Testing New Imputation Methods for Earnings collected by the Survey of Income and Program Participation*

Gary Benedetto and Martha Stinson†

October 19, 2009

Abstract

This paper explores the feasibility and effectiveness of three significant changes to standard Census Bureau methods of imputing earnings in the Survey of Income and Program Participation (SIPP). Currently imputation is performed by stratifying the data based on a set of analyst-chosen characteristics, randomly sorting within each sub-group, and choosing a donor based on the nearest neighbor. We investigated the possibility of using a model-based approach, supplementing survey-collected job and demographic characteristics with administrative earnings data, and using multiple imputation as proposed by Rubin. We modeled monthly earnings from January 2004 to December 2005 using the SIPP 2004 panel linked to W-2 tax records extracted from the Social Security Master Earnings file. We used linear regression techniques to estimate a posterior predictive distribution that is the distribution of earnings conditional on all observed characteristics (including administrative earnings). From this distribution, we took four draws to create four imputed values per case with missing earnings. We compare results using original versus new imputed values from several standard analyses in order to assess the impact of our new method. In particular, we looked at coefficients in a classic earnings regression, trends in income changes over time, the moments of the cross-sectional earnings distribution for a particular month, and income for a small sub-sample of respondents with a high imputation rate. The four imputed values allow us to calculate variance estimates using Rubin's multiple imputation variance formulae and to assess the impact of imputation on the significance of our results.

Key Words: SIPP, imputation, earnings, administrative data, multiple imputation

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau or any of the project sponsors. All of the data used in this paper are confidential data.

†U.S. Census Bureau and U.S. Census Bureau, corresponding author contact information: martha.stinson@census.gov, U.S. Census Bureau, HQ-6H138E, 4600 Silver Hill Road, Washington, D.C. 20233, 301-763-5296

1 Introduction

Over the past two years, the Census Bureau has begun to redesign one of its major surveys, the Survey of Income and Program Participation (SIPP). As well as making siginificant changes to the data collection instrument and methods, Census is also considering how to better process the data once it is collected. In particular, imputation methods have come under scrutiny because they have not been significantly changed since the inception of the SIPP. The goal of this paper is to explore the feasibility and effectiveness of alternative methods of imputation. We chose one important income variable, monthly SIPP earnings at the job level, and investigated three siginificant changes to the imputation procedure for this variable. First, we used a model-based approach. Second, we supplemented survey-collected job and demographic characteristics with administrative earnings data. Third, we used multiple imputation as proposed by Rubin.

Using these new techniques, we "completed" the missing SIPP earnings data by imputing again. We then compared results using original versus new imputed values from several standard analyses in order to assess the impact of our new method. In particular, we will look at distributions of earnings, correlation coefficents between earnings and administrative data, coefficients in a classic earnings regression, trends in earning changes over time, and average earnings of some particular sub-samples. Using the multiple imputed values, we calculated variance estimates using Rubin's multiple imputation variance formulae and assessed the impact of imputation on the significance of regression coefficients and the variance of mean earnings for particular sub-samples.

2 Background

Imputation in the SIPP is currently performed using a hot-deck technique. The hot deck is described by McBride and McKee [McBride and McKee, 2008] as:

> The hot-deck method essentially involves replacing individual missing data items with reported data from another person or household with similar characteristics. Initially, the input file is sorted by geographical keys: PSU, Segment, and Serial Number; this ensures that neighboring records represent geographically proximate units. Edits and imputations are then performed sequentially by unit for each topical section: demographics, household characteristics, labor force, assets, general income, health insurance, and program participation. Each section is processed completely before the next section is done. A hot deck array is created for each edited variable and is stratified by selected variables such as age, race, sex, etc.. Hot decks are first initialized with pre-defined values (referred to as cold deck values), then loaded with live data by passing through the data one time. The data are then passed a second time with good responses contributing to the hot deck and missing responses allocated from

> the hot deck. Allocation flags are also defined for each edited variable and are set to a 1 when value was allocated from the hot deck. Each hot deck cell will contain exactly one value at any point in the edit: either the cold deck value, or the most recently encountered good value meeting the same criteria for that cell - as defined by the stratifying variables. The hot deck imputation process as currently implemented is fully deterministic: subsequent re-processing using the same file and same edit program will result in identical imputations (McBride and McKee, 2008).

There are several assumptions built into the hot-deck imputation method. First, the stratified matrix of donors must have reasonable cell sizes in order to get good donors. If this is not true, then some method of expanding the cell size in order to find a donor must be used. In practice this means that cold deck values are used. In the case of the SIPP, cold deck values are supplied by Census Bureau subject matter analysts and often represent averages in the population. This essentially creates a more heterogeneous group of donors without any way of controlling for their differing characterisitics. Second, current methods rely on the assumption that the relationship between SIPP variables is the same for everyone, regardless of whether these variables contain missing data or not. For example, if age, gender, and industry are used to create a hot deck matrix for earnings, this assumes that the relationship between these three observed values and earnings is the same for everyone, regardless of whether earnings are reported or not. If this assumption is false and respondents with missing earnings data are somehow different than the rest of the sample, then donated earnings values could skew the earnings distribution. The final assumption is that imputing values does not add additional variance to estimates produced using the "completed" data (i.e. data which no longer has missing values). However Rubin argues that an impute is a draw from a distribution and hence should be considered as a random variable with a variance[Rubin, 1987]. Multiple imputation, or taking multiple draws from this distribution, should be done so that the additional variance can be estimated.

In our new imputation procedure, we use methods which do not require any of the three assumptions above. First, we utilize a model-based approach in order to overcome problems associated with small cells potentially produced by stratification. The model also relies on stratification but as small cells are combined, stratification variables that are dropped can be added into the model directly, thus providing an additional way to control for heterogeneous characteristics amongst a block of otherwise similar respondents. Second, we merge administrative earnings data with our survey data, which provides a second, independent source of information on earnings. This second type of data can be used to differentiate between otherwise observationally equivalent people and account for missing patterns that are not fully random. Finally we impute multiple times and create four implicates. Each implicate has the exact same variables and number of observations. For respondents with non-missing data, their values are identical across the four implicates. For respondents who were

missing earnings, their imputed values vary across the four implicates. Multiple imputation allows users to quantify the variance introduced by the modeling of missing data and to produce more accurate variance estimates of their statistics of interest. We give formuale for calculating this new variance estimate in Section 4.3.

In summary, we will model monthly job-level earnings from January 2004 to December 2005 using the SIPP 2004 panel linked to W-2 tax records extracted from the Social Security Master Earnings file. We will use linear regression techniques to estimate a posterior predictive distribution that is the distribution of earnings conditional on all observed characteristics (including administrative earnings). From this distribution, we will take four draws to create four imputed values per case with missing earnings. Our paper proceeds as follows: first we described our data and which cases we chose to impute in Section 3; second, we describe our methodology in more detail, explaining how we estimate the posterior predictive distribution and take draws, in Section 4. Then we present results in Section 5 and show comparisons between the old and new imputations. Finally we conclude in Section 6.

3 Data

The SIPP is a longitudinal data set that is collected in panels, which are samples of respondents that are re-interviewed at set frequencies over the course of two to four years. The SIPP data that are the focus of this study come from the job-level data collected every four months over the course of the SIPP panel that began in the year 2004. We use data from the first eight interviews, or waves, of this panel. The survey asks respondents to report information about a maximum of two jobs per wave. These jobs are tracked across waves and linked by a common employer identification number. Individuals report industry, occupation, firm size, usual weekly hours, type of job (profit, government, etc.), union participation, and earnings. In the public-use data sets released by the Census Bureau, earnings are reported for these two main jobs at the monthly level. Thus it is possible to construct a time series of monthly earnings at a given job from the beginning of the SIPP panel to the end of the job or the end of the panel, whichever occurs first. We create a data set with one observation per person, per job that contains both person and job characteristics and monthly earnings from January 2004 to December 2005. We then check the months when a job was on-going and determine whether earnings were reported or imputed. Imputation flags indicate when a hot-deck imputation for a specific month's earnings was performed due to item non-response. We return all the imputations to missing but keep the data that were reported by the respondent or a proxy.

We then subset the data to include only individuals who were 15 or older at the time of the job and who had Social Security Numbers (SSNs) that were found in the internal Census Bureau databases and deemed to be good matches to the individuals. We also dropped jobs that were unpaid family jobs or that

were originally imputed by type Z imputation, where a donor record was used for a substantial portion of the job record and hence it would be impossible to discard the earnings impute but keep the job characteristic data. Finally, we consider only months when the job was on-going and when an interview was recorded for the person, either by self or proxy. Thus we do not consider months when a person missed an entire wave because they were not able to be interviewed. Although there is a substantial amount of this kind of missing data, the SIPP has rarely done missing wave imputation and our focus was on changing the methods of imputation, not the scope.

After preparing the SIPP person-job file, we match it to the Detailed Earnings Record (DER) extract from the Social Security Administration Master Earnings File (MEF). The MEF is the official repository for historical W-2 data and is used by SSA to calculate benefit eligibility. The DER extract contains one record per employer per year and has uncapped earnings from Box 1 of the W-2 form, essentially wages, tips, and salary taxable under income tax law. Although both the SIPP and the DER are job-level files, for the purposes of this paper, we matched at the person level using the SSN. We made this choice because matching a specific job from the SIPP to a specific job in the DER cannot be done with certainty. There is no common identifier on the two files at the job level. Thus any matching would have to be done using probabilitic linking and we did not want to introduce this level of complication or an additional source of data error due to mis-matched jobs. Hence we summed all the DER records to the person level and merged on total earnings from all jobs for an individual in the years 1978-2006, as well as a count of the total number of employers in the DER in each year. Earnings and total employers from the DER were some of the main explanatory variables in our modeling. For this purpose, we only use individuals with valid SSNs who matched to the administrative data.

In summary, we imputed earnings for people who were 15+ years old and who had valid SSNs and W-2 records in the DER, for jobs that were not unpaid family jobs or original type Z imputations, and for months when the job was on-going and the person had not missed the entire interview and had missing data due only to item non-response. Table 1 gives summary statistics for six months in our 24 month period.

4 Methodology

We begin with a general description of the theory of multiple imputation and then relate the specifics of how we performed our imputation.

4.1 Theory of Multiple Imputation

Since the late 1970's, the theory and techniques for multiple imputation in order to fill missing data have been developed and refined [Rubin, 1996]. These methods offer an analytically useful set of completed data that allows the analyst

to measure the variability introduced through imputation and properly take that into account in estimating statistics and their measures of uncertainty. Adapting Rubin's notation to our missing data problem, the data can be expressed as Y where Y is a matrix of variables at least some of which contain missing values. Y can be expressed as (Y_{obs}, Y_{miss}) where Y_{obs} represents the observed values of Y and Y_{miss} represents the missing values of Y. The inclusion indicator, I, is a structure equivalent in size to Y with elements equal to 1 where Y is non-missing and 0 otherwise. The database can then be expressed by the joint distribution, $p(Y, I, \theta)$, where θ are unknown parameters. In this case, the missing data mechanism is said to be missing at random if

$$p(I|Y) = p(I|Y_{obs}). \tag{1}$$

To create multiple implicates, draws are taken from the posterior predictive distribution (PPD)

$$p(\widetilde{Y}|Y_{obs}) = \int p(\widetilde{Y}|\theta)p(\theta|Y_{obs})d\theta \tag{2}$$

to produce M multiply-imputed completed data files Y^m where $Y^m = (Y_{obs}, \widetilde{Y}^m)$ for $m = 1, ..., M$. The resulting M data files are individually referred to as completed implicates.

In practice, it is very difficult to estimate the joint likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$, especially in a case such as ours where there are many variables, both continuous and discrete, that relate to eachother in very complex ways. As a result, we chose to estimate a sequence of univariate conditional models. Letting $\mathbf{Y} = [\mathbf{y}_0 \ \ \mathbf{y}_1 \ \ \mathbf{y}_2 \ \ \cdots \ \ \mathbf{y}_K]$ (where $\mathbf{y}_0$ is a subset of columns of Y containing no missing data and each $\mathbf{y}_k$ for $k > 0$ has non-missing elements $\mathbf{y}_{k,obs}$ and missing elements $\mathbf{y}_{k,miss}$), and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \ \ \boldsymbol{\theta}_2 \ \ \cdots \ \ \boldsymbol{\theta}_K]$, the joint likelihood can be factorized as:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = p_1(\mathbf{y}_1|\mathbf{y}_0, \boldsymbol{\theta}_1)\, p_2(\mathbf{y}_2|\mathbf{y}_0, \mathbf{y}_1, \boldsymbol{\theta}_2) \cdots p_K(\mathbf{y}_K|\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{K-1}, \boldsymbol{\theta}_K) \tag{3}$$

Estimating a univariate conditional model for each $\mathbf{y}_k$ permits sequential generation of completed values $\mathbf{Y}^c = [\mathbf{y}_0 \ \ \mathbf{y}_1^c \ \ \mathbf{y}_2^c \ \ \cdots \ \ \mathbf{y}_K^c]$ (where $\mathbf{y}_k^c$ is the same as $\mathbf{y}_k$ except $\mathbf{y}_{k,miss}$ has been replaced by draws $\tilde{\mathbf{y}}_k$). That is, sample $\tilde{\mathbf{y}}_1$ from the posterior predictive distribution of $\mathbf{y}_1$ given $\mathbf{y}_0$**,** then $\tilde{\mathbf{y}}_2$ from the posterior predictive distribution of $\mathbf{y}_2$ given $\mathbf{y}_0$ and $\tilde{\mathbf{y}}_1$, etc. Doing this independently M times results in completed implicates, $\mathbf{Y}^{c1}, ..., \mathbf{Y}^{cM}$.

4.2 Implementation

In practice, we do the earnings imputation in two steps. For months that were in sample according to the criteria described in Section 3, we first imputed whether the individual had positive earnings or not. While it is a relatively rare event for individuals to have zero earnings while still holding the job, it does happen. In January 2004, about 5% of in-scope months had zero earnings

(see Table 1 for more details). Because it is difficult to model distributions with significant mass at any particular point, we decided to first impute the fact of positive earnings and then, for individuals who were imputed to have positive earnings, we modeled the distribution of the positive dollar amount. We describe each of these two steps below.

4.2.1 Indicator for positive earnings

After some experimentation, we decided that the overall size of the group of zero-earners was not large enough to successfully estimate the PPD using logistic models and we switched to using Bayesian Bootstrap. The Bayesian bootstrap (BB) was originally defined by Rubin[Rubin, 1981]. As explained therein, the BB is used to simulate the posterior distribution of the parameter whereas the regular bootstrap simulates the sampling distribution of the parameter. Whereas a conventional bootstrap assumes that the sample cumulative density function (CDF) is equal to the population CDF, the BB properly accounts for the uncertainty of the sample CDF. The following is a generic description of the Bayesian bootstrap.

Let X $(n \times k)$ be the source data matrix and Y $(s \times k)$ be the target data matrix. This means that we want to construct an $s \times k$ Bayesian bootstrap sample from an $n \times k$ matrix of source data. Each BB replicate ℓ is a unique $Y^{(\ell)}$.

1. Draw $n-1$ random variables from $U(0,1)$.
2. Sort u_i ascending and let $u_{(i)}$ denote the order statistics from lowest to highest. Define $u_{(0)} = 0$ and $u_{(n)} = 1$.
3. For $i = 1, \ldots, n$, let $\hat{p}_i = u_{(i)} - u_{(i-1)}$.
4. For $j = 1, \ldots, s$ sample with replacement from the rows X using $\hat{p}_i$ as the probability of selecting row i. Place the sampled row into Y_j.
5. Repeat from step 1 for as many BB replicates as desired.

In other words, beginning with a data matrix, X, that contains values for the k variables of interest, this process assigns a probability of choosing a given observation from X to provide data to a corresponding observation in Y for the k variables. The set of probabilities constitutes a non-parametric representation of the PPD from which the sampling is done. In a conventional bootstrap, because of the assumption that the sample CDF is equivalent to the population CDF, each observation in X would be assigned probability $\frac{1}{n}$ of being chosen. There would be no uncertainty in what probability would be assigned to a given observation. However, the Bayesian bootstrap accounts for the fact that the sample CDF is not the population CDF and hence does not assign equal probability to each observation.

In our case, $k = 1$ and the variable of interest, $posearn_{yrmnth}$, is an indicator variable. Suppose that for 96% of the sample of individuals, $posearn_{yrmnth} = 1$

and that $posearn_{yrmnth} = 0$ for the remaining 4%. In a conventional bootstrap, with each individual assigned a probability of $\frac{1}{n}$ of being chosen, the CDF used for sampling would always give $posearn_{yrmnth} = 1$ a 0.96 probability and $posearn_{yrmnth} = 0$ a 0.04 probability. The resulting target matrix Y would not necessarily have a realized 96%/4% frequency distribution for the two values for $posearn_{yrmnth}$ but all the bootstrap samples would have been drawn from such a distribution. In a Bayesian bootstrap, when each source record is assigned a unique probability whose expected value is $\frac{1}{n}$, the CDF used for BB sampling might have 95% versus 5% probability of drawing $posearn_{yrmnth} = 1$ or 0. The next BB might have 97% versus 3%. The variation in the BB probabilities reflects the fact that the sample proportion of 96% in X is an estimate of the probability that $posearn_{yrmnth} = 1$.

It is essential to the success of the Bayesian bootstrap in accurately replicating statistical properties of the data that the observations in a given source (donor) group and a given target (donee) group be as homogenous as possible. Hence we chose grouping variables such that the rows of (X, Y_{obs}) could be assumed to come from the same joint distribution within each group defined by the unique combinations of values of the grouping variables. We created G_1 initial groups based on the values of the variables in the grouping variable list. One of the main advantages of the Bayesian bootstrap is that the group sizes do not have to be as large as groups where parametric modeling is done. Thus we began with a relatively long list of initial grouping variables. However we then imposed a minimum group size of 20. We used BB to impute for all the groups that met the size criteria, then dropped one or more grouping variables, re-grouped based on the new list, and tested for group size again. We repeated this process until we had imputed for all the necessary observations. Our largest list of grouping variables included: month in SIPP sample (values of 1 to 4 that tell which indicate the month in the interview wave); indicator for positive administrative earnings in that year plus one- and two-year lag and one-year lead (i.e. did respondent have positive administrative earnings in year -1, year -2, and year+1); categorical variable that gives combination of count of SIPP and DER jobs[1]; 3 category race variable; male; 11 category age variable[2],

[1] Categories were:
1. 1 DER job/1 SIPP job
2. 2 DER jobs/2 SIPP jobs
3. >2 DER jobs/ > 2 SIPP jobs
4. 0 DER jobs/ > 0 SIPP jobs
5. 1 DER jobs/ > 1 SIPP job
6. 2 DER jobs/ > 2 SIPP jobs
7. > 2 DER jobs / SIPP jobs > DER jobs
8. 2 DER jobs / < 2 SIPP jobs
9. > 2 DER jobs / SIPP jobs < DER jobs

[2] Age categories were:
1. 15<=age<18
2. 18<=age<22
3. 22<=age<25
4. 25<=age<30
5. 30<=age<35

and when available, one-month, two-month, and 12-month leads and lags of indicator for positive monthly SIPP earnings (i.e. did respondent have positive earnings in month t-1, month t-2, month t-12, month t+1, month t+2, month t+12); indicator for positive usual weekly hours worked; and indicator for usual weekly hours worked reported as "variable" instead of amount. Our shortest list included only month in SIPP sample and indicator for positive administrative earnings .

4.2.2 Earnings

When individuals were imputed to have positive earnings for an in-scope month, we then used regression-based modeling to estimate the PPD. We again chose a set of grouping variables and defined G_1 groups. Within each group, we ran a regression of earnings on a chosen set of explanatory variables using only observations where earnings were non-missing. This regression produced a set of estimated parameters $\theta = (\beta_1 ... \beta_k, \sigma^2)$. As with the BB, the important insight is that these parameters are only estimates of their population counterparts because they are based on a sample. Hence these parameters have distributions. In order to take a draw from the PPD and assign an imputed value, we first take a draw from the distribution of each of the parameters. We then take a draw from the distribution of the error term. For individuals missing earnings, we then calcuate a predicted value using the draws for the regression coefficients and the error term and the observed values for the explanatory variables. This predicted value is the new imputed value. By taking multiple draws from the distributions of the estimated parameters θ, we could obtain multiple predicted values and hence multiple imputes.

In practice we again imposed a minimum group size of 100 or 10 times the number of explanatory variables, whichever was greater, and when groups were too small, we dropped grouping variables from the list, and re-combined individuals into larger groups based on the shorter list. Our largest list of grouping variables included: month in SIPP sample; indicator for positive administrative earnings in that year; categorical variable that gives combination of count of SIPP and DER jobs; one-month lead and lag of indicator for positive monthly SIPP earnings; 4 category age variable[3]. Our shortest list included only month in SIPP sample. When variables were dropped from the the grouping list, they were included in the list of explanatory variables. Our initial list of explanatory variables was: age, race black, race other, male, one- and two-year lags,

6. 35<=age<40
7. 40<=age<45
8. 45<=age<50
9. 50<=age<55
10. 55<=age<62
11. 62<=age

[3] Age categories were:
1. 15<=age<25
2. 25<=age<50
3. 50<=age<62
4. 62<age

and one-year lead of indicator for positive administrative earnings, one- and two-year lags, current year, one-year lead of actual administrative earnings, and one-, two-, and twelve-month leads and lags of indicators of positive monthly SIPP earnings and actual monthly SIPP earnings.

4.3 How to Use Multiple Completed Implicates

We now explain how to perform analysis on the multiple implicates of completed data. Starting with the notation from the previous section, suppose interest focuses on a completed data estimand Q which is a function of Y and has dimensions $(k_q \times 1)$. This estimand can be any computable, vector-valued function of the data. For example, it could be the average value of Y, many moments of Y, conditional moments of some columns of Y given other columns Y, parameters of a model relating columns of Y, percentiles of the distribution of Y, and so on. The essential feature of Q is that it is computable from completed data on the population and, therefore, is not random. In the context of this paper, consider the example of average earnings in January 2004. If we had complete earnings data on every individual in the United States, we could calculate the national average with certainty.

Estimates of Q are random because they are based on $D = [Y_{obs}, I, R]$ (where R contains the information regarding the SIPP sampling design), which involves both sampling from the finite population and incomplete observation of Y in the sample. We can only calculate an estimate of the average January 2004 earnings because of the sampling involved with the SIPP and because not all SIPP individuals provided January 2004 earnings data. Even if all SIPP individuals in our sample reported January 2004 earnings, the sample design of the SIPP would still make this average a random variable. We will call the completed data estimator $q(D)$ and its variance estimator $u(D)$. Notice that because of the definition of completed data, q and u depend only on (Y_{obs}, R) and not on I. The analyst is assumed to have an inference system for $q(D)$ and $u(D)$. In particular, completed data inference can be based on $(q(D) - Q) \sim N(0, u(D))$, which may be exact or an approximation but is assumed to be appropriate in what follows.

In the classic Rubin missing data application[Rubin, 1987], Y_{miss} is imputed M times by sampling from $p(Y_{miss}|Y_{obs})$, the posterior predictive distribution of Y_{miss} given D. The completed data consist of M implicates $D^{(m)} = \{D, Y^m\}$, where Y^m is as defined in the previous section. Continuing the example of January 2004 earnings, we estimate the posterior predictive distribution of missing January 2004 earnings conditional on everything else we observe about the individual (surrounding years' income, gender, race, marital status, etc.). We sampled four times and created four implicates $D^{(1)}$, $D^{(2)}$, $D^{(3)}$, and $D^{(4)}$, each of which consists of original non-missing January 2004 earnings data (D) and imputed January 2004 earnings $(\widetilde{Y}^1 ... \widetilde{Y}^4)$. Inference is based on the following formulae:

$$\text{statistic calculated on each implicate file:} \quad q^{(m)} = q\left(D^{(m)}\right).$$

In our example the function q is the average of January 2004 earnings across all individuals in the sample. This average is calculated separately for each implicate and then averaged across implicates as the next formula indicates:

$$\text{average of the statistic across implicates:} \quad \bar{q}_c = \sum_{m=1}^{M} \frac{q^{(m)}}{M}.$$

In order to draw proper inferences about $\bar{q}_c$, the correct variance measure must be used. The variance of $\bar{q}_c$ has two parts. The first part is commonly referred to as the "between-implicate" variance, defined by the following formula:

$$\text{variance of the statistic across implicates:} \quad b_c = \sum_{m=1}^{M} \frac{\left(q^{(m)} - \bar{q}_c\right)\left(q^{(m)} - \bar{q}_c\right)'}{M-1}$$

The measure b_c tells how much variation has been introduced by the multiple draws from the posterior predictive distribution. The second component of the overall variance of $\bar{q}_c$ is calculated by averaging the within implicate variance across implicates. We define the variance of $q^{(m)}$ for each implicate m and the average across implicates as follows:

$$\text{variance of the statistic on each implicate file:} \quad u^{(m)} = u\left(D^{(m)}\right)$$

and

$$\text{average variance of the statistic across implicates:} \quad \bar{u}_c = \sum_{m=1}^{M} \frac{u^{(m)}}{M}.$$

In our continuing example of January 2004 earnings, $u^{(m)}$ is the sampling variance of average earnings (defined as $\frac{s^2_{earn}}{N}$ where N is the number of observations in Y) for each implicate m. The total variance of January 2004 earnings is then calculated as a weighted sum of the between implicate variance and the average within implicate variance, defined as follows:

$$\text{total variance of the average statistic across implicates:} \quad T_c = \bar{u}_c + \left(1 + \frac{1}{M}\right) b_c$$

When N and M are large, inference is based on $(\bar{q}_c - Q) \sim N\left(0, T_c\right)$. When M is moderate and the estimator $\bar{q}_c$ is univariate (*i.e.,* $k_q = 1$), inference is based on $(\bar{q}_c - Q) \sim t_{\nu_c}\left(0, T_c\right)$, where the degrees of freedom ν_c are defined as

$$\nu_c = (M-1)\left(1 + \frac{\bar{u}_c}{\left(1 + \frac{1}{M}\right) b_c}\right)^2$$

Proofs and further details can be found in [Rubin, 1987] and [Rubin, 1996].

5 Results

We begin with Table 1 which shows a summary of the number of observations that were imputed for January 2004, June 2004, December 2004, January 2005, June 2005, and December 2005. Of jobs that were on-going in a given month, between 9% and 10% had imputed earnings in the original SIPP data. For the approximately 90% of non-imputed job-months, between 91% and 95% had positive earnings. For the imputed job-months, using the original imputes, between 95% and 98% had positive earnings. The confidence intervals for these means are also reported in Table 1 and show that the difference between the two groups is statistically significant. A higher percentage of imputed cases have zero earnings. Our revised imputed values for these job-months show the opposite result. Between 88% and 94% of job-months have positive earnings, a lower percentage than the non-imputed cases. In January 2004 this difference is not siginficant since the confidence intervals overlap, but in every other month, the confidence interval for revised imputes lies below that for the non-imputed cases. Still the revised imputations are closer to the original data than the original imputations.

Mean earnings range from $2950 to $3178 across the job-months where no imputation was done. The mean of original imputed earnings ranges from $2819 to 3156 and, with the exception of January 2004, the point estimate of the mean is lower than in the non-imputed data. Mean earnings using our revised imputed values range from $2768 to $3405. Confidence intervals for the different means show that except for January and June 2004, the mean for the original imputes was lower than the mean for the non-imputed cases by a statistically significant amount. However the confidence intervals for the mean always overlap between the two types of imputes, showing that the differences here are not statistically significant. The size of the confidence intervals for our revised imputes reflects the additional variance that we measure by imputing multiple times. The standard deviation of earnings is higher in our revised imputed data than in either the original imputes or the non-imputed job months except for January 2004 where the original imputes have the highest standard deviation.

In Figure 1, we show a graphical representation of the distribution of earnings at jobs that were on-going in January 2004, weighted by the person-month weight for January 2004. The yellow line represents the original earnings distribution and the pink line represents the distribution with the new imputed values. These lines are very similar which is mostly due to the fact that imputed values make up a relatively small percentage of the overall cases. Figure 2 shows the weighted distribution of earnings by imputation group: non-imputed earnings (yellow line), original imputed earnings (pink line), and revised imputed earnings (blue line). Here the differences are more obvious. Our new imputation method shifted the distribution for imputed values to the left. It also smoothed it considerably, most likely due to the modeling. Figure 3 shows the weighted distribution of earnings summed to the person-level (i.e. earnings from all jobs in a given month summed together) and then summed across months in the

year. We included only individuals who did not have any missing waves in a year in order to insure they would have an earnings report of zero or positive in every month of the year, thus allowing us to accurately calculate annual earnings. Here again we see that the distributions with imputed and non-imputed cases together are very similar regardless of whether original or revised imputes were used. We also graph the distribution of annual administrative data which is shifted substantially to the left of SIPP reported and imputed earnings.

Table 2 shows how the SIPP earnings in January 2004 correlate with administrative earnings for 2004, and how that relationship compares between the imputed cases and the non-missing cases. The correlations in this table are calculated without any weights and only for the subset of the population that successfully matches to administrative records. There are five pairs of columns of data. Each pair contains correlation coefficients of SIPP earnings to administrative earnings for non-missing cases and imputed cases broken down by strata defined by gender and race; as a result, the left column (computed on non-missing data) in each pair is identical. The first pair of columns looks at how the old imputations compare to the non-missing cases. The correlation coefficients are lower in general for the old imputations than for the non-missing data, suggesting that the imputations have a weaker relationship with administrative earnings than the reported earnings. The next four pairs of columns show how the imputed values from each of the four implicates using the new imputation method compares to the non-missing data. The correlation to administrative earnings is much stronger for the new imputations than for the old imputations, and appears to be more similar to the correlations found with the reported earnings. The bottom row of the table quantifies this by showing the correlation coefficient across gender and race strata of the correlation coefficients between SIPP and administrative earnings. The idea here is to measure how the correspondence between survey and administrative earnings relates between imputed and non-imputed cases. The larger correlations for the new imputations imply that the model-based method preserves the relationship to administrative earnings found in the non-missing data to a greater degree than the old method. This is not a surprising result, since the new imputations used administrative earnings as conditioning variables, but it is evidence that earnings may correlate differently with other observable survey responses differently for those who report earnings than for those who do not report earnings. Therefore, access to an external source of data can aid imputations in the cases where one suspects data may not be missing at random.

In Tables 3a and 3b, we do a comparison of the old and new imputed values for a small sub-sample of the data. We chose to look at Black women, ages 18-25, living with a parent because of the relatively high rate of earnings imputation for this group. In Table 3a, we show weighted average annual earnings for 2004 for non-imputed cases and both types of imputed cases. For our revised imputations, we show the average by implicate and the average across the four implicates. Mean annual earnings using the old imputed values is substantially higher than either the mean of the non-imputed cases or the mean of the revised imputed cases. The variance of the mean is also much larger using

the original imputed data. This translates into a large confidence interval and an imprecise estimate of the mean. To illustrate how the variance calculation described in Section 4.3 works, we report the variance estimates of each of the four implicates and then the average. This is the average within-implicate variance estimate. We then report the between-implicate variance estimate and the total variance using the previously given formulae. Finally we report standard errors of the mean for each of the three groups: non-imputed, original imputed, and revised imputed. Imputation clearly adds variance. If we had performed only one imputation using our modeling methods, our variance estimate would have been that listed for IMP1 in the second row of Table 3a. But as can be seen when the total variance is reported, this would have been an underestimate of the true variance. Having multiple implicates allows us to calculate the between implicate variance and incorporate this piece into the overall variance calculation. Interestingly, our modeling seems to have been better than the original hot-deck for this group and so inspite of multiple imputation, the variance estimate is still lower for the new imputes than the original ones. It is also important to note that since hot-decking was only done once, the variance estimate for the original imputes is in reality too small because it does not take account of variation introduced due to imputation.

In Table 3b, we combine the imputed and non-imputed cases to show the results for mean annual earnings for the full group. The mean for the full data falls when the new imputes are used and is closer to the mean for the non-imputed group. The standard error of the mean is also smaller using the new imputes, even after properly calculating the total variance to take account of between implicate variance. Thus for this sub-sample of people, our new imputations make a significant change to the mean and its standard error.

In Table 4, we present results from a time-series regression using a GEE model for the error terms, with an AR(1) process within person. The dependent variable is the log wage for a month which we define as monthly earnings divided by usual weekly hours times average number of weeks (4.4). We sorted jobs by start date and total earnings and chose the job which began first and paid the most to be an individual's first dominant employer. If that job ended, we chose the next dominant employer as the job which began next (or was already on-going) and paid the next most. Using this method, we constructed a time series of earnings from dominant employers from January 2004 to December 2005. Our regression model included age, age squared, education level, male, race, labor force experience, job tenure, number of employees at the firm (firm size), and whether the job was covered by a union contract, as well as controls for industry, occupation, and type of job which we do not report in the table.[4] We show coefficients using the original earnings data and the data that includes revised imputes. We also report standard errors and again use the formulae from Section 4.3 to calculate an accurate variance measure. Again it appears that the relatively small number of imputes in the sample overall means that

[4] Type of worker was for-profit, not-for-profit, contingent worker, state and local government, and federal government. Industry was a 12 category classification and occupation was a 6 category classification.

there are no siginificant changes to the regression coefficients or their standard errors. Items in bold are significant at the 5% level and there are no differences in siginificance levels due to the new imputation method.

Finally in Table 5, we consider changes in household income. Earnings are a siginificant portion of income for most households and hence we wanted to explore the impact of our new imputations on income calculations. A 2008 report from the Congressional Budget Office (CBO) [CBO, 2008] estimated changes in household income using multiple SIPP panels and found that changes of +/- 25% or greater were more frequent in the SIPP than in administrative data taken from W-2 records. They looked specifically at imputed cases and found these to have a higher frequency of large income changes than even the non-imputed SIPP data. We attempted to replicate the CBO calculation using both the new and old imputes. We created a new measure of household income that included revised earnings for all household members. Using households where the head was between ages 25 and 55 and where no household member was missing administrative data, we followed CBO and calculated the change between 2004 and 2005 household income (in real terms) as (HTOTINC2005 - HTOTINC2004)/((HTOTINC2005 + HTOTINC2004)/2). We split our sample into a group of households where at least one individual had imputed earnings in at least one month and a group where no individuals had any imputed earnings during either year. Results in table 5 show that for households with imputed data, the revised imputes do not siginficantly alter the percentage of households that experienced a large positive or negative change in income. Both the original and revised imputed data have higher levels of transitions overall than the non-imputed data. Interestingly the positive changes in household income are less frequent in the non-imputed cases whereas the negative changes happen for similar numbers of people. Figure 4 provides a graphical representation of this table to aid in comparing the data types. We were uncertain whether the new imputations produced too much movement in earnings and we needed to iterate our procedure more times in order to have the imputations settle down or whether people with imputed data were somehow more volatile than those without. To help with this judgement, we used our administrative data. We kept non-earned income as reported in the SIPP but replaced SIPP-reported or imputed earnings with administrative earnings in the calculation of household income. We show the results for the group of individuals with reported SIPP earnings and the group with imputed SIPP earnings. Even with the administrative data, the imputed group has a higher overall percentage of households with large income changes than the non-imputed group, leading us to believe that this group is more volatile. What is surprising, however, is that in both groups, the percentage of households with a 25% or more increase in income is larger in the administrative data while the percentage with a 25% or larger decrease in income in smaller. While overall both groups have higher volatility than the administrative data, this appears to be solely the result of too many large drops in income. One possible explanation for this finding is that the SIPP sometimes missed earnings which led to spurious large declines (i.e. reporting error). It is also possible that there is a lag in reporting earnings increases (due

perhaps to seam bias) and this causes annual earnings to be underestimated for individuals who received pay raises.

6 Conclusions

In conclusion, we believe that the three major changes to the SIPP job-level earnings imputation procedure have shown the potential to add value to the SIPP. Modeling as opposed to using hot-deck procedures smooths the distribution and provides seemingly better estimates for small groups with high levels of imputation. It does seem that the earnings data are not missing at random and that the administrative data have helped to improve the imputes. Finally multiple imputation gives more accurate variance estimates but when combined with the improvements from the modeling, does not always create higher levels of variance than the old imputations did.

While we believe that our revised imputation methods have the potential to improve SIPP estimates, we have no evidence that changing imputation methods will alter the overall distribution of earnings. This is an important point to understand about imputation and the use of additional sources of data like W-2 earnings. We have modeled the distribution of self-reported survey earnings, and as such, have mostly preserved that distribution. As shown in Figure 3, the distribution still differs greatly from the distribution of administrative data. In short, using administrative earnings can help with the not-missing-at-random problem but cannot fix overall under-reporting of earnings in the SIPP.

It is sometimes suggested that missing survey data be "imputed" by simply filling in administrative data. This seems likely to alter the distribution of survey earnings in a problematic way. Because this would be done only for some individuals and not for others, later comparisons might not be valid. Suppose, for example, that Group A mostly reported their earnings and Group B did not. Group B received administrative earnings as their "imputed" values. If a researcher then compared Group A to Group B, she might find a large difference in average earnings. But some of this difference would be due to the differences between administrative data and survey data and not to differences between Group A and Group B. The only unbiased comparison between the groups would be to compare administrative data for both groups or survey data for both groups.

There are many possible future refinements that could be made to our model. In particular, it might be both useful and feasible to impute out-of-sample months which would allow complete annual data for a much larger number of respondents. Second, it might be useful to jointly impute other variables concerning jobs, in particular hours worked. Finally with the re-design of the SIPP, imputation methods will undoubtedly need to be tailored to fit the new annual collection scheme and the new method of collecting earnings changes over the course of the year.

References

[CBO, 2008] CBO (2008). Recent trends in the variability of individual earnings and household income. Cbo pub. no. 2996, Congressional Budget Office. http://www.cbo.gov/doc.cfm?index=9507zzz=37668.

[McBride and McKee, 2008] McBride, Z. and McKee, N. (2008). Sipp imputation scheme. Working paper asa-srm presentation, U.S. Census Bureau. http://www.census.gov/sipp/DEWS/ASA-SRM-Imputation

[Rubin, 1981] Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9:130–134.

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

[Rubin, 1996] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Table 1: Summary Statitics for Earnings by imputation group

	Number of Observations		% with Positive Earnings			Confidence Interval of Mean		
Month-Year	No imputes	Imputes	No imputes	Imputes		No imputes	Imputes	
			Original	Original	Revise	Original	Original	Revise
Jan-04	34009	3842	0.952	0.980	0.944	0.949,0.954	0.976,0.984	0.933,0.955
Jun-04	35252	3628	0.912	0.981	0.882	0.909,0.915	0.977,0.986	0.864,0.901
Dec-04	32779	3142	0.920	0.985	0.887	0.917,0.923	0.981,0.989	0.870,0.904
Jan-05	32636	3227	0.914	0.955	0.895	0.911,0.917	0.948,0.962	0.883,0.906
Jun-05	32249	3302	0.925	0.959	0.897	0.922,0.928	0.952,0.966	0.885,0.909
Dec-05	31583	3127	0.936	0.957	0.898	0.933,0.939	0.950,0.965	0.889,0.907

*observation is a person-job-month earnings report or earnings imputation

	Mean earnings			Confidence Interval of Mean			Stand. dev. of earnings		
Month-Year	No imputes	Imputes		No imputes	Imputes		No imputes	Imputes	
	Original	Original	Revise	Original	Original	Revise	Original	Original	Revise
Jan-04	3065.32	3156.61	3405.23	3010.77,3119.88	2800.28,3406.42	3156.11,3654.34	5007.2	9529.8	8771.2
Jun-04	2950.18	2833.49	2854.42	2911.49,2988.87	2754.63,2997.49	2538.17,3170.67	3540.4	3409.3	4561.1
Dec-04	3112.03	2846.60	3103.43	3062.24,3161.82	2708.16,3005.77	2764.30,3442.56	4412.5	3894.1	5665.8
Jan-05	3038.64	2915.50	3034.26	2996.05,3081.23	2707.48,2955.01	2770.44,3298.08	3752.2	3370.2	5746.6
Jun-05	3041.69	2819.04	2768.73	2997.82,3085.56	2671.40,2901.16	2622.10,2915.35	3866.4	3108.3	3875.7
Dec-05	3178.73	2890.09	3047.12	3125.23,3232.24	2723.64,2942.59	2813.44,3280.80	4692.9	2910.9	5029.7

Figure1 Distribution of Job-level Earnings:
Monthly amount January 2004

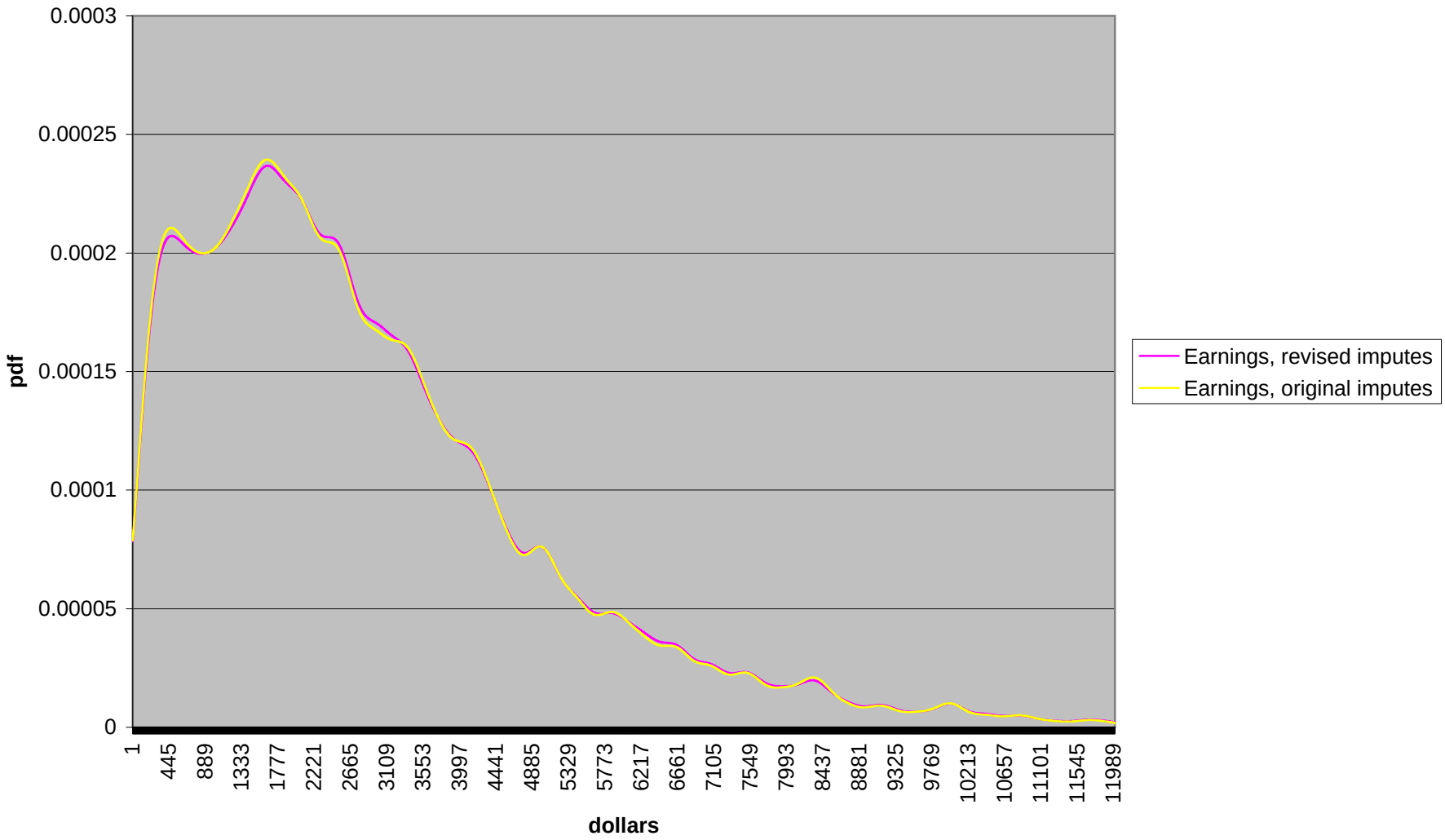

Figure 2 Distribution of Job-level Earnings by imputation group: Monthly amount January 2004

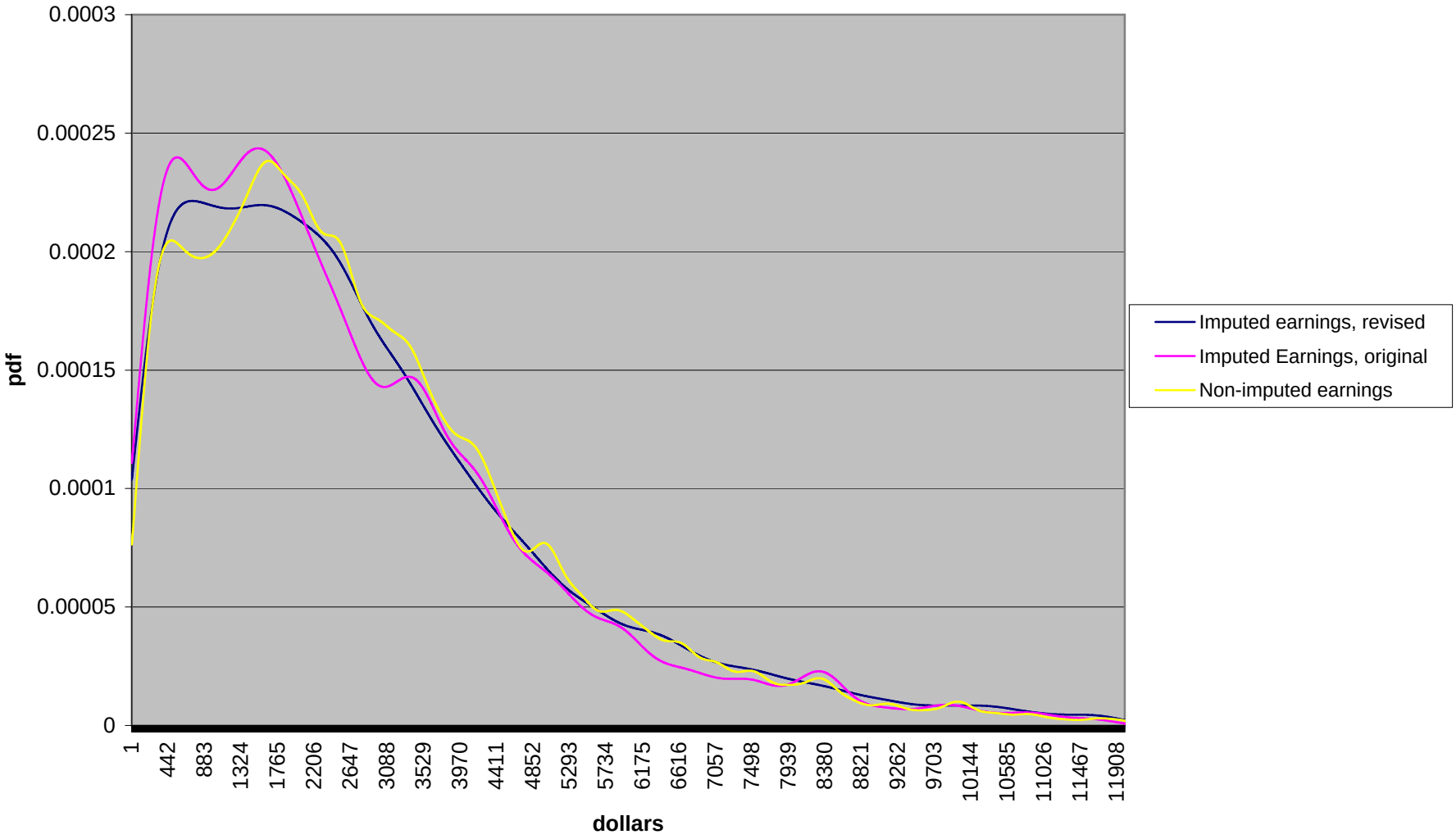

Figure 3 Distribution of Person-level Earnings:
Annual amount 2004

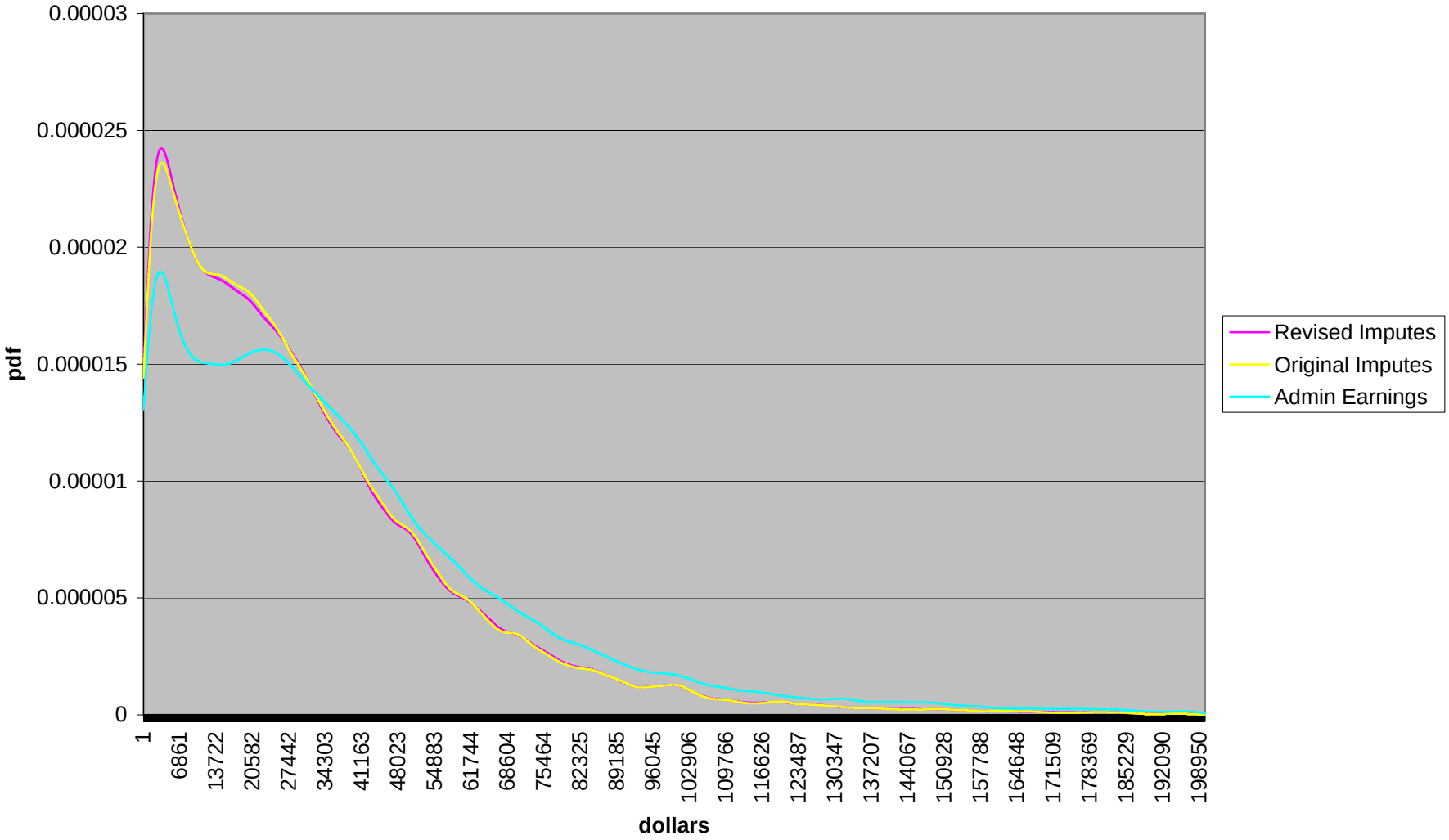

Table 2: Correlation Coefficients of SIPP January 2004 earnings to DER 2004 earnings

		Original SIPP		Implicate 1		Implicate 2		Implicate 3		Implicate 4	
Gender	Race	Non-Miss	Imputed	Non-Miss	Imputed	Non-Miss	Imputed	Non-Miss	Imputed	Non-Miss	Imputed
Female	White	0.62	0.32	0.62	0.71	0.62	0.39	0.62	0.61	0.62	0.61
Female	Black	0.00	0.24	0.00	0.48	0.00	0.61	0.00	0.40	0.00	0.60
Female	Asian	0.74	0.32	0.74	0.69	0.74	0.75	0.74	0.69	0.74	0.67
Female	Other	0.74	0.42	0.74	0.88	0.74	0.81	0.74	0.81	0.74	0.80
Male	White	0.24	0.19	0.24	0.33	0.24	0.52	0.24	0.31	0.24	0.34
Male	Black	0.76	0.29	0.76	0.61	0.76	0.60	0.76	0.42	0.76	0.66
Male	Asian	0.78	0.00	0.78	0.73	0.78	0.79	0.78	0.53	0.78	0.71
Male	Other	0.68	0.60	0.68	0.60	0.68	0.61	0.68	0.81	0.68	0.52
Correlation		0.18		0.75		0.40		0.62		0.56	

Table 3a: Average Annual Earnings 2004, Black Women Ages 18-25, living with parent, by imputation groups

	Mean			Variance of Mean		
	Non-Imputed	Orig Imputed	Revise Imputed	Non-Imputed	Orig Imputed	Revise Imputed
	7120.67	**9730.75**		**564,566**	**2,889,341**	
IMP1			7658.91			1766186
IMP2			7228.03			1897529
IMP3			8139.74			2273244
IMP4			7382.06			1699238
Average			**7602.18**			**1,909,049**

Between Implicate Variance			160,210
Total Variance of Mean			**2,109,312**
Standard Error of the Mean	**751**	**1,700**	**1,452**

Sample Size= 126 non-imputed cases, 37 imputed cases
no missing waves, non-imputed is no imputed months for full year,
imputed group has at least one imputed month

Table 3b: Average Annual Earnings 2004, Black Women Ages 18-25, living with parent,all

	Mean		Variance of Mean	
	Original Data	Revised Data	Original Data	Revised Data
	7744.74		**500,578**	
IMP1		7204.59		424,931
IMP2		7114.39		432,156
IMP3		7320.05		457,748
IMP4		7152.03		422,097
Average		**7197.77**		**434,233**

Between Implicate Variance		8,014
Total Variance of Mean		**444,251**
Standard Error of the Mean	**708**	**667**

Sample Size = 163 cases

Table 4: Estimates from Log Wage Regression
Monthly time series of dominant employers from Jan. 2004 to Dec. 2005

	Coefficient		Std. Error	
Explanatory Variable	Original	New imputes	Original	New imputes
Intercept	**0.958**	**0.900**	0.027	0.030
Age reported in SIPP	**0.050**	**0.053**	0.001	0.001
Age squared	**-0.001**	**-0.001**	0.000	0.000
Educ: HS degree	**0.125**	**0.108**	0.009	0.010
Educ: some college	**0.248**	**0.236**	0.009	0.010
Educ: college degree	**0.552**	**0.526**	0.011	0.012
Educ: grad degree	**0.766**	**0.751**	0.014	0.015
Male	**0.197**	**0.190**	0.006	0.006
Black	**-0.115**	**-0.129**	0.008	0.008
Other race	0.000	-0.007	0.010	0.011
LF Experience (months)	**0.001**	**0.001**	0.000	0.000
Job tenure (months)	**0.001**	**0.001**	0.000	0.000
Firmsize: 25-99	**0.045**	**0.046**	0.009	0.010
Firmsize: 100-499	**0.061**	**0.073**	0.009	0.010
Firmsize: 500-999	**0.062**	**0.059**	0.012	0.014
Firmsize: 1000+	**0.085**	**0.092**	0.007	0.008
Firmsize: missing	-0.106	-0.099	0.055	0.071
Union job	**0.095**	**0.109**	0.008	0.012

*coefficients in bold are siginificant at the 5% level

*dependent variable is log wage = ln of monthly earnings / (usual weekly hours *4.5)

Table 5: Changes in Household Income from 2004 to 2005

Imputed Cases	Nobs	Percentage of HH with +25% change	Percentage of HH with -25% change	Sum
Original SIPP	2850	0.148	0.288	0.436
Revised SIPP	2850	0.150	0.290	0.440
Admin data*	2850	0.214	0.182	0.396
Non-imputed Cases				
Original SIPP	6829	0.098	0.290	0.387
Admin data*	6829	0.169	0.168	0.337

*Admin data is used to replace SIPP earnings, non-earned income is still from SIPP
Includes households with no missing administrative data for HH members,
HH head between ages 25 and 55, and no missing wave data for HH members
Excludes 1st and 99th percentiles of HH earnings distribution

Figure 4: Percentage change in Total Household Income from 2004 to 2005: Imputed and Non-imputed Cases