Statistical Matching Using Propensity Scores:

Theory and Application to the Analysis of the Distribution of Income and Wealth

Hyunsub Kum
Assistant Professor
Graduate School of Public Administration
Seoul National University
San 56-1 Shilim-dong, Gwanak-gu, Seoul,
Korea, 151-742
phone: 82-2-880-5647
hyunsk@snu.ac.kr

Thomas Neal Masterson [corresponding author]
Research Scholar
Levy Economics Institute of Bard College
Blithewood
Annandale-on-Hudson, NY 12504
office: 845-758-7715; fax: 845-758-1149
masterso@levy.org

Abstract

This article considers the usefulness of statistical matching using propensity scores in carrying out economic research. We include an application of the procedure to a statistical match between the 2001 Survey of Consumer Finance (SCF) and the 2002 Annual Demographic Supplement (ADS) of the Current Population Survey (CPS) data sets to demonstrate the procedure and results of the matching. Challenges facing the use of this technique such as the distribution of weights are discussed in the conclusion.

Keywords: Statistical Matching; Survey of Consumer Finances; Annual Demographic Supplement; Distribution of Income and Wealth

JEL Codes: C14 Econometric and Statistical Methods: General: Semiparametric and Nonparametric Methods; C40 Econometric and Statistical Methods: Special Topics: General; D31 Personal Income, Wealth, and Their Distributions.

1. Introduction

Statistical matching has long been used in economics in research on the distribution of income and wealth. The procedure has become much more feasible with advances in computer technology and the development of statistical packages that make use of them. Much of the existing literature consists of theoretical elaborations of the method or analysis of the results of applications of statistical matching. This paper provides a concrete example of an application of statistical matching from the production of the Levy Institute Measure of Economic Well-being (LIMEW) with assessment of the quality of the match.

The earliest uses of statistical matching in economics were in the government statistical offices of the US and Canada. In the US, Okner [18] and Ruggles and Ruggles [26] and in Canada, Alter [2] are early examples of statistical matching applications. The purpose of these early efforts was to produce comprehensive household income estimates. Due to the limited technology available at the time, methods were relatively crude (involving sorting records on several variables and matching them according to their rank). Rässler provides a detailed history of the field [18, 49-52].

While interest in extended income measures has increased recently, producing such measures still requires the combination of household surveys with differing sets of desirable information. The main question addressed in this paper is how best to integrate the information from different household surveys into a single data set that is representative of the population as a whole when no single source of data has all of the information required for the desired undertaking. This is distinct from the question of how to handle missing values. The nature of the problem is not item non-response or censoring due to confidentiality concerns. Rather, the problem is that no single survey collects all of the information needed to address certain research questions. Statistical matching is one of the available methods that can be employed to address this problem.

When combining sets of data together into a new single data set, the structure of the research project will generally determine the choice of a specific statistical matching procedure. For example, many recent medical studies use statistical matching to construct a control group when one was unavailable in the observational study. That type of matching will require a different procedure than the one we will use in the application in this paper, in which we want to create a combined but synthetic data file that is representative at the level of the U.S. national population. The latter case is more restrictive because we need a matching procedure that preserves at least the marginal distributions of the variables of interest. Also this procedure will have to be able to handle the fact that the various sets of data employed in statistical matching are taken from surveys with varying sample designs and weighting schemes.

The rest of the paper is arranged as follows. The second section presents a review of literature concerning statistical matching. The third section outlines a procedure for constrained statistical matching using propensity scores to match the 2001 Survey of Consumer Finances (SCF) and the 2002 Annual Demographic Supplement (ADS) as an example. The fourth section discusses properties of the resulting synthetic data set. The fifth and final section summarizes our findings, draws conclusions, and lays out challenges yet to be dealt with in the procedure.

2. Overview of Statistical Matching

Statistical matching (or data fusion, as it is called in other discipline in Europe) is by now a widely used technique in producing empirical studies. The method is used in many observational studies in the medical literature [16], [23], [24]. In addition to the numerous examples in the field of economics cited by Rässler [20], there are studies by Radner [19], Wolff [33], Wolff and Zacharias [32], Greenwood [7], [8], Wagner [30], Brodaty, Crépon, and Fougère [4], Keister [11], [12], the Urban-Brookings Tax Microsimulation [22], and the 2003 Congressional Budget Office report on income tax burdens [5]. Finally, the U.S. Census Bureau implicitly uses statistical matching in its 'hot-decking' methodology for imputing missing values [31].

In the standard statistical matching framework, one has two data files, file A and file B, with a set of common variables **Z**. File A contains variables **X** that are not available in file B, and file B contains variables **Y** that are not available in file A. One needs a data file with variables **X**, **Y**, and **Z** together, but that kind of data file is unavailable from a single source. Although the variable **Z** is common in both data files, there is no identification information such as social security number (SSN) to link records exactly from each other, or in fact file A and B might be different samples from the same population. One must then combine the two files in such a way that the distributions of the variables of interest (for instance, **X** and **Y**) remain as unchanged as possible. The resulting synthetic file is not appropriate for regression analysis, however, as the matching procedure will not necessarily reproduce the functional form of the conditional distribution of **Y** on **X** [27].

The basic assumptions of statistical matching are straightforward. We assume that **X** (observed only in the recipient file), **Y** (observed only in the donor file), and **Z** (common both in the recipient and donor files) are multivariate random variables with a joint probability or density function $\mathbf{f}_{\mathbf{xyz}}$, and that no single file has information on **X**, **Y**, and **Z** together. Also we assume that the records in both files are drawn randomly and independently of each other from the same population. In other words, both samples to be matched are regarded as a single-source random sample from the underlying population. Combining the two files is only possible if the specific variables, **Y** and **X**, are conditionally independent given the common variables **Z = z**. This criterion is called the *Conditional Independence Assumption* (CIA).

(1) Constrained Statistical Matching

In practice, statistical matching techniques break down into two broad categories: unconstrained statistical matching (USM) and constrained statistical matching (CSM). USM uses a distance function to find the nearest neighbor in the donor file for each record in the recipient file [19]. This procedure allows individual donor records to be selected multiple times or not at all, which can lead to very different empirical marginal distributions of **Y** or empirical conditional distributions of **Y** given **Z** in the statistically matched file compared with those in the original donor file. Thus, USM is not appropriate for our application, although USM has been more widely used.

CSM, contrary to USM, requires that the weights (and records) in each file be fully used according to the following constraints in which file A has *n* and file B has *m* records [21]:

$$\sum_{j=1}^{m} w_{ij} = w_i, \quad for \quad i = 1\, to\, n \tag{1}$$

and

$$\sum_{i=1}^{n} w_{ij} = w_j, \text{for j} = 1 \text{ to m.} \tag{2}$$

In equations (1) and (2), w_{ij} is the weight of the record in the synthetic file created by matching the *i*th record in file A with the *j*th record in file B. The advantage of this method is that all of the records in both files are represented in the matched synthetic file by using up the weights attached to each record. In other words, the empirical multivariate distribution (the marginal distribution, for instance) of the variables in the donor file is replicated in the statistically matched file. These constraints also imply that there may be more records in the synthetic file than in the recipient file, because recipient records may have to be split up to match more than one donor record. Further constraints are:

$$w_{ij} > 0, \quad \forall ij \tag{3}$$

and

$$d_{ij} \geq 0, \quad \forall ij \tag{4}$$

d_{ij} is the distance between the *i*th record in file A and the *j*th record in file B. In order to achieve this result, the weighted population totals in the donor and recipient files should be equalized and the distance between two matched records must be minimized. The latter condition can be stated:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} d_{ij} \tag{5}$$

The matching procedure should minimize the sum of the weighted distances between matched pairs in the synthetic file. Because matching in the CSM context is carried out without replacement, the distances between matched records in CSM will generally be larger on average than in the USM case [21].

Another feature of CSM is that records are matched according to their rank according to a constructed score, rather than the absolute values of Z or a distance measure itself. This is why CSM is frequently called an imputation on rank and why linear programming approaches have been employed in earlier applications of this method. The main disadvantage of CSM, however, is due to the nature of rank order matching: some matches may be made over large distances that are unacceptable or undesirable to researchers. Additional steps can be taken to minimize this problem.

(2) Matching Algorithm

One significant feature of many statistical matching algorithms is that they address the dimensionality problem involved in multivariate analysis by reducing the matching to one constructed variable (for instance, distance, predicted mean or propensity score as discussed below). This reduction is a very important advantage for our purpose because in our context a large number of differently weighted common variables should be considered in the search for nearest neighbor matches. Moreover, separate files may show different empirical distributions of the common variables due to the various sampling designs across files -- over-sampling special population groups or different sampling strata and clusters. Therefore, dimensionality reduction in statistical matching is not only an attractive but also a necessary feature.

Three approaches are available for dimensionality reduction: a distance function; predictive means; and propensity scoring. A distance function produces a geometric measure of distance between two observations. If we wish to use the Euclidean distance, for instance, the distance function is given by

$$d(z_i, z_j) = \sqrt{\sum_{k=1}^{K} g(z_k)(z_{ki} - z_{kj})^2} \qquad (6)$$

where $g(z_k)$ is an individual weight that allows us to give extra influence to covariates that we believe are more important, z_{ki} is the kth common variable in the recipient file (**A**), and z_{kj} is the kth common variable in the donor file (**B**). In the *predictive mean matching* (PMM) framework, the target variable is regressed on common variables in both files and the predicted value is used to rank records in each file. This algorithm is widely used and was adopted in the Urban-Brookings micro-simulation model, for example [22]. *Propensity score statistical matching* (PSSM) is often used in observational studies to generate suitable control groups that are similar to the treatment groups when a randomized experiment is not available [25]. In the PSSM framework, a propensity score is produced from a maximum likelihood estimation of a record being in the recipient file on the common variables. Gu and Rosenbaum [9] show that propensity score matching produces matched samples that are more balanced than the use of the Mahalanobis distance function or propensity score with a Mahalanobis caliper if there are many covariates and large imbalances in the covariates between data sets. Therefore, we adopt this approach in our construction of a matching algorithm, and give more elaboration on that below.

Assuming that the conditional independence assumption holds, the variables observed only in one file are conditionally independent from the assignment (**T**) to this file given the covariates $\mathbf{Z = z}$ (that is, $\mathbf{f_{X|T,Z} = f_{X|Z}}$ and $\mathbf{f_{Y|T,Z} = f_{Y|Z}}$), then we can say that the assignment of the records (**T**) to each file is *strongly ignorable* given the covariates $\mathbf{Z = z}$ (i.e., randomization). Rosenbaum and Rubin [23] prove that if the assignment (**T**) is strongly ignorable given $\mathbf{Z = z}$, then it is also strongly ignorable given any balancing score **b(z)**, (that is, $\mathbf{f_{X|T,b(z)} = f_{X|\,b(z)}}$ and $\mathbf{f_{Y|T,b(z)} = f_{Y|\,b(z)}}$). Here a balancing score **b(z)** is defined as a function **b** of the observed covariates **Z**. Following this logic, we can conclude that the distributions of the covariates for recipient and donor files are also identical if the balancing scores in both files are identical. In this regard, matching based on identical common variables can be regarded as an extreme type of PSSM, using **Z** itself as a balancing score (**b(z)=z**). Various types of balancing scores can be constructed and the propensity score is one of them [23].

There are several requirements in constructing a matching algorithm. First, based on the conditional independence assumption, the two separate files (donor and recipient) should not differ significantly in terms of the common variables. This difference is minimized by harmonization, which will be discussed later. Second, to account for the different scales of the common variables, it is recommended to standardize continuous and even ordinal variables to a mean of zero and a standard error of one [20]. Third, the algorithm may use all or some of the common variables to match each recipient record with at least one donor record. A subjective weight for each common variable can be used to incorporate the subjective importance of each variable. Fourth, for some of the common variables (strata or cohort variables) a perfect match is required. This requirement is satisfied by the segmentation of the data and by confining the matching to within these segments. Finally, because of differences in sample weights, one donor record may be used for multiple recipient records. In order to limit the number of times an individual donor is used, a penalty can be placed on selected donor records using the distance function. But this restriction may lead to a loss in variability or sample size, so abandoning certain matches for a better match is inevitable.

3. Propensity Score Statistical Matching Procedure with Application to SCF 2001 and ADS 2002 Matching

In this section we provide a detailed application taken from our work on the Levy Institute Measure of Economic Wellbeing (LIMEW).[1] We use constrained statistical matching (CSM) based on estimated propensity scores to produce the synthetic data set from which the LIMEW is constructed. The matching algorithm uses propensity scores to rank observations within pre-specified segments and then matches records from the donor

[1] The Levy Institute Measure of Economic Wellbeing (LIMEW) is a comprehensive income measure. It includes elements such as earnings, income from wealth, household production and public consumption. The construction of the LIMEW data files requires the integration of many sources of information about households such as the Current Population Survey's Annual Demographic Supplement for household demographic and income data, the Survey of Consumer Finances for household wealth data, the American Time Use Survey for household production data, Income Tax Models for household tax data, and administrative data for public consumption. For more information, visit
http://www.levy.org/limew.aspx

data file to records in the recipient data file by rank. The working procedure is elaborated here.[2]

(1) Description of SCF and ADS files

The two data sets used in this application of statistical matching are the 2001 Survey of Consumer Finances (SCF) and the March 2002 Current Population Survey Annual Demographic Supplement (ADS). Both surveys are nationally representative and have been used by many researchers as major sources of information on wealth holdings (SCF) or income (ADS) of households, but have never been used together. This gap in the literature motivates us to combine these two data sets using statistical matching.

The SCF, a triennial survey carried out by the Federal Reserve Board, in addition to demographic information, includes great detail on the components of wealth such as bonds, stocks, money market accounts, certificates of deposit, mutual funds, checking and saving accounts, real estate and so forth. It also contains information on various types of individual debt, which allows us to calculate net worth at the level of the primary economic unit within each household. The data set contains records for 4,442 households and missing values have been multiply imputed so that there are 22,210 records in total. The sampling frame is also important to emphasize. Because the distribution of wealth is highly skewed, a simple random sample would under-represent those households with high wealth, yielding biased estimates of wealth in the U.S. [3]. In addition, a survey of

[2] We used STATA Special Edition, version 9.2 to produce the analysis presented here. The algorithm we use is our own, coded as a STATA ado file. Details about STATA can be found at http://stata.com.

this type is likely to suffer from the problems of nonrandom non-response, especially among those with high amounts of wealth. These problems, difficult to eliminate perfectly, are addressed by using a dual-sampling frame, in which higher wealth households are over-sampled using a wealth index and adjusted using aggregate data on household wealth [1], [13], [14], [34]. In this project, we treat the SCF file as a donor to transfer information on wealth to the ADS file as a recipient.

The ADS is an annual survey carried out by the Census Bureau to examine the labor market situation and it is the most widely used household survey data to extract information on income and demographics in the U.S. The data set has 78,200 household records in total after cleaning up some anomalies [29]. Compared to the SCF, the ADS has a fat tail at the lower part of the income distribution due to its original purpose of monitoring changes in the labor market. Also, since the survey unit in the SCF is the primary economic unit, with respondents not necessarily being the household head, and because race is asked only of the respondent we have to assume that the race of the head is the same as the race of the respondent, an assumption that gets harder to justify as time passes [15]. So during the matching, additional care needs to be taken for these underlying differences between the two data sets. We believe the Conditional Independence Assumption holds in this case, since both surveys are random samples independently drawn from the same population.

(2) Data Preparation and Harmonization

Preparation for PSSM (or statistical matching in general) involves much work on the separate files. We align the common variables in both files to each other in terms of definitions and measurement, so that at the very least the two files do not differ significantly in terms of the common variables. For example, the age variable in the SCF file has values between 18 and 95, while the corresponding variable in ADS file has values between 15 and 80 for householders. So we truncate the age variable at 18 and 80 in both files. Also, the occupation code in the SCF public-use file is not the 3-digit Census occupation code. It has been recoded to a 1-digit code. Thus, we convert the occupation code in the ADS to match the SCF code. Harmonization across the common variables (**Z**) in both files in this way makes the joint distributions of the common variables in each file be as close as possible to each other, maximizing the quality of the match.

We also check that the distributions of the common variables are comparable. Since the data sets we use are intended to be representative at the national level, we expect there to be very close correspondence between the two files in terms of the common variables. Exceptions to this rule are generally the result of nonexact correspondence between the actual records the two files have and this inevitably introduces error into the matching procedure due to mismatched samples.[3]

[3] Matching tax records with census data, aside from the question of different samples, provides a good example. Tax records include variables such as return type and marital status that are similar to but distinct from the information in the census (which never includes information on tax return type). Return type must then be assigned to the records in the census data, using assumptions that limit the categories that can be assigned. Married Filing Separately can never be adequately assigned, since there are no criteria appropriate to the task. O'Hara solves this problem by using a rule of thumb that simply assumes that married couples file joint returns [16].

(3) Weight Adjustments and Segmentation

After harmonization, we adjust the sum of the attached weights for records (weighted population totals) in the donor (SCF) file so that they are comparable with those in the recipient (ADS) file. Frequently, the recipient and donor files are not from the same year, which means that the sum of weights will be different due to population changes. We adjust weights by multiplying weights in the donor file by the ratio of the sum of weights in the recipient file to the sum of the weights in the donor file. This transformation allows all donor records to be matched to recipient records by splitting their weights. This weight adjustment could cause the means and variances of the variables in the synthetic matched file to be different from those of the donor file, so we need to compare the distributions after the match.

We then separate the data within each file into several discrete segments. This segmentation is used either because matches between certain types of records should be avoided or because matches between certain types should be required or both. In our application family type, elder status, race, homeownership, and household income are selected as strata variables and the combination of these lead to 120 discrete cells in each file. This choice is made because differences between these subpopulations are the main interest of our research. When strata variables are defined and segmentation is done accordingly, propensity scores can be estimated separately or unique propensity scores can be constructed for different cells.

Second, segmentation with balanced weights is desirable: the weighted counts of observations within cells should be balanced as much as they can be between the two files. Table 2 shows the distribution of weighted observations by cell and source file. As can be seen, there is inexact correspondence between cells, even though the surveys were done only one year apart. For example, in the ADS white, married-couple elderly homeowners number 7.83 million, while in the SCF this cell has 9.77 million members. The differences are due to differences in the sampling frame as well as the slight shift in demographics we could expect to happen in one year. Because of the imbalances, collapsing across cells in later stages will be required in order to exhaustively match the records in the two files.

(4) Propensity Score Estimation

To estimate propensity scores, we add the outcome variable (**T=1**) for all records in the recipient file and the outcome variable (**T=0**) to donor file, and join the files by stacking the records. The selection of the specific common variables in the logistic regression model for propensity scores should be made carefully to maximize explanatory power. This is because the validity of PSSM relies on the power of the common variables to act as good predictors that can be transformed into effective propensity scores. In the SCF-ADS match, we use gender, age category, education category, race, and occupation of the household head, as well as homeownership, family types, household size, household income, existence of property income, existence of self-employed income, existence of transfer income, and (adjusted gross) household income to estimate the propensity score.

Specifically, logistic regression models are run with several variations with the dependent variable (**T**) and the selected common variables (**Z**) as independent variables. First, an overall model is estimated with all the selected common variables as independent variables to get an overall propensity score. After that, different models with respect to the included independent variables are constructed within different cells, created by different combinations of strata variables, to estimate cell specific propensity scores. In order to get a tighter fit in matching (with respect to income class, for instance), additional segmentation is done. That is, subcells within each cell are constructed and estimations of the propensity score are carried out after screening out the subcells where no propensity scores can be estimated. So we need to estimate one overall model, cell specific models, and subcell specific models here.

The propensity score is defined as:

$$e(z_i) = P(T = 1 \mid Z = z_i) = g(z_i^{'}\beta), \qquad (7)$$

the conditional probability of a record *i* to belong to a certain group (**T = 1**) given the covariates (**Z = z**). The estimated propensity score is defined accordingly,

$$\hat{e}(z_i) = g(z_i^{'}\overset{\wedge}{\beta}) = \frac{1}{1+e^{-z_i^{'}\overset{\wedge}{\beta}}} \qquad (8)$$

The individual propensity scores $\hat{e}(z_i)$ are the predicted values from the logistic regression output for β. However, we use different subjective weights for each parameter to incorporate the subjective importance of the independent variables. For the subcell cases, however, subjective weights are not critical because the main variables of interest are already included as strata variables (although more elaboration can be added if necessary).

After running each model, all records for each file are sorted by estimated propensity score $\hat{e}(z_i)$ (in ascending order) and attached weight (in descending order). Under this sorting scheme, we assign records with larger weights in the donor file to multiple records in the recipient file until all of their weight has been used up. As we will see below, however, this will not exhaustively match all of the records in both files, requiring additional estimation of propensity scores by relaxing the restriction of perfect matches by strata variables.

(5) Statistical Matching Algorithm

The matching procedure begins with the separation of the combined file back into donor and recipient files according to their original membership. Then matching is performed in an iterative and hierarchical process: first, matching is done between records of the donor and recipient files by subcell, separately; second, the unmatched subcell leftover records are collapsed into the corresponding cells and matching is carried out within each cell separately; and third, the unmatched cell leftover records are collapsed and matching is carried out across combinations of subsets of the strata variables or their variants to use up the attached weights for each file.

An important point here is that it is inevitable to collapse cells across strata variables. Although almost 90% of weights are exhausted after second matching step (see Table 3 for the breakdown by round for this match), we still need to sacrifice the perfect-match requirement within cell in order to use up all of the weights in the data sets (the main restrictions of constrained matching). Therefore, several additional considerations are employed in matching procedure. First, searching for nearest neighbors is done through comparison of forward and backward search results, and splitting weights is followed with some buffering for flexibility.[4] Second, several records that have weights

[4] In our case, we regard weight differences of 100 between corresponding donor and recipient records as acceptable matches.

that are too small to be matched with corresponding records are combined as groups and then adjusted following their proportion to the within group total.[5]

4. Properties of the Statistical Match

Under the constrained matching scheme, all marginal distributions are supposed to be identical before and after matching. Only the joint distributions of variables not jointly observed may be different. Following this logic, statistical matching is regarded as successful if the marginal and joint empirical distributions of **X** given **Z** that are observed in the statistically matched file are nearly the same or similar to those of the donor file. This criterion is based on the assumption that discrepancies should not be large between two independent random samples drawn from the same population. Although there are other proposed tests to check the validity of statistical matching, comparing the marginal and joint distribution is the only available test in practice [20].

In this study, the empirical marginal distributions of the imputed variables Y in the resulting matched file are compared with their empirical marginal distributions of the donor file to evaluate the similarity of both files through the calculation of Lorenz coordinates, Gini coefficients, quantile values and their respective ratios. Also the weighted mean and median values for **Y** by each strata variable are computed and compared between the donor and matched files. The **Y** variables in our case are five classes of assets (value of primary and secondary residential housing, other nonfinancial assets, liquid assets, other financial assets, and retirement assets), two classes of debt

[5] This procedure provides alternative matched variables with some variations that can be compared with originally matched variables, and we can pick one of them at the quality check stage.

(mortgages and home equity lines of credit on primary and secondary residential housing, and other debt), and net worth (the sum of assets minus the sum of debts). Figure 1 shows the ratio of the average value in the matched file to the average value in the donor file for each of these variables. Each variable has two ratios; the first, “scaled” ratio reflects the adjustment made in the matching procedure for those observations that were dropped due to small weights, while the “unscaled” ratio refers to the unadjusted values. In all cases, the “unscaled” ratios are closer to unity, so we choose to incorporate these values into the final synthetic file, and for the rest of the discussion we will refer to the “unscaled” values only.

Figures 2 through 4 provide comparisons of the wealth variable in the original data set (**SCF2001**) and in the matched data set (**IMP1**). As we can see, the distribution of net worth in the matched data set is very close to that of the original data set. Figure 2 shows the Lorenz curves for the two distributions. They are, in fact, too similar for this level of detail to be very revealing. Figure 3 shows the distribution of wealth for each of eight cells, differentiated by race, homeownership, and age. The box plots give us confidence that the marginal distributions have been well preserved in the statistical matching process. The notable exception is that the upper tails of the distribution have not been transferred. This is because the records in the donor file from the upper tail have small weights, and so are frequently collapsed with neighbors in the matching procedure. Finally, Figure 4 shows the density functions of wealth for the imputed and original data sets. Again, they appear to be identical at this level of detail.

While the preceding analysis sheds some light on the similarities between the imputed and original data sets, closer examination of the marginal distributions for all of the variables is required for complete confidence in the results. Figure 5 and Table 4 provide a detailed comparison of the empirical marginal distribution of variables in the matched file to those in the donor data set by the strata variables we identified above: race, age, family type, homeownership, and income class. Figure 5 summarizes the ratios of the average net worth in the imputed data set to the source data set for each category of our strata variables plus education. The best results are for the cases of race, age, and homeownership. The family type and income class ratios vary a bit more, but are mostly close to unity. In Table 4, the comparison is of mean and median values of net worth.

In Table 4 we can see that the means in the imputed data set are, for the most part, quite close to those in the source data set. We can see though that the gap between white and nonwhites is understated in the matched data set as compared to the donor data set. This pattern is attributable to the fact that the matching does not perfectly capture the upper tail of the distribution of wealth in the SCF (for the reasons mentioned above).

The case of income categories reveals an interesting skew in the results: wealth is less unequally distributed along the income distribution in the synthetic data set than in the SCF. However, it is important not to overstate the significance of this pattern. For those households with less than $20,000 income, the average net worth in the synthetic data set is fifteen percent higher than in the SCF. However, this amounts to a little under $10,000 in additional wealth (compare this to the absolute difference for elder households of $22,000 less wealth on average in the matched data set than in the SCF). For the most part, the average values of all the variables are quite similar in the matched data set to their corresponding values in the SCF for all income classes.

Table 4 shows close correspondence between the imputed and the donor data sets by race and homeownership status, with nonwhite renters’ net worth lower and both nonwhite and white owners' net worth larger on average in the synthetic data set than in the original. Some of the largest examples of divergence in the synthetic data set are between white and non-white, married couple and female-headed households. All have their net worth inflated in the matched data set, on average, while nonwhite male-headed households have theirs understated. The proportions between female-headed and married couple households are well preserved (for example, the ratio of nonwhite, female-headed average net worth to married couple average net worth is 0.223 in the matched data set, compared to 0.208 in the SCF), while the same is not as true for white, male-headed households (the ratio is 0.578 in the matched data set, compared to 0.490 in the SCF).

In summary, this application of statistical matching has resulted in a synthetic data set that preserves pretty well the marginal empirical distribution of the wealth variables in the donor data set. Some variation is observed, due for the most part to differences in the sample frames between the two data sets.

5. Conclusions

Statistical matching is an attractive procedure for empirical exercise. The data required to answer even basic questions is often not available in one survey data set, and carrying out additional comprehensive survey requires a great amount of costs and time. Especially when it comes to the situation that one is interested in past rather than future, the ability to combine sets of data can be seductive. However, care must be taken whenever two sets of data are combined in this manner. If the assumption of conditional independence is violated, the resulting analysis will be compromised, because the joint distribution of the variables in the synthetic data set will be substantially different from that of the target population.

