Kum, Hyunsub; Masterson, Thomas

**Working Paper**

# Statistical matching using propensity scores: theory and application to the Levy Institute measure of economic well-being

Working Paper, No. 535

**Provided in Cooperation with:**
Levy Economics Institute of Bard College

*Suggested Citation:* Kum, Hyunsub; Masterson, Thomas (2008) : Statistical matching using propensity scores: theory and application to the Levy Institute measure of economic well-being, Working Paper, No. 535, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY

This Version is available at:
https://hdl.handle.net/10419/31492


Working Paper No. 535

# Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being

by

**Hyunsub Kum**
**Thomas Masterson**
The Levy Economics Institute of Bard College

May 2008



The Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levy.org

**ABSTRACT**

This paper summarizes the background, type, logic, and working procedure of the statistical matching used in the Levy Institute Measure of Economic Well-Being (LIMEW) project to combine the various data sets used to produce the synthetic data set with which the LIMEW is constructed. We use the match between the 2001 Survey of Consumer Finances and Annual Demographic Survey of Current Population Survey 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 Classifications:** C14; C40; D31

## 1. INTRODUCTION

Statistical matching is a technique used to link records in two separate data sets in cases when exact matching of individual records (record linkage) is not possible due to confidentiality restrictions on the data available. Statistical matching uses variables common to both data sets to identify similar records that can be linked in order to generate a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets. In particular, the associations between variables never jointly observed are often the main motivation for interest in such a complete, but synthetic, data set. This interest compels one to seek the best possible (in lieu of exact) match between records in the two data sets

The Levy Institute Measure of Economic Wellbeing (LIMEW) is an extended income measure meant to be as comprehensive as possible. Thus, it includes elements not often incorporated in to measures of economic wellbeing, such as money income, household production, and public consumption. It also treats wealth differently; instead of incorporating income from wealth directly, a household's net home and nonhome wealth are converted into imputed rental income and imputed annuity income. The construction of such a comprehensive and complex measure 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, because no single source of data has all of the information required for this undertaking.

Combining these sets of data together into a new single measure, however, requires some particular statistical matching procedure that must satisfy another concern, driven in part by the structure of the research project. That is, we want a measure of economic wellbeing that is representative at the level of the U.S. national population. This means that we need a matching procedure that preserves at least the marginal distributions of the variables of interest and this procedure will have to be able to handle the fact that the various sets of data are taken from surveys with varying sample designs and weighting schemes.

This paper presents an application of such a matching procedure on which the LIMVEW is constructed. The rest of the paper is arranged as follows. The second section presents a review of literature concerning statistical matching. The third section outlines the statistical matching procedure used in the generation of the LIMEW with the results of matching the 2001 Survey of Consumer Finances (SCF) and the 2002 Annual Demographic Survey (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 Europe) is by now a widely used technique in producing empirical studies. The method is used in many observational studies in medical literature (Little and Rubin 2000; Rubin and Thomas 1992, 1996; Rosenbaum and Rubin 1983). In addition to the numerous examples in the field of economics cited by Rässler (2002), there are studies by Radner (1981), Wolff (2000), Wolff and Zacharias (forthcoming), Greenwood (1983, 1987), Wagner (2001), Brodaty, Crépon, and Fougère (2001), Keister (2000, 2003), the Urban-Brookings Tax Microsimulation (Rohaly, Carasso, and Saleem 2005), and the 2003 Congressional Budget Office report on income tax burdens (CBO 2003).

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 file with variables **X**, **Y**, and **Z** together, but that kind of file is unavailable from a single source. One must then combine the two files in such a way that the distributions of the variables of interest remain as unchanged as possible.

Let's assume that file A (which exclusively has **X**) is the recipient file and file B (which exclusively has **Y**) is the donor file, because we want to transfer the information about **Y** in donor file B to the recipient file A. Combining these two files is usually carried out by using some distance function to assess the similarity between the records in each file. The distance function, constructed from the common variables **Z** in both files, is used to search for the most

similar donor record for each recipient record (so-called nearest neighbor matching). Then the variables of the donor (**Y**) are added to the recipient file leading to a new and complete (**X**, **Y**, **Z**), but synthetic statistically matched, file. Once the data are matched, the analysis proceeds as if the artificial matched file were a real sample representative of the true population of interest.

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 $\mathbf{f_{xyz}}$, 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).

### A. 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 (Radner 1981). In this case, the main criterion is similarity between records in donor and recipient files, and the matching is performed by imputing the nearest possible record among the closest records. As a result, this procedure allows multiple selections or no selection (unmatched) of donor records, 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. Due to these problems, it is not appropriate for us to employ USM in our statistical matching task although USM has been more widely used.

CSM requires that the weights (and records) in each file be fully used according to the following constraints (Rodgers 1984) in which file A has *n* and file B has *m* records:

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

and

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

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. In order to achieve this result, the distance between two matched records must be minimized and the weighted population totals should be equalized between the donor and recipient files. These are necessary conditions for CSM.

In order to equalize weighted population totals in both files, a weight-split or duplication procedure across donor and recipient records should be undertaken during the matching procedure until the weights in both files are exhausted. Note that because matching in the CSM context is carried out without replacement, the distances between matched records in the CSM will generally be larger on average than in the USM case (Rodgers 1984).

In CSM, records are matched according to their rank 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. Consequently, additional steps must be taken to minimize this problem.

## B. Matching Algorithms

### *i. Distance Functions*

Under the assumptions stated above, statistical matching can be initiated. The main task here is to search for a donor record $\mathbf{B_j(Y_j,Z_j)}$ whose observed values of the common variables $\mathbf{Z_j}$ are identical or closest to those $\mathbf{Z_i}$ of the recipient record $\mathbf{A_i(X_i,Z_i)}$. Usually this searching process is carried out using an algorithm based on nearest neighbor matching by calculating a distance function. If we wish to use the Euclidean distance, the distance function is given by

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

where $\boldsymbol{g(z_k)}$ is an individual weight that allows us to give extra influence to covariates that we believe are more important, $\boldsymbol{z_{ki}}$ is the $k^{\text{th}}$ common variable in the recipient file (**A**), and $\boldsymbol{z_{kj}}$ is the $k^{\text{th}}$ common variable in the donor file (**B**). Then for every record of the recipient file, distance measures using all donor records are calculated and selection of the donor record that has the smallest distance is made.[1]

There are several caveats in constructing such an 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 (Rässler 2002). Third, the algorithm of a distance function may use all or some of the common variables to find for each recipient record at least one donor record whose distance is minimal. A subjective weight for each common variable can be used during this step to incorporate the subjective importance of each variable. Fourth, for some of the common variables a perfect match is required. This is done by segmentation and restricting records matching within these segments. Finally, one donor record may be used for multiple recipient records, and in order to limit the number of times a donor is taken, a penalty can be placed on 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.

*ii. Predictive Mean Matching*

In the *predictive mean matching* (PMM) framework, the search for the nearest neighbor record is carried out using regression estimation rather than a specific distance function. The procedure is as follows. First, one or some of the **Y** variables in the donor file are regressed on the common variables **Z** and the predicted values of **Y** ($\boldsymbol{\hat{Y}_d}$) are obtained. Then, using the same **Z**

[1] The previous matching algorithm for the LIMEW project was based on this distance-minimization procedure.

variables in the recipient file and parameter estimates from the donor file, predicted values of **Y** ($\hat{\boldsymbol{Y}}_{\boldsymbol{\gamma}}$) are calculated in the recipient file. The records in each file are sorted in descending order by the predicted values of **Y** ($\hat{\boldsymbol{Y}}_{\boldsymbol{d}}$ in the donor file and $\hat{\boldsymbol{Y}}_{\boldsymbol{\gamma}}$ in the recipient file), and corresponding records from each file based on the predicted values of **Y** are matched. Note that usually the closest record to be matched is defined in terms of the most similar predicted value of Y in the USM context, but in the CSM context matching is done based on the rank order of the predicted values. Thus, segmentation with balanced weights is again critical here. The caveats that arise in distance minimizing matching also apply here. This algorithm is widely used and was adopted in the Urban-Brookings micro-simulation model (Rohaly, Carasso, and Saleem 2005), for example.

*iii. Propensity Score Statistical Matching*

*Propensity score statistical matching* (PSSM) is often used in observational studies to generate suitable control groups that are similar to treated groups when a randomized experiment is not available (Rubin and Thomas 1996). PSSM refers to a multivariate method used to construct control groups that have similar distributions on many covariates compared with treated groups. One significant feature of PSSM is that it reduces the dimensionality problem involved in multivariate analysis by reducing the matching to one constructed variable—the propensity score. 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—oversampling special population groups or different sampling strata and clusters. In this case, PSSM's dimensionality reduction is an attractive alternative.[2]

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 $\mathbf{Z} = \mathbf{z}$ (that is, $\mathbf{f}_{\mathbf{X}|\mathbf{T},\mathbf{Z}} = \mathbf{f}_{\mathbf{X}|\mathbf{Z}}$ and $\mathbf{f}_{\mathbf{Y}|\mathbf{T},\mathbf{Z}} = \mathbf{f}_{\mathbf{Y}|\mathbf{Z}}$), then we can say that the assignment of the

[2] The predicted mean matching algorithm also reduces the dimensionality of the match, but implements it in a different way.

records (**T**) to each file is strongly ignorable given the covariates **Z = z** (i.e., randomization). Rosenbaum and Rubin (1983) prove that if the assignment (**T**) is strongly ignorable given **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 (Rosenbaum and Rubin 1983). Gu and Rosenbaum (1993) 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 matching algorithm.

## 3. PROPENSITY SCORE STATISTICAL MATCHING PROCEDURE WITH APPLICATION TO SCF 2001 AND ADS 2002 MATCHING

To sum up, the statistical matching procedure used in the LIMEW project is constrained statistical matching (CSM) based on estimated propensity scores. The matching is desirable in a sense, since each file contains survey weights that make them representative of the population as a whole,[3] and we will use up the weights in each file during the matching. The matching algorithm uses propensity scores to rank observations within prespecified segments and then matches records from the donor data file to records in the recipient data file by rank. The working procedure is elaborated here.

[3] We use the set of all U.S. households as the population of this research.

### A. 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, 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 as well as demographic information, which allow us to calculate net worth values at the household level. 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 United States (Avery, Elliehausen, and Kennickell 1988). In addition, a survey of this type is likely to suffer from the problems of nonrandom nonresponse, especially among those with high amounts of wealth. These problems, hard to be eliminated perfectly, are addressed by using a dual-sampling frame, in which higher wealth households are oversampled using a ”wealth index” (Kennickell 2001, 2003) and adjusted using aggregate data on household wealth (Aizcorbe, Kennickell, and Moore 2003; Yamokoski and Keister 2006). 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 United States. The data set has 78,200 household records in total after cleaning up some anomalies (U.S. Census Bureau 2002). 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. So during the matching, additional care needs to be taken for these underlying differences between the two data sets.

## B. Data Preparation and Harmonization

Preparation for PSSM (or statistical matching in general) typically involves much work on the separate files. The common variables in both files have to be aligned to each other in terms of definitions and measurement, and their distributions should be made comparable so that at the very least the two files do not differ significantly by means of the common variables. This is necessary because statistical matching is founded on the assumption that the two separate files are randomly and independently drawn from the same population, although each file is in fact produced for different purposes and so based on different designs. Both the creation of the matching cells (or segments) and the calculations of the propensity score for ranking records within cells in the matching procedure employ these common variables as main criteria. In our case, for instance, the age variable in the SCF file has values between 18 and 95, while the corresponding variable in ADS file has values between 0 and 90. So we need to truncate the age variable at 18 and 90. 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 must similarly convert the occupation code in the ADS to match the SCF code. Harmonization across the common variables (**Z**) in both files in this way is required to make the joint distributions of the common variables in each file be as close as possible to each other.

Another concern is the similarity of the distributions of the common variables in the SCF and ADS files. 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.[4]

[4] 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. Thus, we follow rule of thumb (for example, those in O’Hara 2004) that simply assumes that married couples file joint returns.

**Table 1. Comparison of ADS and SCF file in Compositions**

| 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% |

After cleaning and harmonizing the files, 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 up.

### C. Weight Adjustments and Segmentation

After harmonization, we need to 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 expanding 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 with splitting of their weights. Although 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, in practice we find that they are very close to each other in our case.

The next step is to separate the data within each file into several discrete cells. 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. Analogously, to cluster analysis, the data are classified into N matching cells, which are identically defined for the donor and recipient files in a mutually exclusive and exhaustive way, and matching is allowed only within the same cell. Two caveats are important. First, this segmentation primarily depends on the purpose of the research or the researcher's subjective considerations. For instance, a female record should be matched only with another female record if the sex variable is critical in the research. In this case, the sex variable is regarded as a strata variable and perfect matches across files are expected. Of course, more variables and their combinations can be used in this way and this segmentation tends to narrow the distance (or variability) between records and allows for a tighter match. In the case of SCF and ADS matching, 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. In other words, the weighted counts of observations within cells should be balanced as much as they can be between the two files. Because matching is allowed only within the same cell, unbalanced segmentation will result in unused records in the matching procedure. In practice, it is hard to achieve perfectly

balanced weights in segmentation due to the various combinations of strata variables. Table 2 shows the distribution of weighted observations by cell and source file. As can be seen, there is not exact correspondence between cells, even though both surveys are from the same year. The differences are due to the differences in sampling frame. Thus, collapsing across cells in later stages will need to be done in order to exhaustively match the records in the two files.

**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 |

### D. Propensity Score Estimation

The selection of the specific common variables in the logit (or probit) model to estimate propensity scores should be made carefully to maximize the explanatory power. This is because the validity of PSSM relies heavily on the power of the common variables to act as good predictors that can be transformed into effective propensity scores. Of course, an important subset of the common variables will always be reserved to segment the data (as strata) based on the subjective concerns of the research at hand. In the SCF-ADS match, we use sex, homeownership, family types, age category, education category, race, household size, occupation, 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.

