262	* 0.7
Naturalized citizen	1,954	11.3	1,986	11.5	2,032	11.7	32	0.2	78	0.4
Not a citizen	5,904	26.8	5,955	27.0	6,088	27.6	51	0.2	* 184	* 0.8

This table shows the imputation results without administrative (SRMI) and with administrative data (DER SRMI) in the model. Income in 2010 dollars. Households and people as of March of the following year. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see ftp://ftp2.census.gov/programs-surveys/cps/techdocs/cpsmar11.pdf. Standard errors calculated using replicate weights. Multiple imputation formulas to account for imputation uncertainty used for SRMI and DER SRMI only.

*Significant different from zero at the 90 percent confidence level. Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement.