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

Charles Hokayem, Trivellore Raghunathan, and Jonathan Rothbaum*

September 3, 2015

Abstract

Abstract: The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) serves as the data source for income, inequality, and official 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 through multiple imputation. 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.

* Hokayem: Visiting Assistant Professor of Economics, Centre College. Email: charles.hokayem@centre.edu. Raghunathan: Professor, Department of Biostatistics, Michigan University School of Public Health. Email: teraghu@umich.edu. Rothbaum: Economist, Social and Economic Housing Statistics Division, U.S. Census Bureau. Email: jonathan.l.rothbaum@census.gov.

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 Background of the 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. 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.

Table 1 provides an example of this sequential procedure for the allocation group that is only missing earnings from longest job. The table shows the set of match variables and the number of categories at each level. The table also shows the number of categories used to define each match variable at each level. An empty cell indicates the variable is no longer in the set of match variables. For example, the first column shows 16 variables are initially used to match the nonrespondent observation to a donor. If a match is not found, the second level drops presence of children and labor force status of spouse while collapsing the number of categories for race (3 categories to 2 categories), age (9 categories to 6 categories), years of school completed (6

[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.

categories to 5 categories), and type of residence (3 categories to 2 categories). This process of dropping and collapsing match variables continues until the only match variables remaining are sex, years of school completed, weeks worked, and class of worker. This sequential match procedure always ensures a match. The last row of the table gives the number of cells created by the match variables at each level. The 16 match variables used in the first level produces over 620 billion cells while the four match variables in the last attempt produces 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. This imputation procedure uses eight allocation groups. Moreover, the set of match variables is smaller, consisting solely of variables from the basic monthly CPS. 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 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 hot deck procedure does implicitly assume an underlying regression model of the missing income variable on interactions of the match variables. 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 last row of Table 1 illustrates how sparse the cells can be.

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 $\left(Y_2, Y_3, \ldots, Y_p\right)$. 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 earnings are also not capped by Census as are ASEC earnings and income. 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). Workers in the DER file are uniquely identified by a Protected Identification Key (PIK) assigned by the Census Bureau. The PIK is a confidentiality-protected version of the Social Security Number.

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 Security Number. The 2011 ASEC-DER match rate is 89.7 percent. A worker can appear multiple times in the DER file if they have several jobs. 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 also count the total number of jobs held. 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 all missing income for individuals aged 15 and older (156,849 individuals). We impute income for 20 categories: wage and salary earnings, self-employment earnings (farm and nonfarm), 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.

[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.

As discussed in Section 2, there are two reasons that income information could be missing in the CPS ASEC, item non-response and supplement non-response.[4] In Table 2, 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.

4.1 Challenges

There are a number of challenges to implementing SRMI in the CPS ASEC. First, many income types do not follow a normal distribution or any simple transformation of a normal distribution. Second, we impute a large number of variables that are related to income, including the aforementioned recipiency and values for each income type, as well as occupation groups and time worked (weeks in the last year and hours per week). Third, we must select predictors for the modelling of each income variable from a very large set of possible covariates in the CPS ASEC. In this section, we discuss how we address each of these issues.

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. As an example, Figure 1 shows the histogram, kernel density estimation, and normal approximation for interest income using three different transformations, Panel A: log transformation, Panel B: Tukey-gh transformation (He and Raghunathan, 2006), and Panel C: an empirical normal transformation (Woodcock and Benedetto, 2009). Both the log and empirical normal transformation result in approximately normally distributed interest income. However, some income types, such as self-

[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.

employment income, allow negative values. As a result, we use the empirical normal transformation instead of the log transformation as it both ensures normality in all cases and is not affected by the presence of negative values.

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 12.94% 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 categories.[5] We separated these 11 occupation groups into a series of binary categories connected by the tree structure shown in Figure 2. 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

[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).

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

4.2 SRMI Steps

In this section, we will discuss the steps of the SRMI process used to impute missing income in the CPS ASEC.

1. **Normal transformation** – Transform all non-categorical continuous variables to normal distribution with the empirical normal transformation used in Woodcock and Benedetto (2009).
2. **Create all interaction terms** – Create the interaction terms with the transformed variables.
3. **First model-selection stage** – Stepwise model selection for each separate variable to be imputed to prune list of potential interaction term predictors as discussed above.
4. **Transform to original scale** – Return all transformed variables to their original scale
5. **SRMI** – With each iteration of the SRMI do the following steps:

[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 the single marital status variable.

a. **Normal transformation** – same as above. This transformation will change with each iteration as the distribution of incomes changes after imputation.
b. **Calculate any derived variables that are used as predictors –** These derived variables include individual dummy variables, household, and spouse variables. This also updates the derived variables to reflect the imputed values from the previous iteration.
c. **Create all interaction terms** – Create the interaction terms with the transformed variables. This also updates the interaction terms to reflect the imputed values from the previous iteration.
d. **Impute variables sequentially** – For each variable to be imputed (such as wage earnings), do the following:
    i. Select a random sample by Approximate Bayesian Bootstrap.
    ii. Stratify the sample by race (non-Hispanic White, non-Hispanic Black, and Hispanic) and gender.
    iii. For each stratum, select the list of predictors to include using stepwise selection on the pruned list. This is the second model-selection stage.
    iv. Impute the missing values within each stratum using logistic or 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]
e. **Transform to original scale** – return all variables to their original scales.
f. **Repeat to create five implicates** – Each implicate was created with five iterations.

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 (step 5.d.i.). This allows us to approximate the uncertainty in the model selection process (step 5.d.iii) and the

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

uncertainty in the parameter values in the imputation model itself (the logistic or OLS regression in step 5.d.iv).

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.

4.3 Diagnostics

In order to evaluate the SRMI process, we create a number of diagnostic tables and figures. First, to evaluate the variable pruning (Step 3), we plot a histogram of the residuals of the selected model along with the estimated kernel density and normal distributions. This allows us to evaluate whether there are large deviations from the assumption that the errors are conditionally normally distributed. In Figure 3, this is shown for wage from longest job and Social Security income, which are the two largest sources of aggregate income. In both cases, the assumption of conditional normality appears to hold reasonably well. This holds for each of the imputed income variable diagnostic (not shown). For each predictor in the pruning regression, we also plot the residuals against each right hand side variable. For continuous variables, we generate a scatter plot, and for binary or categorical variables, we generate a box plot. These plots help evaluate model misspecification. Examples of these diagnostic plots are shown in Figure 4 for wage earnings from longest job.[8] In both Panel A (wage residuals plotted against age) and B (wage residuals plotted against education level), the conditional normality assumption and the conditional expectation of the residual are reasonable.

In order to evaluate the amount of information our models add to the predictions, we also document the R^2 values for each regression. In Table 3, we show these values for the SRMI and DER SRMI models, as well as the percentage increase in the R^2 from adding the DER data to the model. For example, the pseudo-R^2 for whether an individual had earnings is 0.38 in the SRMI

[8] To avoid disclosure of responses, Figure 3 and Figure 4 show results from the public-use data. However, the results are nearly identical for the internal CPS ASEC with and without the administrative data.

model and 0.57 in the DER SRMI, a difference of 51%. For the value of wages from the longest job, the SRMI R^2 is 0.71 compared to 0.87 for the DER SRMI (22% difference). For nearly all recipiency and value models, the DER data improves the prediction.

We also test whether the imputations are reasonable under the missing at random (MAR) assumption by implementing a propensity score diagnostic proposed by Raghunathan and Bondarenko (2007). In this diagnostic for each implicate, we first regress response (R=1) on the variables selected in the first-stage selection model to predict response propensity based on observed characteristics. We estimate the response propensity for each individual as the average predicted propensity across the five implicates. We then regress the modeled variable (for example normalized wage earnings) on the predicted response propensity. If the imputation are reasonable under the MAR assumption, the distribution of residuals from this regression should follow the same distribution for the respondents and non-respondents. In Figure 5, we apply this diagnostic to Wage of Longest Job for three groups: 1) respondents, 2) item non-respondents, and 3) supplement non-respondents. In Panels A and B, we plot the kernel density of the distribution of transformed wages for the SRMI and DER SRMI imputation. In Panels C and D, we plot the diagnostic showing the distribution of residuals of the regression of wages on the predicted response propensity described above.

Finally, we examine the extent to which wage earnings in each of the imputation methods, including the hot deck, matches the administrative earnings. In Figure 6, 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.

5 Results

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 4. In Table 5 and Table 6, 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 was used for the

calculation of the 2010 report.[9] 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.

We show a modified QQ plot to compare the final distribution of household income in the hot deck, SRMI, and DER SRMI in Figure 7. For this figure, we calculate the average household income at each percentile. We then plot the difference between each SRMI 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 less than the hot deck and the DER SRMI estimate is over $800 less.[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.

In Table 4, we show how the SRMI and DER SRMI affect estimates for median household income. 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). 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. 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, 25-34 year-olds, and those without a disability.

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, the imputation uncertainty increases the standard error by only 26% in the SRMI and 11% in the DER SRMI.

---

[9] The weights used in this paper are balanced to 2000 Census controls and correspond to the one in the 2010 report. The differences between the SRMI model and the hot deck are not statistically significant whereas the differences between the DER and hot deck are, see Table 4.

[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.

Table 5 shows poverty estimates and comparisons between the hot deck and SRMI file. The SRMI estimates are lower only for individuals that did not work (1.9%). SRMI poverty estimates are higher than the hot deck for unrelated individuals (0.9%), Blacks (0.8%), workers between 18-64 (0.9%) and full-time year-round workers (1.5%).

The results differ somewhat for the DER SRMI and the hot deck estimates of poverty, shown in Table 6. 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%), non-citizens (0.8%), the Northeast census region (0.5%), urban areas (inside MSAs, 0.4% and inside principal cities, 0.6%), and nearly all worker and disability types.

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. 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. 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.

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.

Future work will produce SRMI estimates of male-female earnings differentials along with differentials by race. 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.

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.

Groves, Robert. 2001. Survey Nonresponse. New York: Wiley.

He, Yulei, and Trivellore E. Raghunathan. "Tukey's gh distribution for multiple imputation." *The American Statistician* Volume 60, Issue 3 (2006).

He, Yulei, et al. 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*.

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, et al. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27 (1): 85-96.

Raghunathan, Trivellore and Irina Bondarenko. 2007. "Diagnostics for Multiple Imputation." *SSRN Working Paper Series*.

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

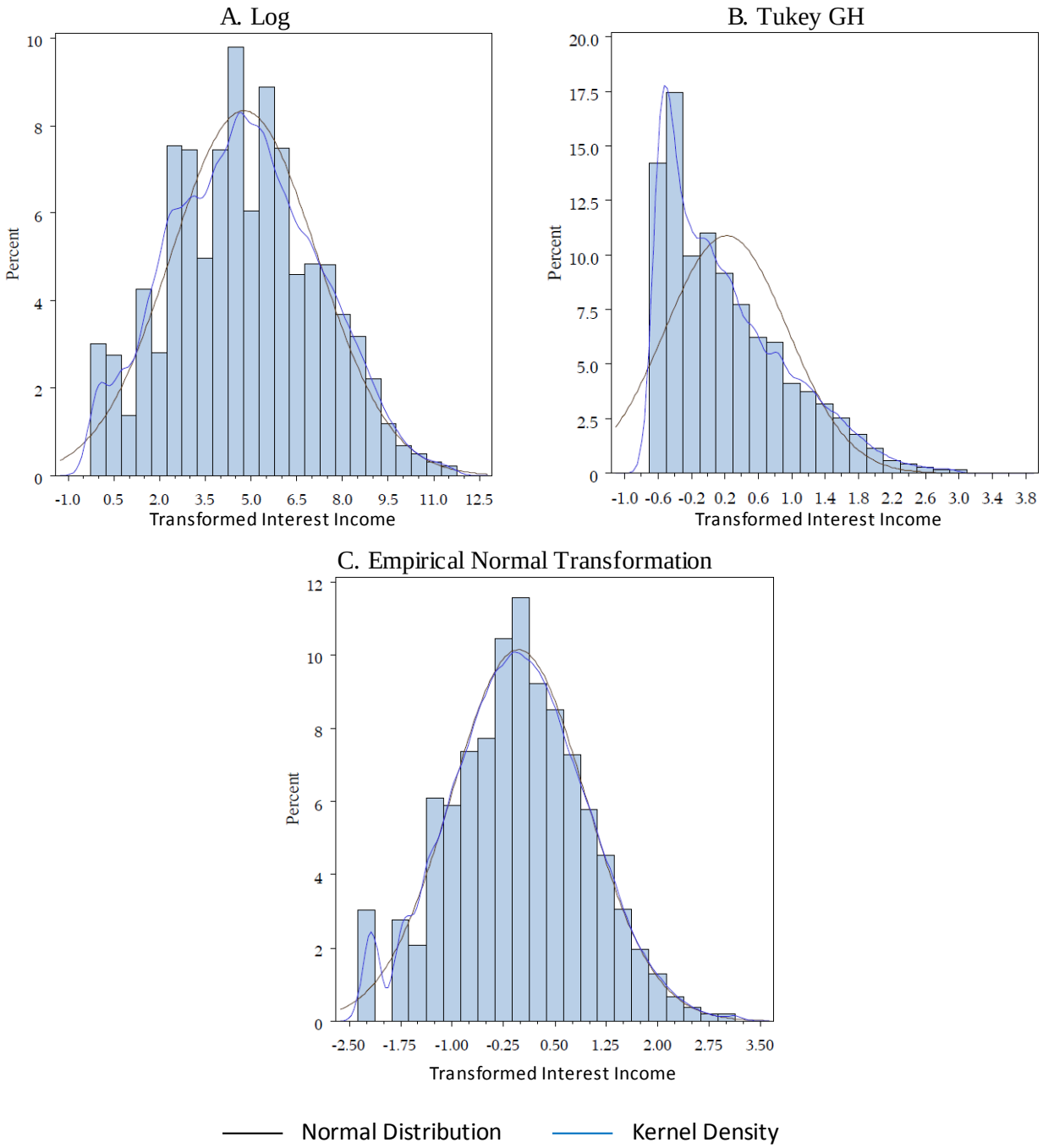
NORC at the University of Chicago. 2011. "Assessment of the US Census Bureau's Person Identification Validation System." Final Report presented to the US Census Bureau.

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, "Distribution-preserving statistical disclosure limitation", *Computational Statistics & Data Analysis*, Volume 53, Issue 12, (2009).

Figure 1: Comparison of Transformations of Interest Income (Public Use Data)



This figure shows histograms with 25 bins for transformed interest income. Panel A shows the log transformation, Panel B shows the Tukey GH approximation, and Panel C shows the empirical normal transformation used in this paper. In each Panel, the kernel density estimation as well as the estimated normal distribution of the transformed income is also shown. For disclosure reasons, all data used in this figure are from public-use CPS ASEC data.

Figure 2: Imputation of Occupation Groups as a Series of Binary Variables

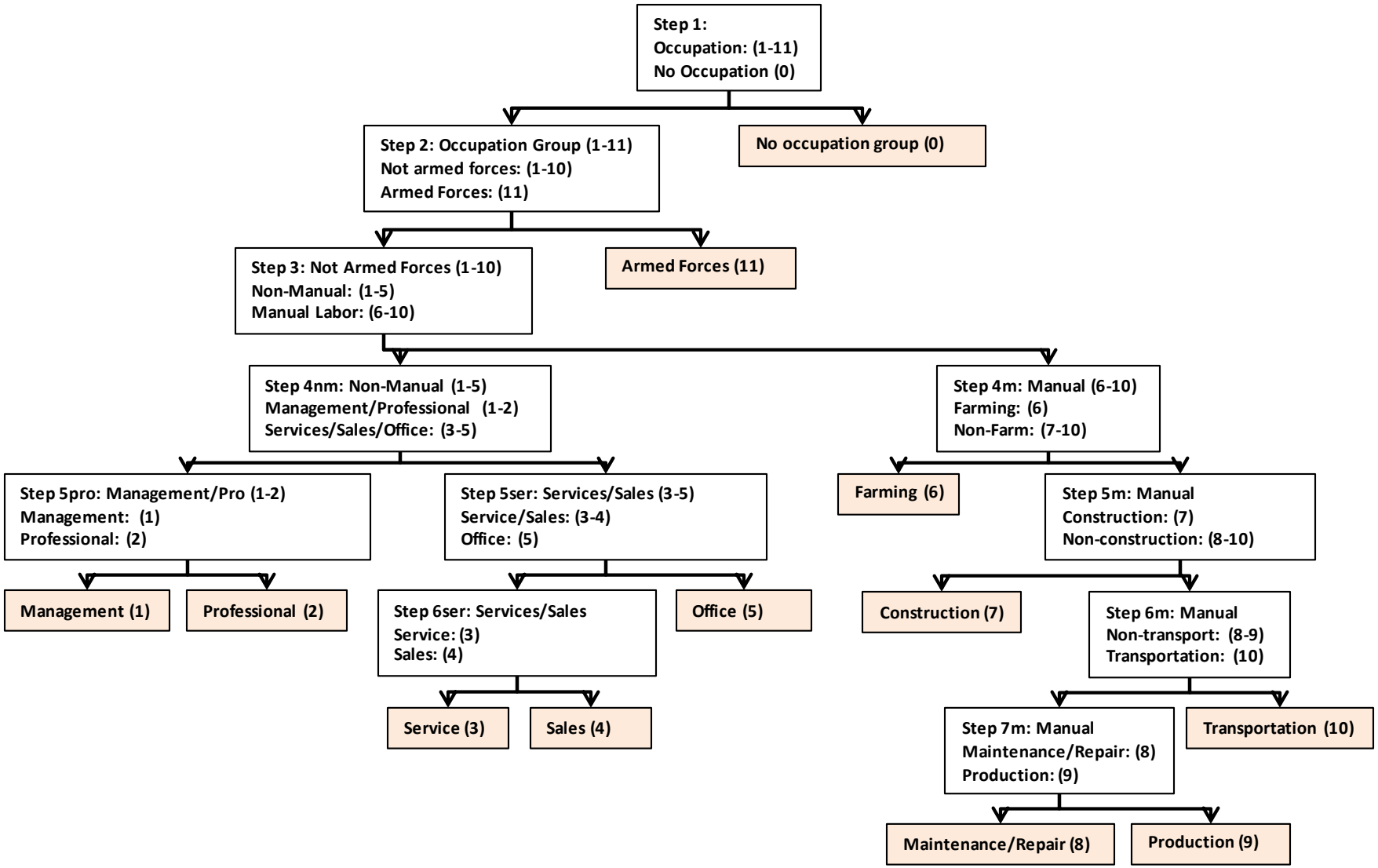
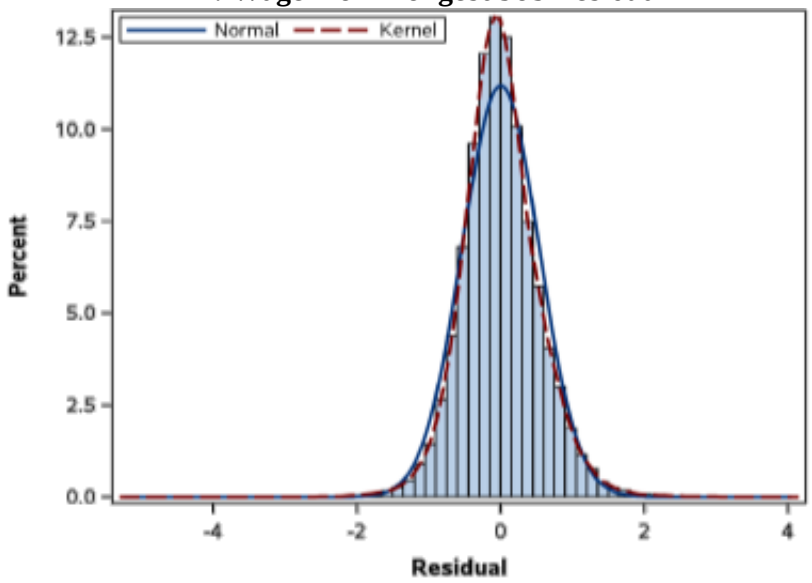
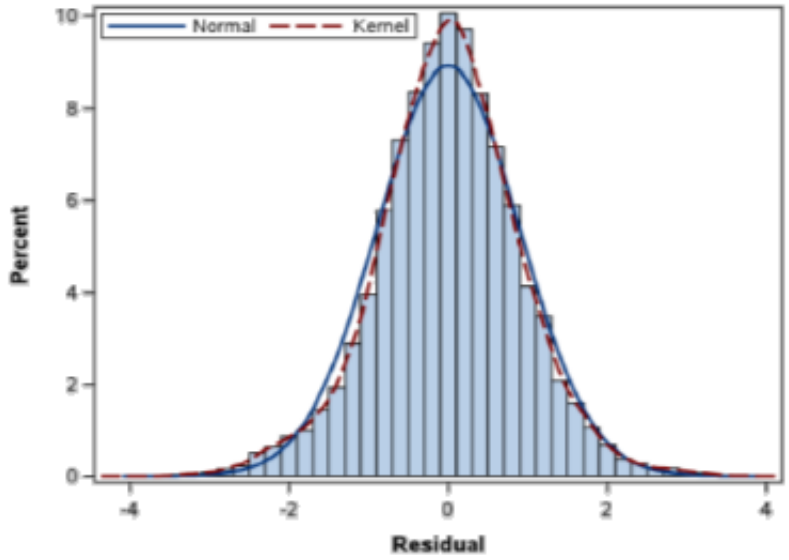


Figure 3: Model Selection Diagnostics: Distribution of Residuals

A. Wage from Longest Job Residual

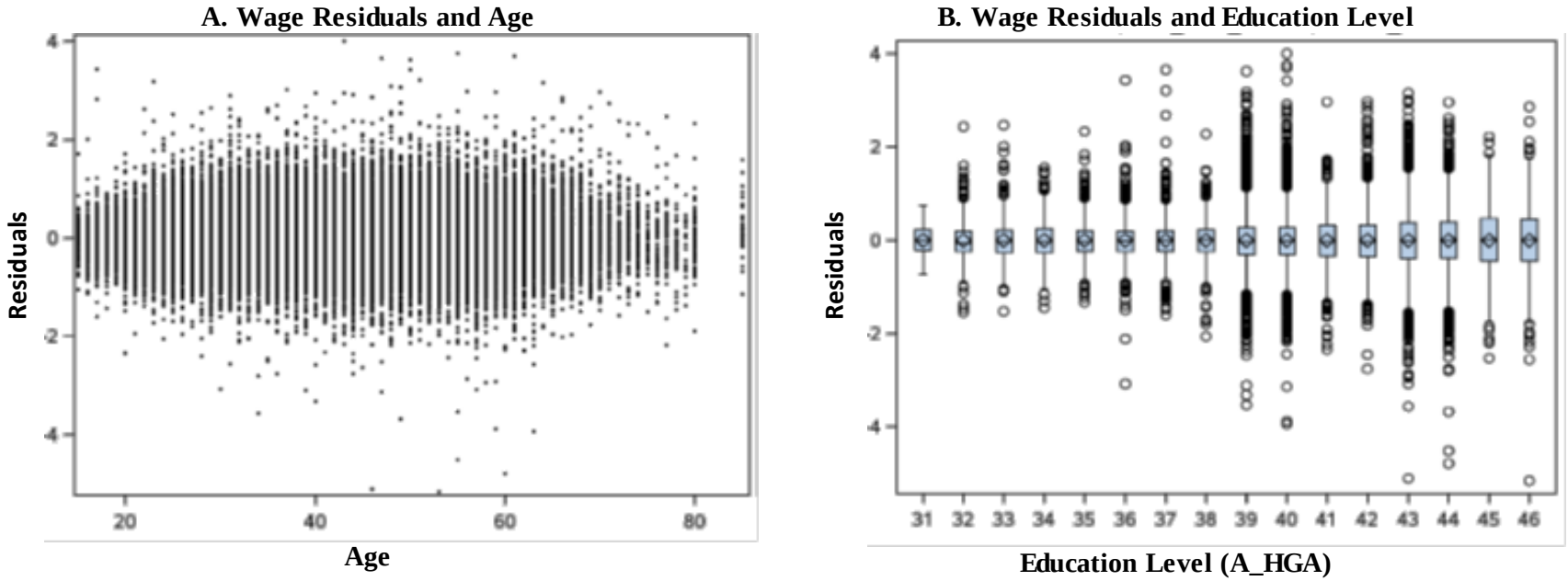


B. Social Security Income Residual



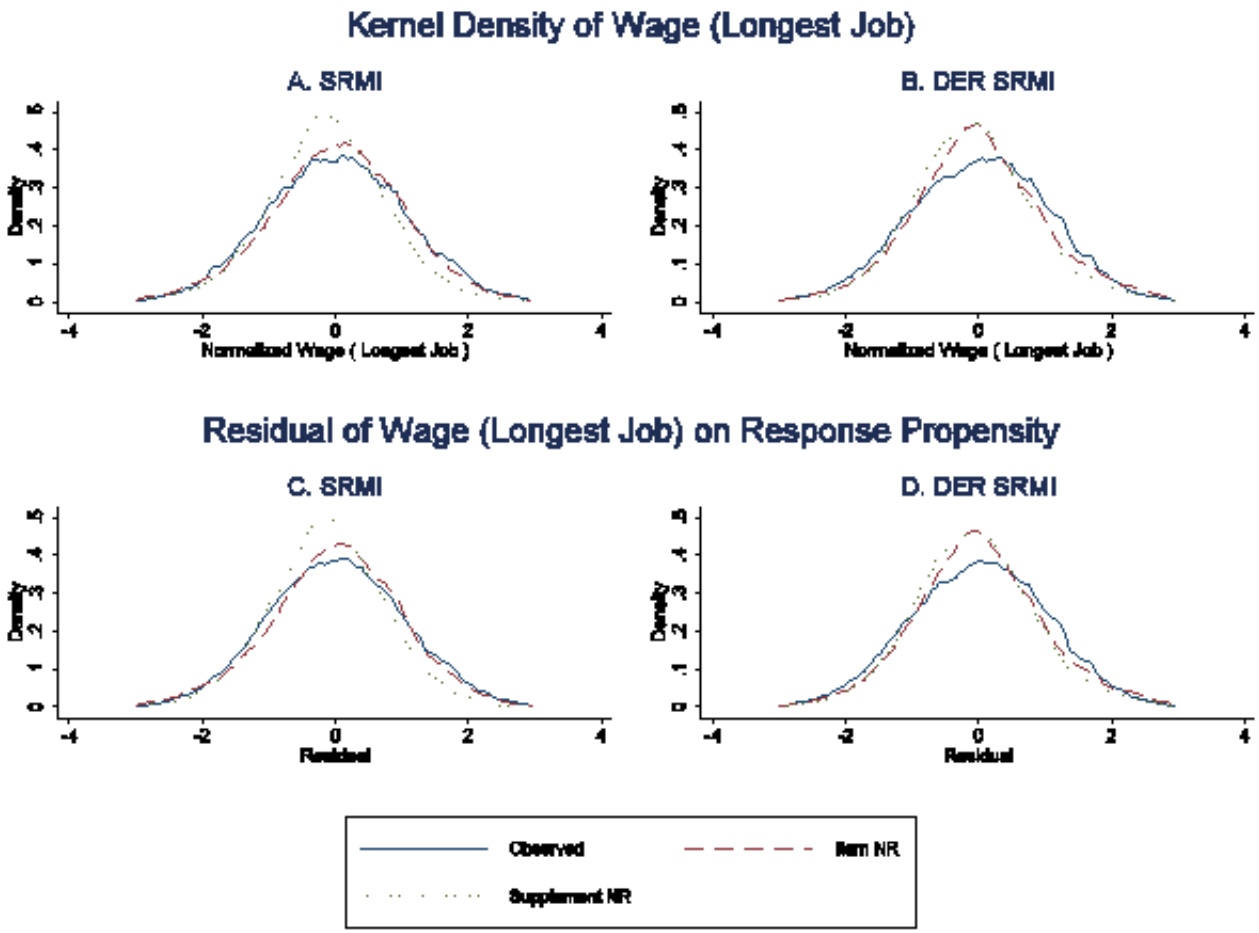
This figure shows histograms for the residuals of the model pruning regressions (discussed in Section 3). Each panel shows the residual term for from the regression of transformed income on the uninteracted predictors selected interactions terms. Panel A shows the residual for wage earnings from longest job. Panel B shows the residual for Social Security income. In both cases, the assumption of conditional normality appears to hold reasonably well. For disclosure reasons, all data used in this figure are from public-use CPS ASEC data.

Figure 4: Model Selection Diagnostics: Wage Earnings from Longest Job Residuals Conditional on Select Predictors



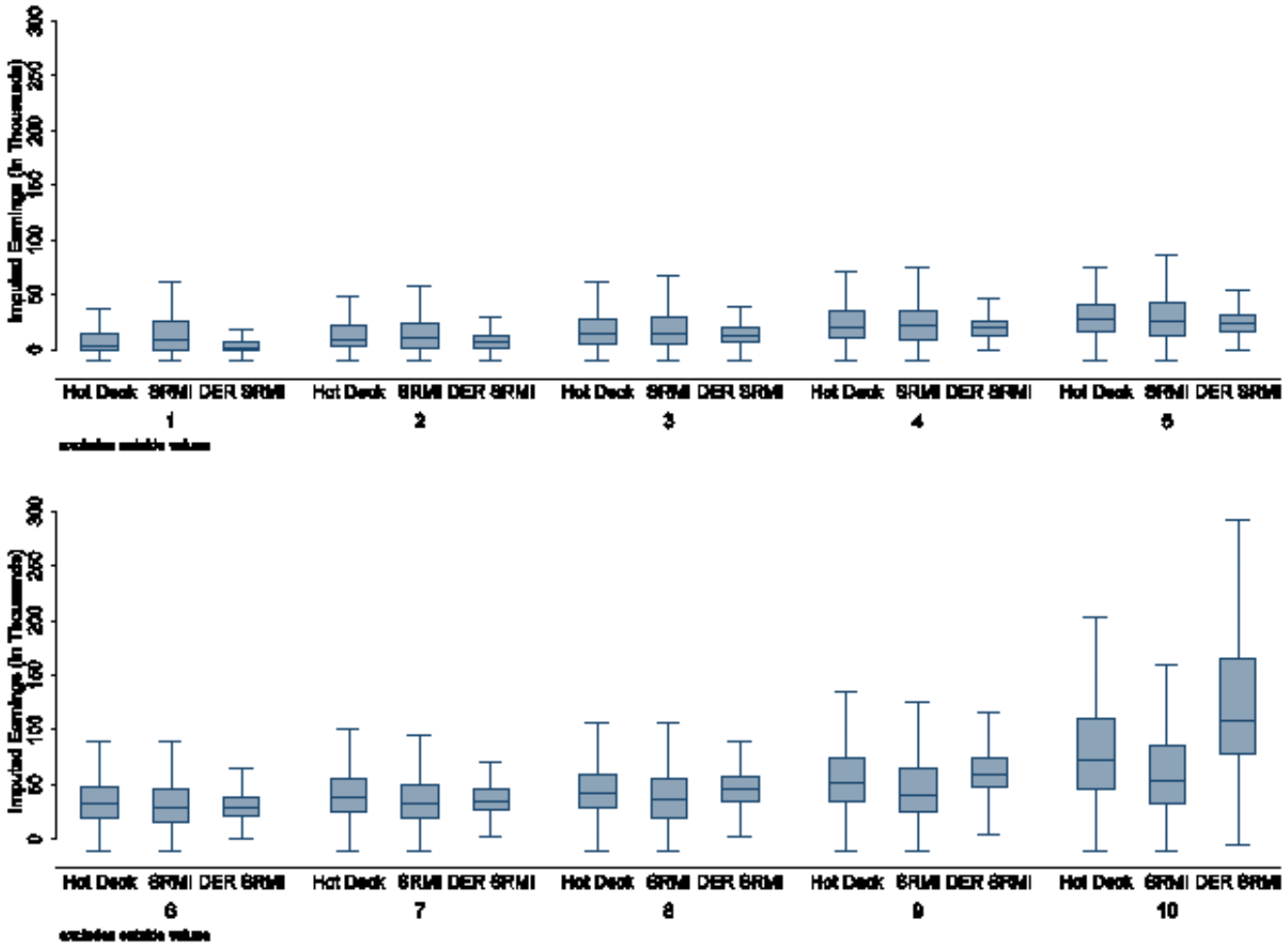
This figure shows the scatter plot of the residuals of the model pruning regressions (discussed in Section 3) against select predictors. These diagnostic plots help evaluate the appropriateness of the model. For example, if the expectation of the residual conditional on a given predictor deviates significantly from zero or the plot residuals appears very non-normally distribution, it is evidence of model misspecification. Panel A shows the residuals from the regression of wage earnings from the longest job plotted agains age. Panel B shows the box plot with 25th and 75th percentile of the residuals and outliers plotted against the 16 education level categories in the CPS ASEC. For disclosure reasons, all data used in this figure are from public-use CPS ASEC data.

Figure 5: Response Propensity Diagnostics for Wage of Longest Job



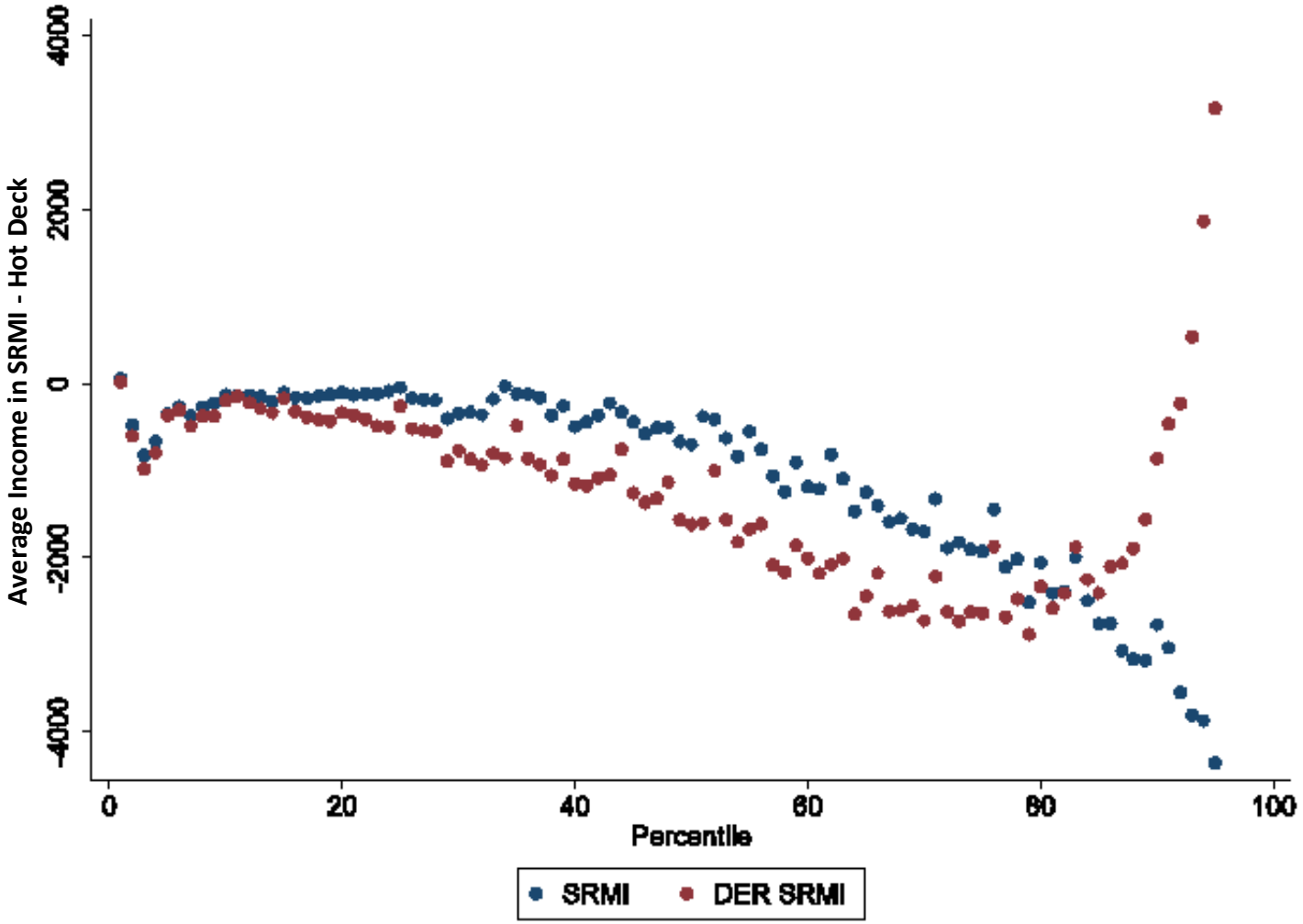
This figure shows the response propensity diagnostics from Raghunathan and Bondarenko (2007). Panels A and B show the kernel density plots of the wage income for three groups 1) respondents (observed), 2) item nonrespondents, and 3) supplement nonrespondents in the SRMI and DER SRMI after the empirical normal transformation. Panels C and D plot the kernel density of the residuals of the transformed w ages regressed on the predicted response propensities for each group based on the selected model for wage earnings. For disclosure reasons, only values between -3 and 3 shown.

Figure 6: Imputed Earnings by DER Earnings Decile



This figure shows box plots of imputed earnings for individuals with positive DER wage earnings by decile in the Hot Deck, SRMI and DER SRMI samples.

Figure 7: 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 and 2) the DER SRMI respectively.

Table 1: CPS Hot Deck Imputation Cell Counts for Missing Earnings from Longest Job

	Match Level					
Match Variable	**1**	**2**	**3**	**4**	**5**	**6**
Sex	2	2	2	2	2	2
Race	3	2	2			
Age	9	6	3	3		
Relationship	7	7	4	4	4	
Years of School Completed	6	5	5	4	4	4
Marital Status	4	4				
Presence of Children	3					
Labor Force Status of Spouse	3					
Weeks Worked	5	5	4	4	4	4
Hours Worked	3	3	3	3	2	
Occupation	528	528	66	66	66	
Class of Worker	5	5	5	3	3	3
Other Earnings	8	8				
Type of Residence	3	2	2			
Region	4	4				
Transfers payments receipt	2	2	2	2		
Number of Donor-Recipient Cells	**620,786,073,600**	**17,031,168,000**	**3,801,600**	**456,192**	**50,688**	**96**

This table shows the calculation of the number of possible cells at each match level in the hot deck for earnings from longest job. In the first match level, there are two categories for gender, three for race, 9 for age, etc. Multiplying the number of categories for each variable yields $2 * 3 * 9 * ... = 621$ billion cells. For a donor (with earnings value information) to match a recipient (missing earnings value), they must be in the same category for all variables at that match level. If no donor exists for a given recipient, the hot deck moves to the next match level to find a donor.

Table 2: 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 3: Model Diagnostics – R^2 of Prediction Model on Observed Responses

	Recipiency (Y/N)			Value		
Variable	**SRMI**	**DER SRMI**	**% Difference (DER SRMI-SRMI)/SRMI**	**SRMI**	**DER SRMI**	**% Difference (DER SRMI-SRMI)/SRMI**
Earnings	0.38	0.57	51.4			
Wages vs. Self-Employment	0.09	0.21	138.3			
Farm vs. Non-farm Self-Employment	0.45	0.51	11.6			
Weeks Worked				0.47	0.49	2.7
Hours Worked				0.44	0.45	2.2
Wages				0.71	0.87	21.6
Non-farm Self-Employment				0.55	0.70	25.5
Farm Self-Employment				0.85	0.47	-45.0
Other Wages	0.15	0.30	96.1	0.33	0.47	41.7
Other Self-Employment	0.16	0.17	8.6	0.62	0.66	7.2
Other Farm Self-Employment	0.19	0.20	4.0	0.46	0.46	-0.1
Unemployment Compensation	0.21	0.22	4.9	0.52	0.55	4.8
Workers' Compensation	0.05	0.05	12.0	0.85	0.86	1.3
Social Security	0.68	0.69	0.6	0.24	0.26	9.5
SSI	0.17	0.18	6.5	0.55	0.61	9.5
Public Assistance	0.06	0.06	8.6	0.60	0.62	4.0
Veterans' Benefits	0.05	0.05	5.5	0.51	0.61	21.1
Survivors' Benefits	0.12	0.12	1.7	0.54	0.58	6.9
Disability Benefits	0.04	0.05	14.8	0.88	0.92	5.2
Retirement Income	0.30	0.31	1.8	0.39	0.41	5.6
Interest	0.31	0.31	1.6	0.23	0.24	4.7
Dividends	0.32	0.32	1.5	0.21	0.23	9.7
Rental Income	0.09	0.10	6.4	0.26	0.31	22.4
Education Assistance	0.19	0.20	1.9	0.24	0.27	16.2
Child Support Income	0.11	0.11	3.1	0.36	0.40	12.6
Alimony Income	0.07	0.07	3.5	0.43	0.47	8.8
Financial Assistance	0.05	0.05	11.0	0.66	0.74	12.9
Other Income	0.03	0.04	31.5	0.54	0.79	46.0

This table shows the regression R^2 (pseudo R^2 for recipiency logistic and R^2 for value OLS) of the first-stage model selection on the observed responses. The Recipiency (Y/N) shows the logistic result for recipiency of a given income type. For wages and self-employment the recipiency regression was 1) for those with earnings did they have wage earnings or self-employment earnings (Wages)?, and 2) for those with self-employment earnings, did they have non-farm or farm self-employment earnings (Self-employment)? All value regressions are on transformed income conditional on recipiency. The third and sixth columns show the percent difference in the R^2 with the DER administrative data in the model for recipiency and value respectively.

Table 4: Median Income by Selected Characteristics: 2011 Hot Deck, SRMI, and DER SRMI (For Income in 2010)

	Hot Deck			SRMI			DER SRMI			Percentage Difference			
	Median income (dollars)			Median income (dollars)			Median income (dollars)			(HD-SRMI)/HD		(HD-DER SRMI)/HD	
Characteristic	Number (thousands)	Estimate	90 Percent CI	Number (thousands)	Estimate	90 Percent CI	Number (thousands)	Estimate	90 Percent CI		Estimate		Estimate
All Households	119,927	49,276	535	118,682	48,740	934	118,682	48,059	806		1.10	*	2.53
Family households	79,539	61,395	437	78,613	61,153	811	78,613	60,452	802		0.40	*	1.56
Married-couple families	58,656	72,495	716	58,036	71,449	938	58,036	70,783	965	*	1.47	*	2.42
Female householder, no husband present	15,235	31,970	596	15,019	32,669	1,140	15,019	32,151	744		-2.13		-0.56
Male householder, no wife present	5,648	49,813	1,510	5,559	48,503	1,879	5,559	47,526	1,834		2.71	*	4.82
Nonfamily households	40,388	29,578	578	40,069	29,331	890	40,069	29,023	888		0.85		1.92
Female householder	21,420	25,365	621	21,234	25,256	723	21,234	25,037	757		0.43		1.31
Male householder	18,968	35,486	789	18,835	34,664	1,458	18,835	34,185	1,442		2.39	*	3.82
White	96,306	51,709	417	96,144	51,330	716	96,144	50,742	670		0.74	*	1.91
White, not Hispanic	83,314	54,460	734	83,471	53,790	1,003	83,471	52,864	865		1.25	*	3.02
Black	15,265	32,124	821	15,065	31,419	830	15,065	31,431	1,013	*	2.24		2.21
Asian	5,212	64,259	2,591	4,747	62,566	2,980	4,747	61,319	1,862		2.72	*	4.80
Hispanic (any race)	14,435	37,631	957	13,665	37,218	1,004	13,665	36,807	755		1.11	*	2.24
Under 65 years	94,190	55,112	571	93,320	54,228	1,269	93,320	53,403	1,033		1.64	*	3.20
15 to 24 years	6,231	28,224	1,418	6,140	28,132	1,626	6,140	27,937	1,554		0.34		1.04
25 to 34 years	19,487	49,877	906	19,572	47,915	1,415	19,572	47,693	1,201	*	4.10	*	4.58
35 to 44 years	21,458	61,418	816	21,250	60,726	1,287	21,250	60,204	1,592		1.14		2.02
45 to 54 years	24,767	62,341	949	24,530	62,282	989	24,530	61,538	1,032		0.10		1.31
55 to 64 years	22,246	56,474	1,099	21,828	55,722	1,563	21,828	54,819	1,486		1.36	*	3.02
65 years and older	25,737	31,461	563	25,362	31,297	640	25,362	31,101	604		0.53		1.16
Native born	103,232	50,154	446	102,647	49,573	962	102,647	49,020	889		1.18	*	2.32
Foreign born	16,695	43,967	1,727	16,036	43,698	1,953	16,036	42,259	972		0.64	*	4.04
Naturalized citizen	8,568	52,945	1,598	8,277	51,995	1,740	8,277	51,472	1,235		1.84	*	2.87
Not a citizen	8,127	36,413	920	7,758	36,692	986	7,758	35,674	987		-0.76	*	2.07