More specifically, logistic regression models are run with a dependent variable (**T**) and the selected common variables (**Z**) as independent variables with several variations. First, an overall model is estimated with all the selected common variables as independent variables to get an overall propensity score. After that, different logistic regression models with respect to the included independent variables are constructed for different cells, which are segmented by 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 run one overall model, cell specific models, and subcell specific models here. Note that neither **X** nor **Y** are used throughout the estimation and matching procedure, distinguishing this approach from the predictive mean matching algorithm.

The propensity score is defined as:

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

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,

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

The individual propensity scores $\hat{e}(z_i)$ are the predicted values from the logistic regression output for β. However, different subjective weights for each parameter can be used to incorporate the subjective importance of the independent variable. This subjective importance of common variables is determined according to the researcher's discretion or the explanatory power of the variables. For the subcell cases, however, subjective weights are not critical because the variables of interest are already included as strata variables (although more elaboration can be added).

After running each model, all records for each file are sorted by estimated propensity scores $\hat{e}(z_i)$ (in ascending order) and weights size (in descending order). Then identifiers for

each record by the level of estimation are assigned for convenience of matching. Under this sorting scheme, a record with a larger weight in the donor file will be split up or duplicated and matched with multiple records in the recipient file until all of its weight has been used up. Note that this is a rank order matching procedure. 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.

**E. 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 some strata variables or their variants to use up the attached weights for each file.

An important point is the order of collapsing cells across strata variables after the second matching step (at which point, typically almost 90% of the weights are exhausted; see Table 3 for the breakdown by round for this example). Our strata variables are family type, elderly, race, homeownership, and household income, corresponding to the subpopulations of interest to us. Thus, we need to select an order to sacrifice the perfect-match requirement on these strata variables in order to use up all of the weights and to preserve marginal distributions (the main restrictions of constrained matching).

**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 |

Experimenting with collapsing cells, however, shows that it is almost impossible to set a strict order beforehand. Instead, it is better to take an ad hoc approach, comparing the propensity scores and attached weights across cells after every matching step to figure out what variable should be discarded. That consideration requires another estimation of propensity scores. Three additional considerations in this algorithm should be noted here. First, searching for the nearest neighbors is done through comparison of forward and backward search results, and splitting weights is followed with some buffering for flexibility.[5] Second, several records that have weights 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.[6] Third, a final adjustment of the values of Y in the synthetic matched file is performed by comparing them to the values of Y in the donor file. This includes readjustment of the minimum or maximum values of Y in the synthetic matched file and unmatched or unused records in the donor or recipient files. Usually these cases stem from the fact that the attached weights are too small to be used in the matching procedure.

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

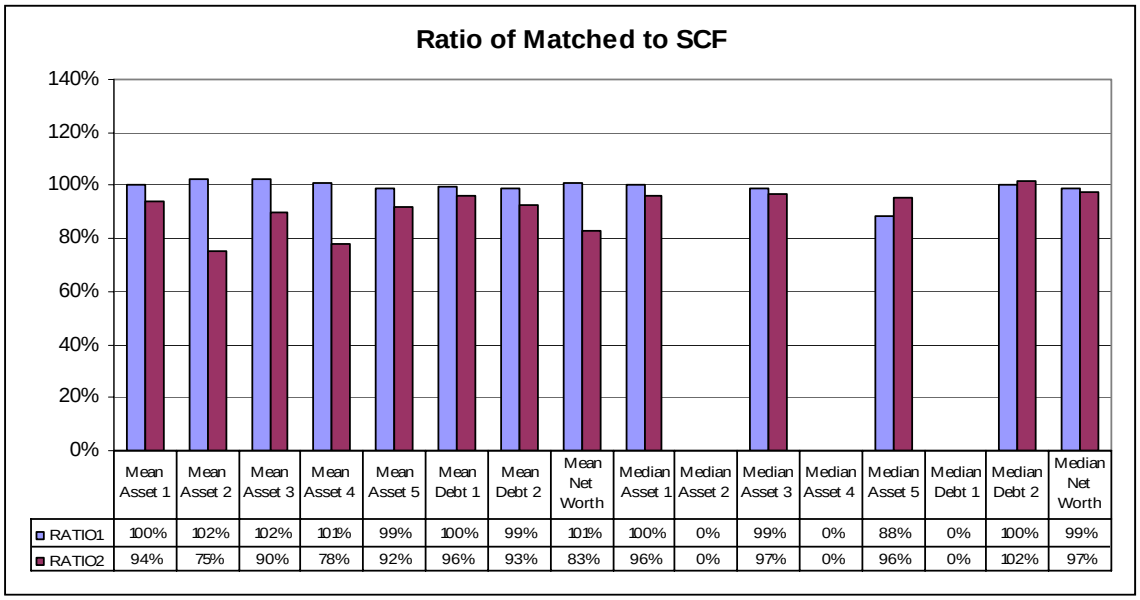
[6] This procedure provides alternative matched variables with some variations that can be compared with originally matched variables, and we can pick one of them at the quality check stage.

## 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 (Rässler 2002).

In this project, 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, decile values and their 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 (mortgages and home equity lines of credit on Asset 1, 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.

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




| | Mean Asset 1 | Mean Asset 2 | Mean Asset 3 | Mean Asset 4 | Mean Asset 5 | Mean Debt 1 | Mean Debt 2 | Mean Net Worth | Median Asset 1 | Median Asset 2 | Median Asset 3 | Median Asset 4 | Median Asset 5 | Median Debt 1 | Median Debt 2 | Median Net Worth |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RATIO1 | 100% | 102% | 102% | 101% | 99% | 100% | 99% | 101% | 100% | 0% | 99% | 0% | 88% | 0% | 100% | 99% |
| RATIO2 | 94% | 75% | 90% | 78% | 92% | 96% | 93% | 83% | 96% | 0% | 97% | 0% | 96% | 0% | 102% | 97% |

Figures 2 through 4 provide comparisons of the net worth 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 logged net worth 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. Finally, Figure 4 shows the density functions of logged net worth for the imputed and original data sets. Again, they appear to be identical at this level of detail.

**Figure 2. Lorenz Curve of Imputed and SCF Net Worth**

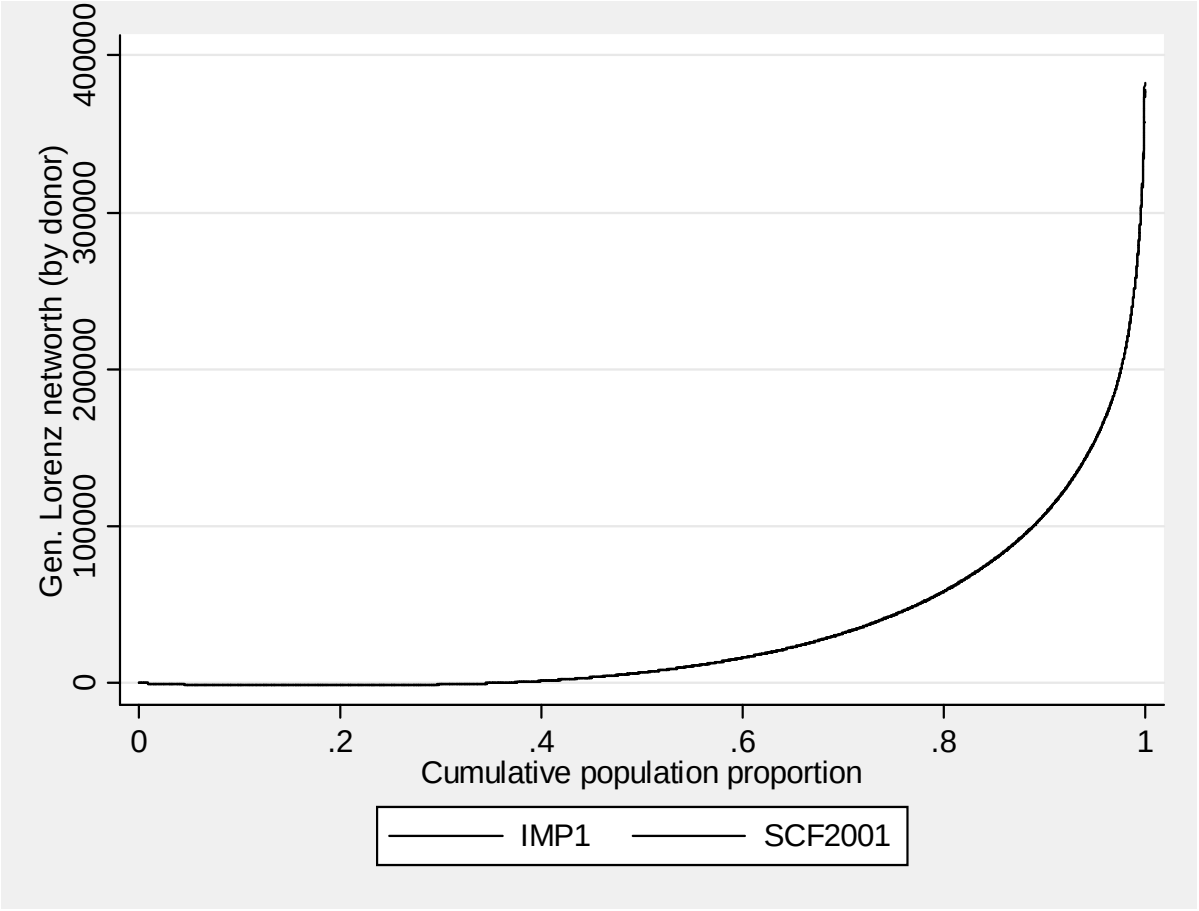


**Figure 3. Distribution of Net Worth by Race, Home Ownership and Age**

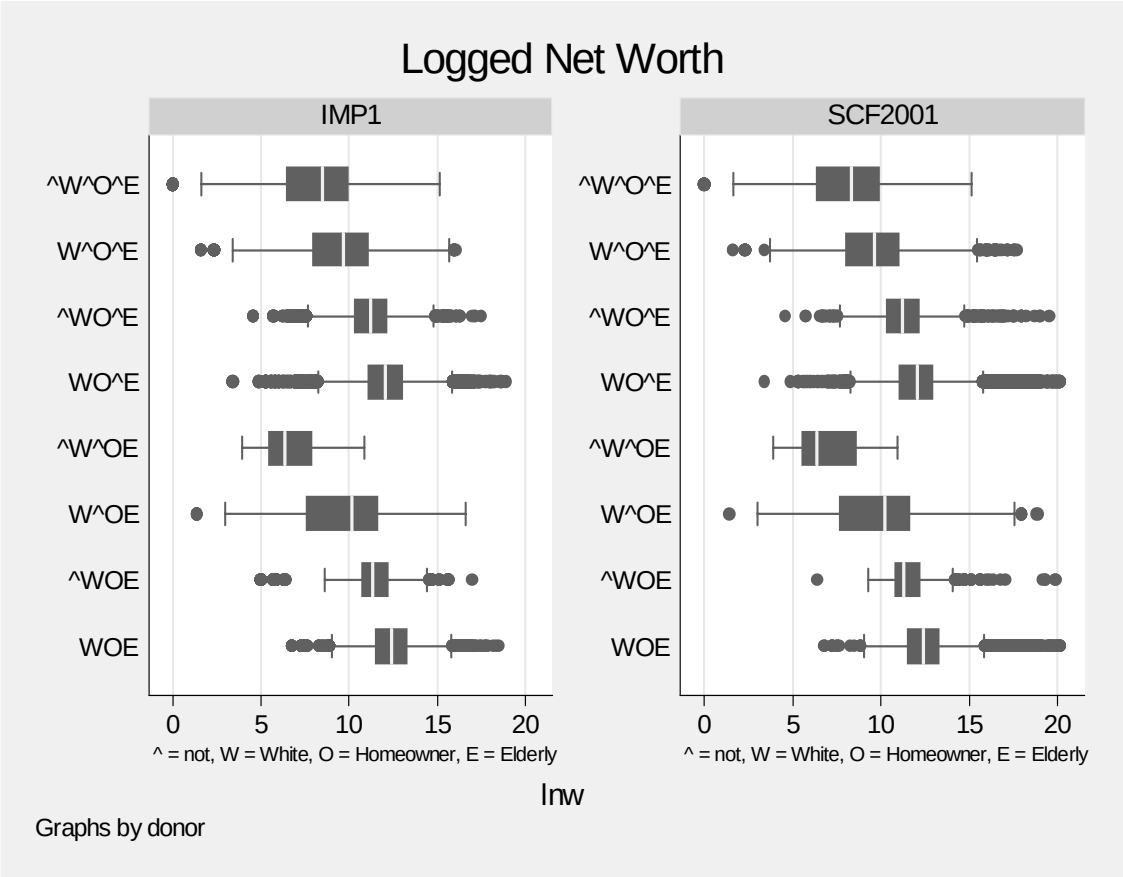
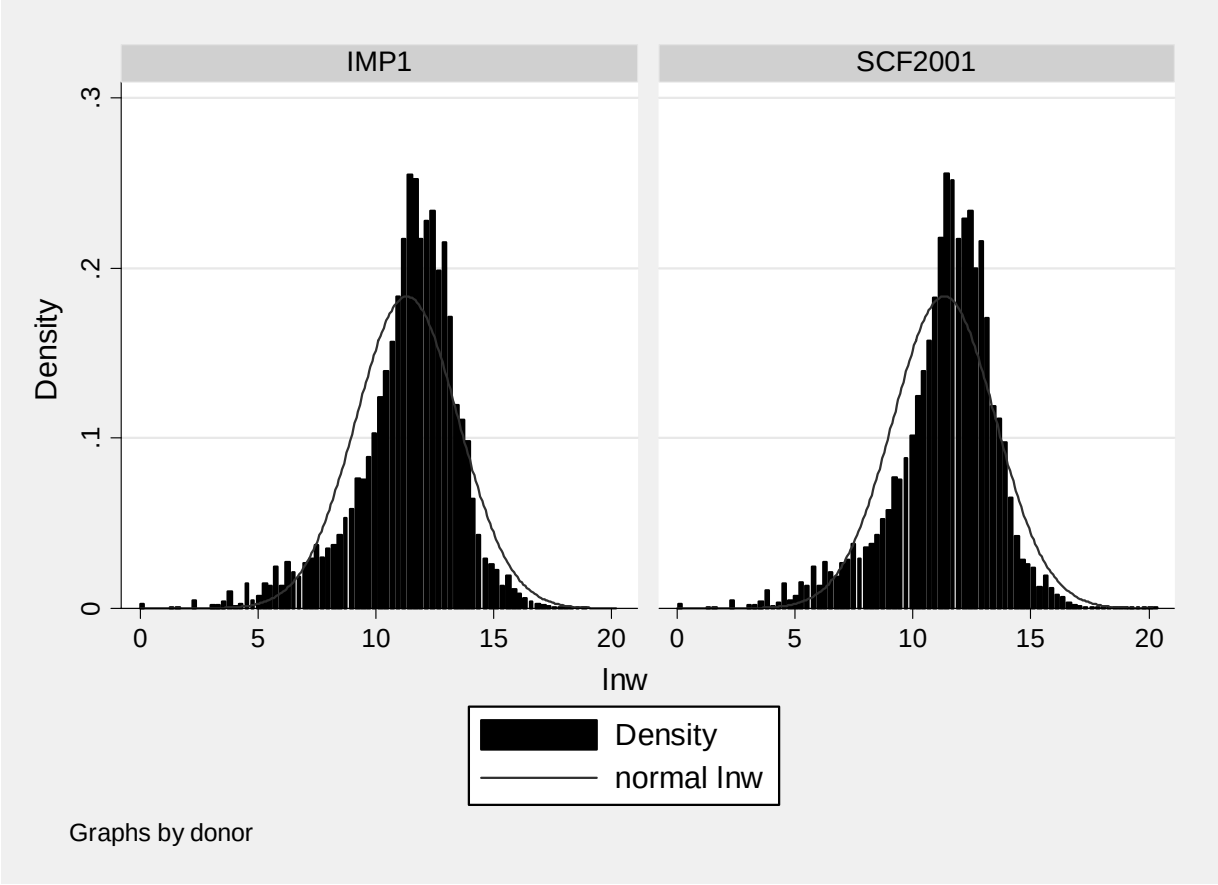


**Figure 4. Distribution of Net Worth in Imputed and SCF Datasets**



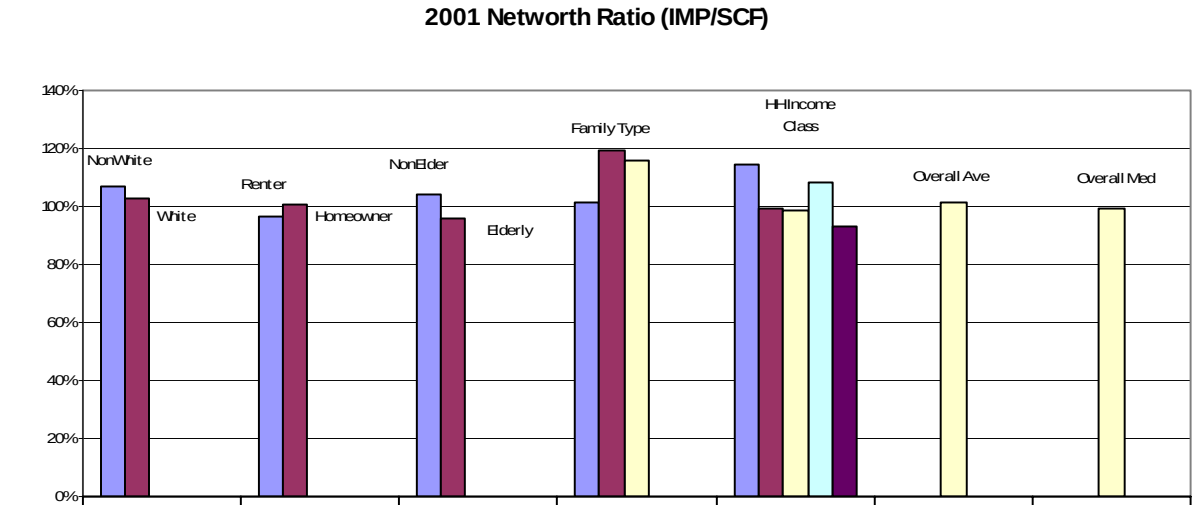
While the preceding analysis sheds some light on the similarities between the imputed and original data sets, closer examination of the marginal distributions for all of the variables is required for complete confidence in the results. Figure 5 and Tables 4 through 8 detail the comparison of the empirical 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 the tables, the comparison is of mean and median values of the eight variables we wish to carry over in the matching process.

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. In some cases, though, we can see that the gap between white and nonwhites is understated in the matched data set as compared to the donor data set. This phenomenon is most marked in the case of Asset 4 (financial assets), the variable that is in fact the most unequally distributed (note that its median value is zero in the SCF). 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 (as can be seen in the box-plot comparison of net worth in Figure 3).

Table 5 breaks down the distribution by elderly status. The ratios are within five percent of unity for all of the variables, with the exception of the average value of Asset 3 and the median value for Debt 1 (mortgages and HELOCs) for nonelders. Table 6 identifies the sources of the large differences between the source and imputed data set in the lowest income class as Asset 4 and 5. This case 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 7 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. Table 8 contains some of the largest examples of divergence in the synthetic data set. White and nonwhite married couples and female-headed households both 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).

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




| | white | homeown | elder | famtype | hhinccl | Overall Ave | Overall Med |
|---|---|---|---|---|---|---|---|
| cat1 | 107% | 96% | 104% | 102% | 115% | | |
| cat2 | 103% | 101% | 96% | 119% | 99% | | |
| cat3 | | | | 116% | 99% | 101% | 99% |
| cat4 | | | | | 108% | | |
| cat5 | | | | | 93% | | |

## Table 4. Ratios of Mean and Median Values by Race

| Average | Asset 1 | Asset 2 | Asset 3 | Asset 4 | Asset 5 | Debt 1 | Debt 2 | Net Worth |
|---|---|---|---|---|---|---|---|---|
| SCF2001 nonwhite | 61,310 | 35,636 | 14,393 | 13,944 | 15,313 | 26,029 | 9,676 | 104,892 |
| SCF2001 white | 141,796 | 148,254 | 45,661 | 125,203 | 62,547 | 45,682 | 14,723 | 463,056 |
| ADS2002i nonwhite | 63,033 | 39,494 | 14,749 | 15,592 | 15,473 | 26,142 | 9,797 | 112,402 |
| ADS2002i white | 144,101 | 153,848 | 47,585 | 129,101 | 63,300 | 46,098 | 14,610 | 477,302 |
| Ratio nonwhite | 102.81% | 110.82% | 102.47% | 111.81% | 101.05% | 100.44% | 101.25% | 107.16% |
| Ratio white | 101.63% | 103.77% | 104.21% | 103.11% | 101.20% | 100.91% | 99.23% | 103.08% |
| Median | | | | | | | | |
| SCF2001 nonwhite | - | - | 1,500 | - | - | - | 1,500 | 7,730 |
| SCF2001 white | 90,000 | - | 8,000 | 150 | 2,200 | - | 2,500 | 104,700 |
| ADS2002i nonwhite | - | - | 1,500 | - | - | - | 1,710 | 8,200 |
| ADS2002i white | 90,000 | - | 8,200 | 200 | 2,450 | - | 2,400 | 107,400 |
| Ratio nonwhite | | | 100.00% | | | | 114.00% | 106.08% |
| Ratio white | 100.00% | | 102.50% | 133.33% | 111.36% | | 96.00% | 102.58% |

## Table 5. Ratios of Mean and Median Values by Age

| Average | Asset 1 | Asset 2 | Asset 3 | Asset 4 | Asset 5 | Debt 1 | Debt 2 | Net Worth |
|---|---|---|---|---|---|---|---|---|
| SCF2001 nonelder | 117,767 | 113,906 | 32,348 | 79,077 | 50,801 | 48,502 | 15,542 | **329,856** |
| SCF2001 elder | 140,948 | 149,784 | 60,195 | 172,266 | 53,250 | 13,018 | 5,982 | **557,444** |
| ADS2002i nonelder | 119,963 | 119,800 | 34,229 | 81,663 | 50,714 | 48,372 | 15,408 | **342,590** |
| ADS2002i elder | 134,623 | 140,472 | 57,570 | 168,518 | 51,320 | 12,029 | 5,428 | **535,312** |
| Ratio nonelder | 102% | 105% | 106% | 103% | 100% | 100% | 99% | **104%** |
| Ratio elder | 96% | 94% | 96% | 98% | 96% | 92% | 91% | **96%** |
| Median | | | | | | | | |
| SCF2001 nonelder | 70,000 | - | 4,370 | - | 1,600 | 3,300 | 4,730 | **51,700** |
| SCF2001 elder | 90,000 | - | 17,310 | - | - | - | - | **150,000** |
| ADS2002i nonelder | 70,000 | - | 4,500 | - | 1,520 | 2,800 | 4,500 | **51,900** |
| ADS2002i elder | 90,000 | - | 16,500 | - | - | - | - | **143,800** |
| Ratio nonelder | 100% | | 103% | | 95% | 85% | 95% | **100%** |
| Ratio elder | 100% | | 95% | | | | | **96%** |

## Table 6. Ratios of Mean and Median Values by Household Income Class

| Average | | Asset 1 | Asset 2 | Asset 3 | Asset 4 | Asset 5 | Debt 1 | Debt 2 | Net Worth |
|---|---|---|---|---|---|---|---|---|---|
| SCF2001 | lt $20K | 42,572 | 12,712 | 9,380 | 9,291 | 4,992 | 6,795 | 4,093 | **68,059** |
| | $20K-$50K | 75,674 | 27,827 | 22,378 | 37,510 | 17,880 | 21,140 | 8,838 | **151,293** |
| | $50-$75K | 121,290 | 47,237 | 33,321 | 57,237 | 45,549 | 45,618 | 13,621 | **245,396** |
| | $75K-$100K | 165,834 | 87,097 | 40,765 | 73,277 | 69,161 | 74,550 | 18,472 | **343,112** |
| | gt $100K | 352,504 | 655,814 | 132,554 | 474,714 | 210,209 | 122,218 | 38,320 | **1,665,257** |
| ADS2001i | lt $20K | 45,980 | 13,335 | 10,131 | 12,914 | 6,239 | 6,525 | 4,383 | **77,936** |
| | $20K-$50K | 75,688 | 26,675 | 22,198 | 37,585 | 17,612 | 21,213 | 8,582 | **149,962** |
| | $50-$75K | 115,158 | 54,510 | 30,972 | 54,577 | 44,079 | 43,257 | 13,386 | **242,652** |
| | $75K-$100K | 161,385 | 114,291 | 40,945 | 77,182 | 63,575 | 67,493 | 17,812 | **372,074** |
| | gt $100K | 329,272 | 607,901 | 130,018 | 445,033 | 193,585 | 115,483 | 34,577 | **1,555,749** |
| Ratio | lt $20K | 108% | 105% | 108% | 139% | 125% | 96% | 107% | **115%** |
| | $20K-$50K | 100% | 96% | 99% | 100% | 99% | 100% | 97% | **99%** |
| | $50-$75K | 95% | 115% | 93% | 95% | 97% | 95% | 98% | **99%** |
| | $75K-$100K | 97% | 131% | 100% | 105% | 92% | 91% | 96% | **108%** |
| | gt $100K | 93% | 93% | 98% | 94% | 92% | 94% | 90% | **93%** |
| Median | | | | | | | | | |
| SCF2001 | lt $20K | - | - | 750 | - | - | - | - | **7,350** |
| | $20K-$50K | 50,000 | - | 3,900 | - | - | - | 2,500 | **37,880** |
| | $50-$75K | 98,000 | - | 8,800 | 500 | 10,700 | 28,000 | 6,700 | **97,500** |
| | $75K-$100K | 131,000 | - | 14,000 | 1,300 | 28,000 | 69,000 | 10,110 | **186,430** |
| | gt $100K | 245,000 | 30,000 | 35,000 | 30,000 | 80,000 | 95,000 | 7,600 | **503,300** |
| ADS2001i | lt $20K | - | - | 860 | - | - | - | - | **10,660** |
| | $20K-$50K | 50,000 | - | 3,840 | - | - | - | 2,100 | **37,880** |
| | $50-$75K | 90,000 | - | 7,600 | 300 | 8,500 | 19,000 | 6,400 | **87,000** |
| | $75K-$100K | 130,000 | - | 11,500 | 1,000 | 22,000 | 57,000 | 8,500 | **168,880** |
| | gt $100K | 225,000 | 15,000 | 32,500 | 22,700 | 67,000 | 85,000 | 6,400 | **447,360** |
| Ratio | lt $20K | | | 115% | | | | | **145%** |
| | $20K-$50K | 100% | | 98% | | | | 84% | **100%** |
| | $50-$75K | 92% | | 86% | 60% | 79% | 68% | 96% | **89%** |
| | $75K-$100K | 99% | | 82% | 77% | 79% | 83% | 84% | **91%** |
| | gt $100K | 92% | 50% | 93% | 76% | 84% | 89% | 84% | **89%** |

## Table 7. Ratios of Mean and Median Values by Race and Homeownership

| Average | | | Asset 1 | Asset 2 | Asset 3 | Asset 4 | Asset 5 | Debt 1 | Debt 2 | Net Worth |
|---|---|---|---|---|---|---|---|---|---|---|
| SCF2001 | nonwhite | renter | 0 | 4361 | 5452 | 6321 | 3662 | 0 | 5325 | **14473** |
| | nonwhite | owner | 130320 | 70839 | 24458 | 22524 | 28427 | 55327 | 14574 | **206668** |
| | white | renter | 0 | 27897 | 15296 | 24702 | 12316 | 0 | 11521 | **68690** |
| | white | owner | 191290 | 190265 | 56260 | 160283 | 80080 | 61627 | 15841 | **600710** |
| ADS2002i | nonwhite | renter | 0 | 3839 | 5580 | 6138 | 3656 | 0 | 5675 | **13539** |
| | nonwhite | owner | 127884 | 76177 | 24183 | 25317 | 27631 | 53038 | 14038 | **214117** |
| | white | renter | 0 | 27793 | 15308 | 23914 | 12451 | 0 | 11184 | **68577** |
| | white | owner | 192935 | 196567 | 58523 | 164747 | 80533 | 61720 | 15771 | **615813** |
| Ratio | nonwhite | renter | **#DIV/0!** | **92%** | **106%** | **96%** | **102%** | **#DIV/0!** | **101%** | **97%** |
| | nonwhite | owner | **102%** | **119%** | **104%** | **175%** | **108%** | **94%** | **100%** | **119%** |
| | white | renter | **#DIV/0!** | **105%** | **95%** | **98%** | **96%** | **#DIV/0!** | **103%** | **100%** |
| | white | owner | **99%** | **103%** | **97%** | **99%** | **100%** | **100%** | **98%** | **100%** |
| **Median** | | | **Asset 1** | **Asset 2** | **Asset 3** | **Asset 4** | **Asset 5** | **Debt 1** | **Debt 2** | **Net Worth** |
| SCF2001 | nonwhite | renter | 0 | 0 | 300 | 0 | 0 | 0 | 400 | **0** |
| | nonwhite | owner | 90000 | 0 | 4700 | 0 | 1200 | 40000 | 5000 | **67460** |
| | white | renter | 0 | 0 | 1580 | 0 | 0 | 0 | 2300 | **1000** |
| | white | owner | 130000 | 0 | 13350 | 1000 | 10000 | 32000 | 2680 | **176200** |
| ADS2002i | nonwhite | renter | 0 | 0 | 310 | 0 | 0 | 0 | 590 | **0** |
| | nonwhite | owner | 90000 | 0 | 4400 | 0 | 600 | 35000 | 5000 | **65000** |
| | white | renter | 0 | 0 | 1500 | 0 | 0 | 0 | 2100 | **1100** |
| | white | owner | 130000 | 0 | 13800 | 1100 | 10000 | 32000 | 2500 | **178750** |
| Ratio | nonwhite | renter | **#DIV/0!** | **#DIV/0!** | **100%** | **#DIV/0!** | **#DIV/0!** | **#DIV/0!** | **113%** | **#DIV/0!** |
| | nonwhite | owner | **100%** | **#DIV/0!** | **98%** | **#DIV/0!** | **50%** | **84%** | **85%** | **100%** |
| | white | renter | **#DIV/0!** | **#DIV/0!** | **97%** | **#DIV/0!** | **#DIV/0!** | **#DIV/0!** | **98%** | **106%** |
| | white | owner | **98%** | **#DIV/0!** | **100%** | **100%** | **95%** | **100%** | **89%** | **99%** |

**Table 8. Ratios of Mean and Median Values by Race and Family Type**

| Average | | | Asset 1 | Asset 2 | Asset 3 | Asset 4 | Asset 5 | Debt 1 | Debt 2 | Net Worth |
|---|---|---|---|---|---|---|---|---|---|---|
| SCF2001 | nonwhite | MC | 92217 | 63222 | 20217 | 17431 | 26129 | 39334 | 12398 | 167484 |
| | nonwhite | FH | 27735 | 5975 | 8156 | 6819 | 4643 | 12343 | 6230 | 34753 |
| | nonwhite | MH | 43885 | 19191 | 10852 | 21730 | 6061 | 16283 | 9569 | 75867 |
| | white | MC | 178753 | 201047 | 55347 | 155917 | 84197 | 60414 | 17948 | 596898 |
| | white | FH | 75074 | 34046 | 26393 | 55478 | 19521 | 17509 | 7966 | 185037 |
| | white | MH | 78473 | 88403 | 31950 | 95694 | 31831 | 22991 | 10711 | 292650 |
| ADS2002i | nonwhite | MC | 95890 | 66867 | 20463 | 20349 | 26405 | 40843 | 12391 | 176741 |
| | nonwhite | FH | 29079 | 6649 | 8423 | 8687 | 4778 | 11899 | 6337 | 39380 |
| | nonwhite | MH | 36886 | 26812 | 11044 | 15777 | 5572 | 12548 | 9291 | 74254 |
| | white | MC | 184109 | 211144 | 59166 | 164399 | 87941 | 63931 | 18254 | 624574 |
| | white | FH | 85506 | 41180 | 28814 | 61758 | 24395 | 19353 | 7952 | 214348 |
| | white | MH | 91717 | 125389 | 35354 | 108221 | 35601 | 23766 | 11976 | 361007 |
| Ratio | nonwhite | MC | 107% | 104% | 104% | 163% | 108% | 101% | 105% | 113% |
| | nonwhite | FH | 114% | 322% | 110% | 169% | 128% | 102% | 101% | 169% |
| | nonwhite | MH | 86% | 129% | 112% | 121% | 121% | 70% | 85% | 117% |
| | white | MC | 102% | 105% | 98% | 102% | 105% | 106% | 103% | 103% |
| | white | FH | 110% | 178% | 112% | 113% | 121% | 96% | 99% | 126% |
| | white | MH | 107% | 98% | 101% | 95% | 104% | 117% | 104% | 99% |
| **Median** | | | **Asset 1** | **Asset 2** | **Asset 3** | **Asset 4** | **Asset 5** | **Debt 1** | **Debt 2** | **Net Worth** |
| SCF2001 | nonwhite | MC | 50000 | 0 | 2400 | 0 | 0 | 0 | 2900 | 25200 |
| | nonwhite | FH | 0 | 0 | 650 | 0 | 0 | 0 | 600 | 500 |
| | nonwhite | MH | 0 | 0 | 1800 | 0 | 0 | 0 | 1190 | 5130 |
| | white | MC | 120000 | 0 | 11500 | 800 | 10000 | 23000 | 4600 | 153080 |
| | white | FH | 42000 | 0 | 3700 | 0 | 0 | 0 | 150 | 52300 |
| | white | MH | 28000 | 0 | 4000 | 0 | 0 | 0 | 1800 | 52000 |
| ADS2002i | nonwhite | MC | 60000 | 0 | 2500 | 0 | 0 | 0 | 3400 | 31550 |
| | nonwhite | FH | 0 | 0 | 600 | 0 | 0 | 0 | 750 | 400 |
| | nonwhite | MH | 0 | 0 | 1500 | 0 | 0 | 0 | 1450 | 3750 |
| | white | MC | 123000 | 0 | 12120 | 1100 | 11000 | 30000 | 5000 | 160550 |
| | white | FH | 53000 | 0 | 4500 | 0 | 0 | 0 | 100 | 63110 |
| | white | MH | 28000 | 0 | 4500 | 0 | 0 | 0 | 1600 | 59500 |
| Ratio | nonwhite | MC | 120% | #DIV/0! | 106% | #DIV/0! | #DIV/0! | #DIV/0! | 100% | 109% |
| | nonwhite | FH | #DIV/0! | #DIV/0! | 110% | #DIV/0! | #DIV/0! | #DIV/0! | 100% | 110% |
| | nonwhite | MH | #DIV/0! | #DIV/0! | 59% | #DIV/0! | #DIV/0! | #DIV/0! | 63% | 15% |
| | white | MC | 102% | #DIV/0! | 101% | 125% | 120% | 130% | 113% | 105% |
| | white | FH | 111% | #DIV/0! | 121% | #DIV/0! | #DIV/0! | #DIV/0! | 44% | 115% |
| | white | MH | 100% | #DIV/0! | 100% | #DIV/0! | #DIV/0! | #DIV/0! | 100% | 114% |

In summary, this application of statistical matching has resulted in a synthetic data set that preserves very 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 extremely attractive procedure for researchers. The data required to answer even basic questions is often not available in one survey data set. Thus, 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. An important qualification is that this is true only at the level at which the less representative of the two datasets is representative. In other words, if one of the data sets is representative at the state level and the other is representative at the national level, the resulting synthetic dataset can only claim to be nationally representative.

Checking the quality of the match is essential, of course, 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. This is a necessary but insufficient indicator of the quality of the match. However, if the *Conditional Independence Assumption* is met, we can be confident that the synthetic data set captures the distribution of the donated variables adequately.

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 effect this problem has is illustrated by the box plots in Figure 3. 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.

**REFERENCES**


Aizcorbe, Ana M., Arthur B. Kennickell, and Kevin B. Moore. 2003. "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances." *Federal Reserve Bulletin* 89(January): 1–32.

Avery, Robert B., Gregory E. Elliehausen, and Arthur B. Kennickell. 1988. "Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finances." *Review of Income and Wealth* 34(4): 339–69.

Brodaty, Thomas, Bruno Crépon, and Denis Fougère. 2001. "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986–1988." in *Econometric Evaluation of Labour Market Policies*, Michael Lechner and Friedhelm Pfeiffer (eds.). Heidelberg: Physica-Verlag.

Congressional Budget Office (CBO). 2003. "Effective Federal Tax Rates, 1997 to 2000." August. Washington D.C.

D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu. 2006. *Statistical Matching: Theory and Practice.* Chichester, England, and Hoboken, NJ: Wiley.

Greenwood, Daphne T. 1987. "Age, Income, and Household Size: Their Relation to Wealth Distribution in the United States." in *International Comparisons of the Distribution of Household Wealth*, Edward N. Wolff (ed.). Oxford, New York, Toronto, and Melbourne: Oxford University Press, Clarendon Press.

————. 1983. "An Estimation of U.S. Family Wealth and Its Distribution from Microdata, 1973." *Review of Income and Wealth* 29(1): 23–44.

Gu, Xing Sam, and Paul R. Rosenbaum. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *Journal of Computational and Graphical Statistics* 2(4): 405–420.

Keister, Lisa A. 2000. *Wealth in America: Trends in Wealth Inequality*. Cambridge, New York, and Melbourne: Cambridge University Press.

————. 2003. “Sharing the Wealth: The Effect of Siblings on Adults’ Wealth Ownership.” *Demography* 40(3): 521–542.

Keister, Lisa A., and Stephanie Moller. 2000. “Wealth Inequality in the United States.” *Annual Review of Sociology* 26: 63–81.

Kennickell, Arthur B. 2003. ”Codebook For 2001 Survey Of Consumer Finances.” Unpublished Manuscript. Washington, DC: Board of Governors of the Federal Reserve System.

————. 2001. “Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances.” Unpublished Manuscript. Washington, DC: Board of Governors of the Federal Reserve System.

Little, Roderick J., and Donald B. Rubin. 2000. “Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches.” *Annual Review of Public Health* 21(1): 121–45.

O’Hara, Amy. 2004. “New Methods for Simulating CPS Taxes.” Unpublished Manuscript. Washington, DC: U.S. Census Bureau.

Radner, Daniel B. 1981. “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(3): 211–42.

Rässler, Susanne. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

Rodgers, Willard L. 1984. “An Evaluation of Statistical Matching.” *Journal of Business & Economic Statistics* 2(January): 91–102.

Rohaly, Jeffrey, Adam Carasso, and Mohammed Adeel Saleem. 2005. “The Urban-Brookings Tax Policy Center Microsimulation Model: Documentation and Methodology for Version 0304.” January 10. Washington, DC: Tax Policy Center. Available at: http://taxpolicycenter.org/publications/template.cfm?PubID=9168

Rosenbaum, Paul R., and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70(April): 41–55.

Rubin, Donald B., and Neal Thomas. 1996. “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” *Biometrics* 52(March): 249–264.

————. 1992. “Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions.” *Biometrika* 79(4): 797–809.

Sutherland, Holly, Rebecca Taylor, and Joanna Gomulka. 2001. “Combining Household Income and Expenditure Data in Policy Simulations.” Cambridge Working Papers in Economics No. MU0110. Cambridge, UK: University of Cambridge.

U.S. Census Bureau. 2002. *Annual Demographic Supplement to the March 2002 Current Population Survey*. Washington, DC: United States Census Bureau.

Wagner, Joachim. 2001. “The Causal Effects of Exports on Firm Size and Labor Productivity: First Evidence from a Matching Approach.” Hamburgisches Welt-Wirtschafts-Archiv Discussion Paper 155. Hamburg, Germany: Hamburg Institute of International Economics.

Wolff, Edward, and Ajit Zacharias. Forthcoming. “The Levy Institute Measure of Economic Wellbeing.” *Eastern Economics Journal.*

Wolff, Edward. 2000. “Recent Trends in Wealth Ownership, 1983–1998.” Working Paper 300. Annandale-on-Hudson, NY: The Levy Economics Institute of Bard College.

Yamokoski, Alexis, and Lisa A. Keister. 2006. “The Wealth of Single Women: Marital Status and Parenthood in the Asset Accumulation of Young Baby Boomers in the United States.” *Feminist Economics* 12(1–2): 167–194.