In cases where the assumption of conditional independence seems appropriate, as in our example, matching can proceed with the confidence that the synthetic data set produced adequately captures the relationship between the variables of interest that are not jointly observed in any of the previously available data sets. Checking the quality of the match is essential in this sense. But if there exists no third source of data against which to check the validity of the synthetic data set, all that is available in terms of quality control is comparison of the conditional distributions of the donated variables in the donor and synthetic data sets. We acknowledge, of course, that this is a necessary but insufficient indicator of the quality of the match.

A problem that has yet to be adequately addressed is posed by the fact of having to use weighted observations (in this type of application). Generally speaking, if the weights on some observations in the donor or recipient data set are very much smaller than the typical weight in the other data set (as in the case of the SCF, in which high-wealth households are oversampled in order to be adequately represented in the completed survey), what can be done to best incorporate this information into the resulting synthetic data set? The box plots in Figure 3 illustrate the effect this problem has. The upper tail of the wealth distribution is attenuated in the process of matching. This may or may not be a severe problem, depending on the application and research purpose. However, if we have reason to believe that significant information about wealth inequality is being discarded in the process of statistical matching, then this problem deserves further attention.

Acknowledgments: Our thanks go to Edward N. Wolff and Ajit Zacharias for their thoughtful comments on earlier versions of this paper, Dimitri Papadimitriou, for his support and Elizabeth Dunn for her editorial assistance.

References

[1] A.M. Aizcorbe, A.B. Kennickell and K.B. Moore, Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances, *Federal Reserve Bulletin* **89** (2003), 1-32.

[2] H. Alter, Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970, *Annals of Economic and Social Measurement* **3** (1974): 373-394.

[3] R.B. Avery, G.E. Elliehausen and A.B. Kennickell, Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finances, *Review of Income and Wealth* **34** (1988), 339-69.

[4] T. Brodaty, B. Crépon and D. Fougère, Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986-1988, in: *Econometric evaluation of labour market policies*, M. Lechner and F. Pfeiffer, eds., Physica, New York, 2001, pp. 85-123.

[5] Congressional Budget Office, Effective Federal Tax Rates, 1997 to 2000, Washington D.C., August, 2003

[6] M. D'Orazio, M. Di Zio and M. Scanu, *Statistical matching: theory and practice,* Wiley, Chichester, England, Hoboken, NJ, 2006.

[7] D.T. Greenwood, Age, Income, and Household Size: Their Relation to Wealth Distribution in the United States, in *International comparisons of the distribution of household wealth*, E.N. Wolff, ed., Oxford University Press, Clarendon Press, Oxford, New York, Toronto and Melbourne, 1987, pp. 121-40,.

[8] D.T. Greenwood, An Estimation of U.S. Family Wealth and Its Distribution from Microdata, 1973, *Review of Income and Wealth* **29** (1983), 23-44.

[9] X.S. Gu and P. R. Rosenbaum, Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms, *Journal of Computational and Graphical Statistics*, **2** (1993), 405-420.

[10] L.A. Keister, *Wealth in America: Trends in wealth inequality*, Cambridge University Press, Cambridge, New York and Melbourne, 2000.

[11] L.A. Keister, Sharing the Wealth: The Effect of Siblings on Adults' Wealth Ownership *Demography* **40** (2003), 521-542.

[12] L.A. Keister and S. Moller, Wealth Inequality in the United States, *Annual Review of Sociology* **26** (2000): 63-81.

[13] A.B. Kennickell, Codebook For 2001 Survey Of Consumer Finances, Unpublished Manuscript, Board of Governors of the Federal Reserve System, Washington, DC, 2003.

[14] A.B. Kennickell, Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances, Unpublished Manuscript, Board of Governors of the Federal Reserve System, Washington, DC, 2001.

[15] S. Lindamood, S.D. Hanna, and L. Bi, Using the Survey of Consumer Finances: Some Methodological Considerations and Issues, Journal of Consumer Affairs 41 (2007), 195-222.

[16] R.J. Little and D.B. Rubin, Causal Effects In Clinical And Epidemiological Studies Via Potential Outcomes: Concepts And Analytical Approaches, *Annual Review of Public Health* **21** (2000), 121-45.

[17] A. O’Hara, New Methods for Simulating CPS Taxes, Unpublished Manuscript, U.S. Census Bureau, Washington, DC, 2004).

[18] B.A. Okner, Constructing a New Data Base from Merging Microdatasets: The 1966 Merge File, *Annals of Economic and Social Measurement* **1** (1972): 325-341.

[19] D.B. Radner, An Example of the Use of Statistical Matching in the Estimation and Analysis of the Size Distribution of Income, *Review of Income and Wealth* **27** (1981), 211-42.

[20] S. Rässler*, Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches,* Springer, New York: 2002.

[21] W.L. Rodgers, An Evaluation of Statistical Matching, *Journal of Business & Economic Statistics* **2** (1984), 91-102.

[22] J. Rohaly, A. Carasso and M.A. Saleem, The Urban-Brookings Tax Policy Center Microsimulation Model: Documentation and Methodology for Version 0304, Tax Policy Center, Washington, DC, January 10, 2005). (http://taxpolicycenter.org/publications/template.cfm?PubID=9168)

[23] P.R. Rosenbaum and D.B. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika* **70** (1983), 41-55.

[24] D.B. Rubin and N. Thomas, Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions, *Biometrika* **79** (1992), 797-809.

[25] D.B. Rubin and N. Thomas, Matching Using Estimated Propensity Scores: Relating Theory to Practice, *Biometrics* **52** (1996), 249-264.

[26] R. Ruggles and N. Ruggles, A Strategy for Matching and Merging Microdatasets, Annals of Social and Economic Measurement **3** (1977), 353-371.

[27] C.A. Sims, Data Matching and Merging: Comment, *Annals of Economic and Social Measurement* **3** (1974), 395-397.

[28] H. Sutherland, R. Taylor and J. Gomulka, Combining household income and expenditure data in policy simulations, *DAE Working Papers*, Number MU0101, Cambridge Working Papers in Economics, January 2001.

[29] U.S. Census Bureau, *Annual Demographic Supplement to the March 2002 Current Population Survey*, Washington, D.C., 2002.

[30] J. Wagner, The causal effects of exports on firm size and labor productivity: First evidence from a matching approach, *Hamburgisches Welt-Wirtschafts-Archiv Discussion Paper* no. 155, Hamburg Institute of International Economics, Hamburg, Germany, 2001.

[31] T.R. Williams, Flexible Matching Imputation: Combining Hot-deck Imputation with Model-based Methodology. In *Proceedings of the Survey Research Methods Section,* American Statistical Association Conference, 2001.

[32] E.N. Wolff and A. Zacharias, The Levy Institute Measure of Economic Wellbeing, *Eastern Economics Journal*, **33** (2007), 443-470.

[33] E.N. Wolff, Recent Trends in Wealth Ownership, 1983-1998, Levy Economics Institute, Working Paper Series, no. 300, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY, 2000.

[34] A. Yamokoski and L.A. Keister, The Wealth of Single Women: Marital Status and Parenthood in the Asset Accumulation Of Young Baby Boomers in the United States, *Feminist Economics* **12** (2006), 167-194.

Table 1. Demographic Characteristics of the SCF and ADS Files

Homeownership	ADS2002	SCF2001
renter	31.9%	32.3%
owner	68.1%	67.7%

Family Type	ADS2002	SCF2001
MC	55.9%	60.3%
FH	27.9%	26.1%
MH	16.2%	13.6%

Elder	ADS2002	SCF2001
nonelder	79.4%	78.9%
elderly	20.6%	21.1%

Race Category	ADS2002	SCF2001
nonwhite	26.1%	23.8%
white	73.9%	76.2%

HH Income Class	ADS2002	SCF2001
lt $20k	22.5%	25.3%
$20k - $50k	33.8%	34.1%
$50-$75k	17.9%	16.9%
$75k -$100k	11.1%	9.6%
gt $100k	14.7%	14.1%

Key: MC = married couple, FH = Female-Headed, MH = Male-Headed; nonelder = less than 65 years old, elder = 65 or older; nonwhite includes the categories black, asian and other.

Table 2. Comparison of ADS and SCF in Weighted Frequency by Cell

ADS	nonelder		elderly	
white	renter	owner	renter	owner
MC	6,288,450	32,560,596	540,057	7,829,036
FH	5,876,830	6,877,571	2,006,559	6,011,014
MH	5,073,342	5,216,404	670,799	1,867,837
nonwhite				
MC	4,970,318	7,565,049	243,752	1,125,922
FH	5,442,342	2,646,253	601,130	1,021,687
MH	2,937,753	1,366,267	246,631	311,856

SCF	nonelder		elderly	
white	renter	owner	renter	owner
MC	7,881,774	34,561,013	1,024,688	9,772,397
FH	5,580,812	6,485,042	2,063,104	4,464,706
MH	4,666,490	4,257,696	339,664	2,216,620
nonwhite				
MC	4,858,145	6,412,491	308,909	1,099,905
FH	5,733,592	2,835,691	756,155	585,731
MH	1,849,371	1,097,773	253,118	192,351

Key: MC = married couple, FH = Female-Headed, MH = Male-Headed; nonelder = less than 65 years old, elder = 65 or older; nonwhite includes the categories black, asian and other.

Table 3. Weighted Distribution of Matched Records by Matching Round

Round	Freq.	Percent
1	98,225,759	89.87
2	3,827,915	3.5
3	301,461	0.28
4	277,796	0.25
5	1,178,157	1.08
6	2,219,868	2.03
7	1,347,519	1.23
8	638,536	0.58
9	34,745	0.03
10	777,852	0.71
11	105,151	0.1
12	362,696	0.33
Total	109,297,455	100

Table 4. Ratios of Mean and Median Wealth by Category

	Average Net Worth			Median Net Worth		
Category	SCF2001	ADS2002i	Ratio	SCF2001	ADS2002i	Ratio
nonwhite	104,892	112,402	107%	7,730	8,200	106%
white	463,056	477,302	103%	104,700	107,400	103%
nonelder	329,856	342,590	104%	51,700	51,900	100%
elder	557,444	535,312	96%	150,000	143,800	96%
lt $20K	68,059	77,936	115%	7,350	10,660	145%
$20K-$50K	151,293	149,962	99%	37,880	37,880	100%
$50-$75K	245,396	242,652	99%	97,500	87,000	89%
$75K-$100K	343,112	372,074	108%	186,430	168,880	91%
gt $100K	1,665,257	1,555,749	93%	503,300	447,360	89%
nonwhite renter	14,473	13,539	94%	-	-	
nonwhite owner	206,668	214,117	104%	67,460	65,000	96%
white renter	68,690	68,577	100%	1,000	1,100	110%
white owner	600,710	615,813	103%	176,200	178,750	101%
nonwhite MC	167,484	176,741	106%	25,200	31,550	125%
nonwhite FH	34,753	39,380	113%	500	400	80%
nonwhite MH	75,867	74,254	98%	5,130	3,750	73%
white MC	596,898	624,574	105%	153,080	160,550	105%
white FH	185,037	214,348	116%	52,300	63,110	121%
white MH	292,650	361,007	123%	52,000	59,500	114%

Key: MC = married couple, FH = Female-Headed, MH = Male-Headed; nonelder = less than 65 years old, elder = 65 or older; nonwhite includes the categories black, asian and other.

Fig. 1 Ratio of Imputed to SCF Values, Unscaled and Scaled

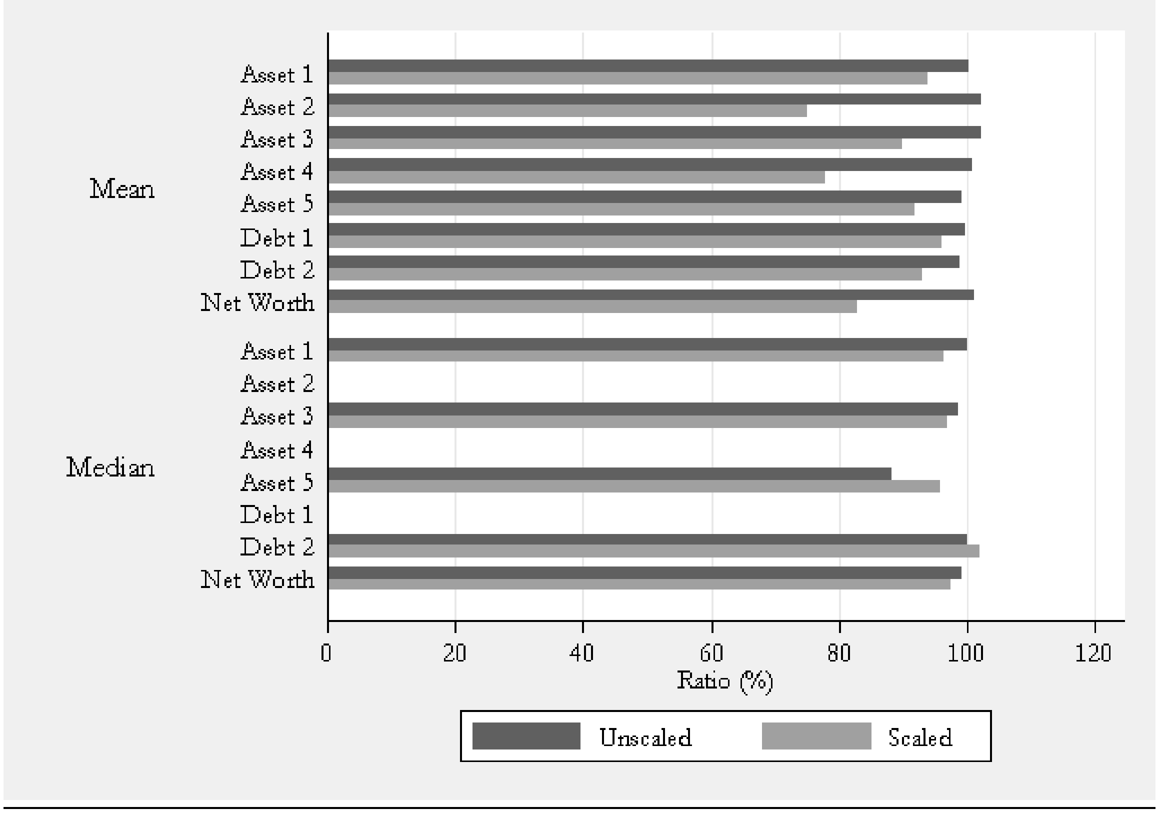


Fig. 2 Lorenz Curve of Imputed and SCF Net Worth

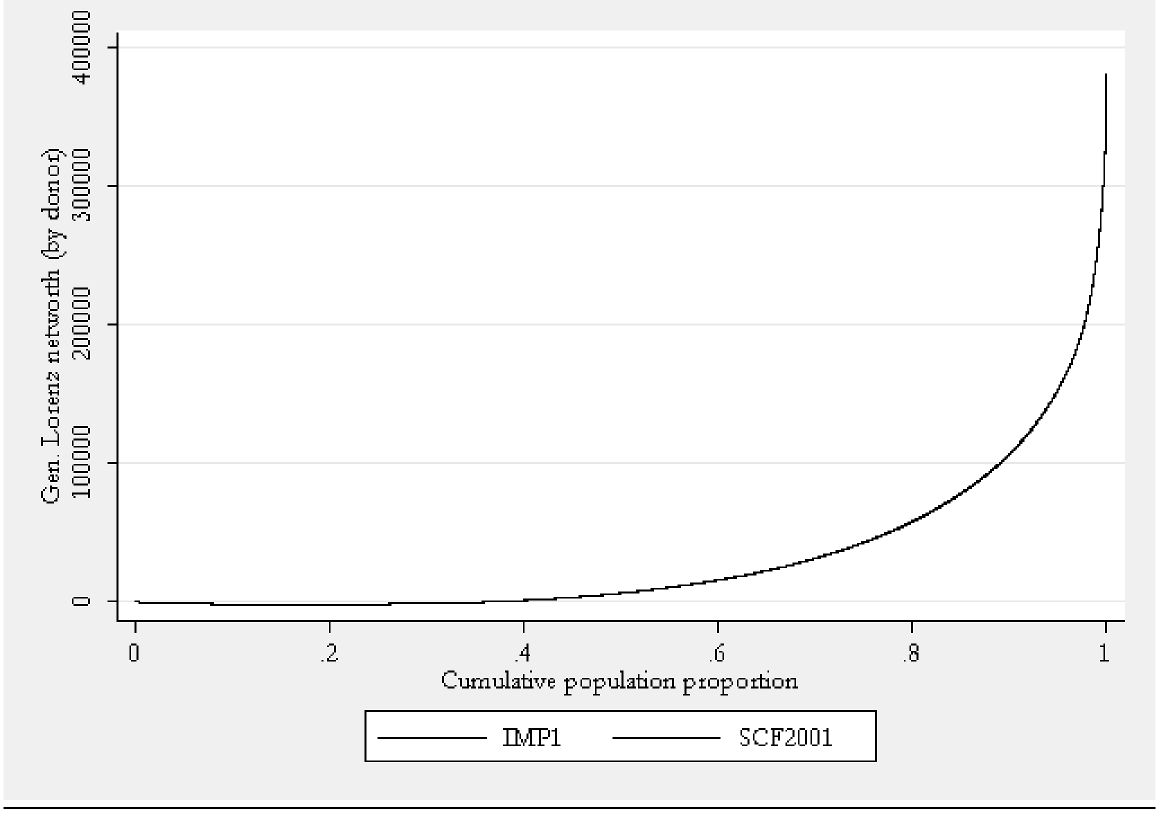


Fig. 3 Distribution of Net Worth (log) by Race, Home Ownership and Age

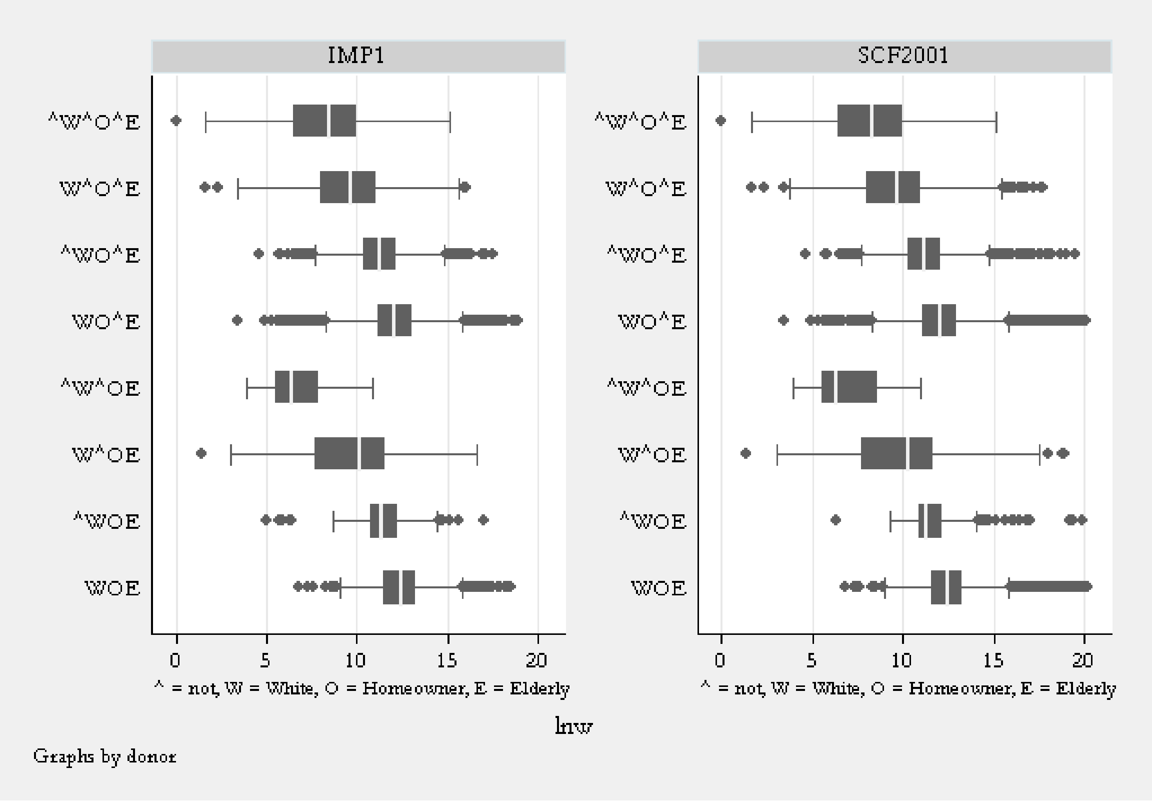


Fig. 4 Distribution of Net Worth (log) in Imputed and SCF Datasets

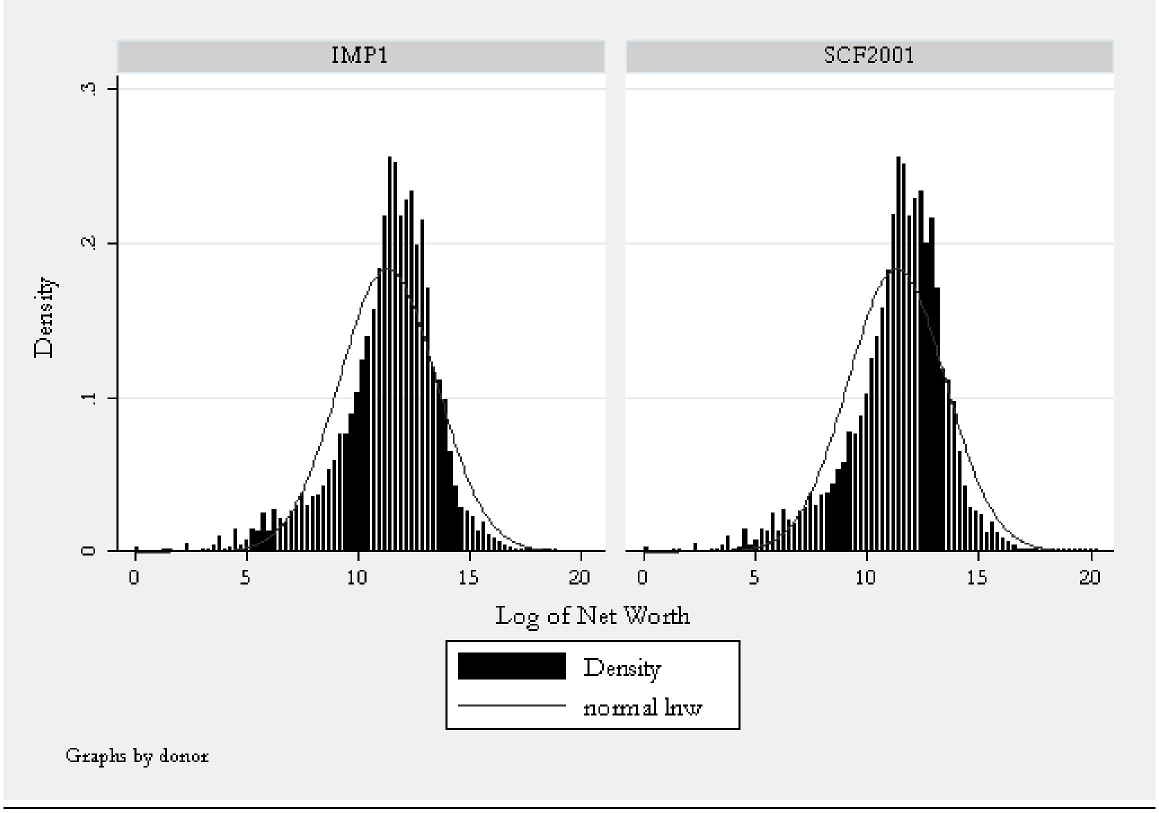
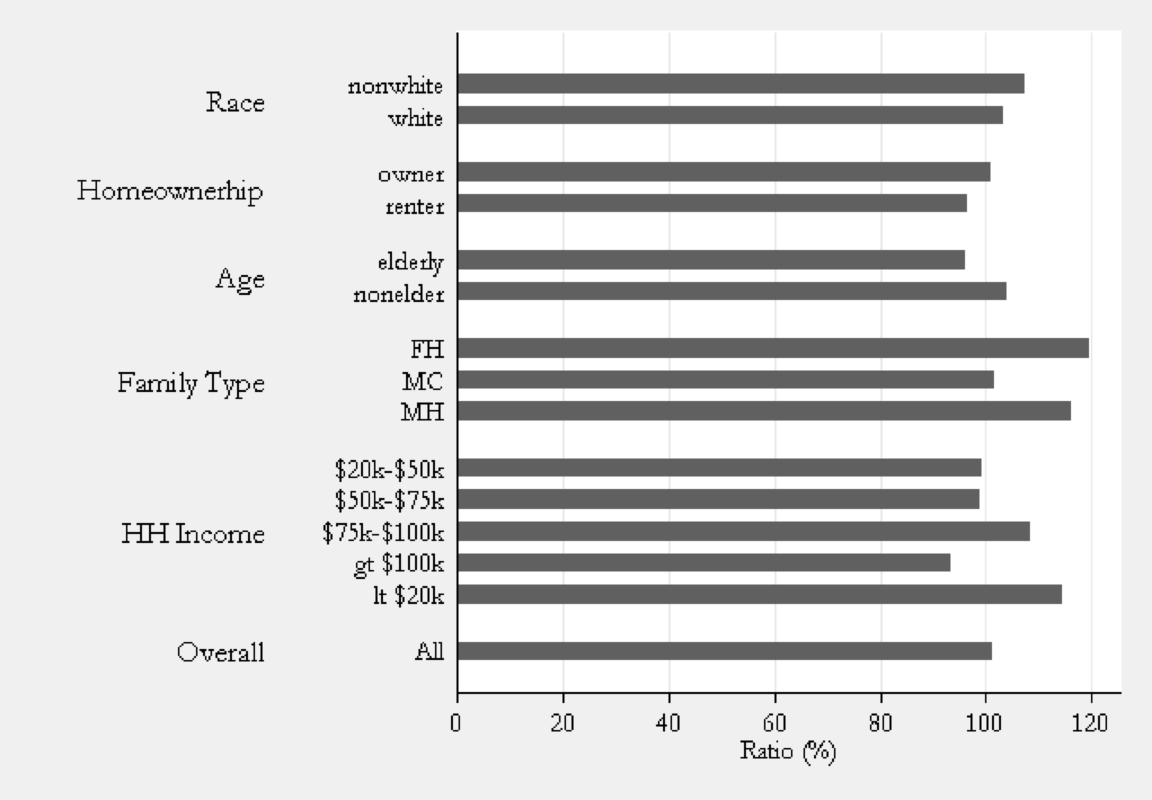


Fig. 5 Ratio of Mean Net Worth in Imputed File to SCF, by Category


Race
nonwhite
white
Homeownerhip
owner
renter
Age
elderly
nonelder
Family Type
FH
MC
MH
HH Income
$20k-$50k
$50k-$75k
$75k-$100k
gt $100k
lt $20k
Overall
All
Ratio (%)