This table shows the SRMI results *without* administrative data 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 used for SRMI only.

Table 4: Median Income by Selected Characteristics: 2011 Hot Deck, SRMI, and DER SRMI (For Income in 2010), Continued

	Hot Deck			SRMI			DER SRMI			Percentage Difference			
	Median income (dollars)			Median income (dollars)			Median income (dollars)			(HD-SRMI)/HD		(HD-DER SRMI)/HD	
Characteristic	Number (thousands)	Estimate	90 Percent CI	Number (thousands)	Estimate	90 Percent CI	Number (thousands)	Estimate	90 Percent CI		Estimate		Estimate
Households with householders aged 18 to 64	93,997	55,175	554	93,132	54,292	1,261	93,132	53,468	1,029		1.63	*	3.19
With disability	8,951	25,496	1,140	8,827	26,421	1,100	8,827	25,366	1,122	*	-3.50		0.51
Without disability	84,632	58,532	720	83,888	57,294	1,012	83,888	56,777	814	*	2.16	*	3.09
Northeast	21,721	52,996	1,686	21,597	52,118	1,200	21,597	51,925	1,024		1.69		2.06
Midwest	26,772	48,241	885	26,669	48,113	1,415	26,669	47,259	1,167		0.27	*	2.08
South	44,912	45,442	864	44,161	45,027	1,171	44,161	44,449	1,153		0.93	*	2.23
West	26,522	52,959	1,267	26,254	52,297	1,014	26,254	51,812	969		1.27	*	2.22
Inside metropolitan statistical areas	100,343	51,124	425	99,266	50,496	755	99,266	50,143	721		1.25	*	1.96
Inside principal cities	39,956	43,874	1,222	39,472	43,264	1,506	39,472	42,741	1,127		1.42	*	2.66
Outside principal cities	60,387	55,996	683	59,793	55,285	966	59,793	54,723	1,114		1.29	*	2.33
Outside metropolitan statistical areas	19,584	40,173	1,021	19,417	40,734	1,072	19,417	39,645	1,317		-1.38		1.33

This table shows the SRMI and DER SRMI results compared to the hot deck. 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 used for SRMI and DER SRMI only.

Table 5: People and Families in Poverty by Selected Characteristics: 2011 Hot Deck Imputation and SRMI

Characteristic	Hot Deck					SRMI					Difference in Poverty (SRMI-HD)/HD		
	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Number		Percent
PEOPLE													
Total	306,098	46,343	842	15.1	0.3	306,101	46,687	790	15.3	0.3	343		0.1
Family Status													
In families	250,152	33,120	728	13.2	0.3	250,197	32,961	707	13.2	0.3	-159		-0.1
Householder	79,529	9,400	218	11.8	0.3	79,561	9,401	189	11.8	0.2	1		0.0
Related children under 18	72,581	15,598	364	21.5	0.5	72,587	15,558	330	21.4	0.4	-39		0.0
Related children under 6	23,891	6,037	197	25.3	0.8	23,891	6,013	138	25.2	0.6	-24		-0.1
In unrelated subfamilies	1,680	774	115	46.1	4.8	1,680	756	85	45.0	4.1	-17		-1.0
Reference person..	654	283	42	43.2	4.7	654	275	30	42.1	3.8	-8		-1.2
Children under 18	933	469	73	50.3	4.9	933	458	53	49.1	4.3	-11		-1.1
Unrelated individual	54,245	12,449	369	23.0	0.6	54,251	12,969	245	23.9	0.4	* 520	*	0.9
Race[3] and Hispanic Origin													
White alone	240,023	31,083	675	13.0	0.3	239,976	31,063	597	12.9	0.2	-20		0.0
White alone, not Hispanic.	194,850	19,251	550	9.9	0.3	194,757	18,973	369	9.7	0.2	-278		-0.1
Black alone	39,277	10,746	410	27.4	1.0	39,283	11,045	253	28.1	0.6	* 299	*	0.8
Asian alone	15,614	1,899	175	12.2	1.1	15,610	1,897	126	12.2	0.8	-2		0.0
Hispanic (of any race)	50,970	13,522	427	26.5	0.8	50,973	13,897	413	27.3	0.8	375		0.7
Sex													
Male	149,768	20,893	469	14.0	0.3	149,719	21,122	399	14.1	0.3	230		0.2
Female	156,427	25,451	473	16.3	0.3	156,394	25,564	436	16.3	0.3	114		0.1
Age													
Under 18 years	73,860	16,286	366	22.1	0.5	73,874	16,227	341	22.0	0.4	-59		-0.1
18 to 64 years	192,438	26,499	557	13.8	0.3	192,460	26,848	513	14.0	0.3	349		0.2
65 years and over	39,759	3,558	162	9.0	0.4	39,771	3,611	119	9.1	0.3	53		0.1
Nativity													
Native	266,703	38,485	796	14.4	0.3	266,697	38,746	663	14.5	0.2	261		0.1
Foreign born	39,408	7,858	297	19.9	0.7	39,405	7,941	233	20.2	0.5	83		0.2
Naturalized citizen	17,338	1,954	120	11.3	0.7	17,342	1,986	109	11.5	0.6	32		0.2
Not a citizen	22,062	5,904	271	26.8	1.1	22,061	5,955	179	27.0	0.7	51		0.2

This table shows the SRMI results *without* administrative data 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 used for 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 5: People and Families in Poverty by Selected Characteristics: 2011 Hot Deck Imputation and SRMI, Continued

Characteristic	Hot Deck					SRMI					Difference in Poverty (SRMI-HD)/HD			
	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Number		Percent	
Region														
Northeast	54,728	7,038	325	12.9	0.6	54,710	7,165	252	13.1	0.5	127		0.2	
Midwest	66,018	9,216	404	14.0	0.6	66,034	9,162	254	13.9	0.4	-55		-0.1	
South	113,690	19,123	573	16.8	0.5	113,672	19,240	409	16.9	0.4	118		0.1	
West	71,723	10,966	451	15.3	0.6	71,696	11,120	325	15.5	0.4	154		0.2	
Residence														
Inside metropolitan statistical areas	258,333	38,466	925	14.9	0.3	258,333	38,833	788	15.0	0.3	367		0.1	
Inside principal cities	98,798	19,532	584	19.8	0.5	98,825	19,953	471	20.2	0.4	420		0.4	
Outside principal cities	159,506	18,933	741	11.9	0.4	159,538	18,880	521	11.8	0.3	-54		0.0	
Outside metropolitan statistical areas	47,771	7,877	542	16.5	0.7	47,762	7,854	330	16.4	0.5	-24		-0.1	
Work Experience														
Total, 18 to 64 years	192,438	26,499	557	13.8	0.3	192,460	26,848	513	14.0	0.3	349		0.2	
All workers (18 to 64 years)	143,709	10,462	280	7.3	0.2	143,693	11,760	331	8.2	0.2	*	1,298	*	0.9
Worked full-time year-round	95,580	2,600	119	2.7	0.1	95,676	4,080	208	4.3	0.2	*	1,480	*	1.5
Not full-time year-round	47,999	7,862	245	16.4	0.5	47,989	7,680	176	16.0	0.3		-182		-0.4
Did not work at least one week	48,788	16,037	432	32.9	0.7	48,795	15,088	278	30.9	0.5	*	-948	*	-1.9
Disability Status[5]														
Total, 18 to 64 years	192,438	26,499	557	13.8	0.3	192,460	26,848	513	14.0	0.3	349		0.2	
With a disability	14,975	4,196	194	28.0	1.0	14,973	4,140	115	27.7	0.6	-56		-0.4	
Without a disability	176,542	22,227	494	12.6	0.3	176,588	22,628	472	12.8	0.3	401		0.2	

This table shows the SRMI results *without* administrative data 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 used for 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 6: People and Families in Poverty by Selected Characteristics: 2011 Hot Deck Imputation and DER SRMI

Characteristic	Hot Deck					DER SRMI					Difference in Poverty (DER SRMI-HD)/HD			
	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI		Number		Percent
PEOPLE														
Total	306,098	46,343	842	15.1	0.3	306,170	47,481	711	15.5	0.2	*	1,138	*	0.4
Family Status														
In families	250,152	33,120	728	13.2	0.3	250,210	33,758	579	13.5	0.2		638		0.3
Householder	79,529	9,400	218	11.8	0.3	79,559	9,623	182	12.1	0.2		223		0.3
Related children under 18	72,581	15,598	364	21.5	0.5	72,578	15,790	284	21.8	0.4		192		0.3
Related children under 6	23,891	6,037	197	25.3	0.8	23,892	6,139	141	25.7	0.6		102		0.4
In unrelated subfamilies	1,680	774	115	46.1	4.8	1,680	773	77	46.0	3.4		0		-0.1
Reference person..	654	283	42	43.2	4.7	654	281	28	43.0	3.3		-2		-0.2
Children under 18	933	469	73	50.3	4.9	933	467	48	50.0	3.6		-2		-0.3
Unrelated individual	54,245	12,449	369	23.0	0.6	54,250	12,949	249	23.9	0.4	*	500	*	0.9
Race[3] and Hispanic Origin														
White alone	240,023	31,083	675	13.0	0.3	239,982	31,850	538	13.3	0.2	*	767	*	0.3
White alone, not Hispanic	194,850	19,251	550	9.9	0.3	194,790	19,549	457	10.0	0.2		298		0.2
Black alone	39,277	10,746	410	27.4	1.0	39,284	10,929	320	27.8	0.8		183		0.5
Asian alone	15,614	1,899	175	12.2	1.1	15,611	1,948	116	12.5	0.7		49		0.3
Hispanic (of any race)	50,970	13,522	427	26.5	0.8	50,967	14,083	273	27.6	0.5	*	561	*	1.1
Sex														
Male	149,768	20,893	469	14.0	0.3	149,743	21,485	380	14.3	0.2	*	593	*	0.4
Female	156,427	25,451	473	16.3	0.3	156,375	25,996	377	16.6	0.2	*	545	*	0.4
Age														
Under 18 years	73,860	16,286	366	22.1	0.5	73,873	16,472	288	22.3	0.4		186		0.3
18 to 64 years	192,438	26,499	557	13.8	0.3	192,499	27,327	498	14.2	0.3	*	828	*	0.4
65 years and over	39,759	3,558	162	9.0	0.4	39,782	3,681	142	9.3	0.3		123		0.3
Nativity														
Native	266,703	38,485	796	14.4	0.3	266,748	39,361	668	14.8	0.2		876		0.3
Foreign born	39,408	7,858	297	19.9	0.7	39,408	8,120	189	20.6	0.4	*	262	*	0.7
Naturalized citizen	17,338	1,954	120	11.3	0.7	17,343	2,032	84	11.7	0.5		78		0.4
Not a citizen	22,062	5,904	271	26.8	1.1	22,062	6,088	163	27.6	0.6	*	184	*	0.8

This table shows the SRMI results *with* administrative data 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 used for 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 6: People and Families in Poverty by Selected Characteristics: 2011 Hot Deck Imputation and DER SRMI, Continued

Characteristic	Total	Hot Deck				DER SRMI				Difference in Poverty (DER SRMI-HD)/HD			
		Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Total	Number in Poverty (Thousands)	90 percent CI	Percent in Poverty	90 percent CI	Number		Percent	
Region														
Northeast	54,728	7,038	325	12.9	0.6	54,710	7,306	218	13.4	0.4	* 268		* 0.5	
Midwest	66,018	9,216	404	14.0	0.6	66,048	9,405	309	14.2	0.5	189		0.3	
South.	113,690	19,123	573	16.8	0.5	113,682	19,406	386	17.1	0.3	283		0.3	
West.	71,723	10,966	451	15.3	0.6	71,707	11,364	357	15.8	0.5	397		0.5	
Residence														
Inside metropolitan statistical areas	258,333	38,466	925	14.9	0.3	258,339	39,484	728	15.3	0.3	* 1,019		* 0.4	
Inside principal cities	98,798	19,532	584	19.8	0.5	98,814	20,113	446	20.4	0.4	* 580		* 0.6	
Outside principal cities	159,506	18,933	741	11.9	0.4	159,545	19,372	536	12.1	0.3	438		0.3	
Outside metropolitan statistical areas	47,771	7,877	542	16.5	0.7	47,763	7,996	344	16.7	0.5	119		0.2	
Work Experience														
Total, 18 to 64 years	192,438	26,499	557	13.8	0.3	192,499	27,327	498	14.2	0.3	* 828		* 0.4	
All workers (18 to 64 years)	143,709	10,462	280	7.3	0.2	143,699	11,740	301	8.2	0.2	* 1,278		* 0.9	
Worked full-time year-round	95,580	2,600	119	2.7	0.1	95,689	3,973	189	4.2	0.2	* 1,373		* 1.4	
Not full-time year-round	47,999	7,862	245	16.4	0.5	47,987	7,767	172	16.2	0.3	-95		-0.2	
Did not work at least one week	48,788	16,037	432	32.9	0.7	48,795	15,587	291	31.9	0.5	* -450		* -0.9	
Disability Status[5]														
Total, 18 to 64 years	192,438	26,499	557	13.8	0.3	192,499	27,327	498	14.2	0.3	* 828		* 0.4	
With a disability	14,975	4,196	194	28.0	1.0	14,973	4,322	119	28.9	0.6	* 126		* 0.8	
Without a disability	176,542	22,227	494	12.6	0.3	176,604	22,937	451	13.0	0.3	* 711		0.4	

This table shows the SRMI results *with* administrative data 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 used for 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.