Statistical Matching: A New Validation Case Study

Deborah D. Ingram, National Center for Health Statistics, John O'Hare and Fritz Scheuren, Urban Institute, and Joan Turek, Office of the Assistant Secretary for Planning Evaluation, Health and Human Services
Joan Turek <jturek@osaspe.dhhs.gov>

Key Words: constrained and unconstrained matching, conditional independence, total survey error

Abstract: This paper employs recently available data from the National Survey of American Families (NSAF) to assess the effect of violations of the conditional independence assumption on associations of health-related measures and income measures observed in a data set constructed by statistically matching the National Health Interview Survey (NHIS) and the Current Population Survey (CPS). This assessment is possible because NSAF includes health questions from the NHIS and income questions from the CPS. Contingency tables constructed from NSAF are compared with similar tables from the statistically-matched CPS/NHIS data sets. Chi-square tests of independence applied both to the NSAF and CPS/NHIS contingency tables assess the extent to which associations between health and income variables were preserved in the statistically matched file. Implications for the use of the statistically matched data sets are explored and conjectures about the importance of failures in the conditional independence assumption in practical microsimulation modeling settings are made.

1. Introduction

Statistical matching is one way information from two or more data sets can be combined to allow for the types of analyses that would be impossible from one input data source alone. In many cases, statistical matching relies on the strong assumption of conditional independence –i.e., given the values of variables common to both data sets, variables found only on the first of the two data sets are independent of variables found only on the second data set. A violation of this assumption is difficult to detect in practice because corroborative evidence from alternative data sources generally does not exist.

This paper assesses the quality of a statistical match between the March 1996 Current Population Survey (CPS) and the 1995 National Health Interview Survey (NHIS). The CPS is conducted by the Census Bureau for the Bureau of Labor Statistics. The NHIS is conducted by the National Center for Health Statistics.

To carry out the validation study, we employ recently available data from the 1997 National Survey of American Families (NSAF) to directly address conditional independence between health-related measures from the NHIS and income measures from the CPS. This is possible because NSAF contains both health-related measures and income measures. Multidimensional contingency tables constructed from NSAF are compared with similar tables from the statistically-matched CPS-NHIS data set. Statistical tests applied to cross-product ratios allow for the direct testing of conditional independence.

Organizationally, the present paper is divided up into 6 sections, beginning with this introduction. Section 2 provides an overview of statistical matching. In section 3, the basic design of the statistical match that was carried out between the NHIS and CPS is described. Section 4 provides a brief summary description of the three data sources being used. We assume the reader is familiar with the NHIS and CPS and employ this assumed familiarity to describe NSAF which draws many of its questions from these two surveys. Section 5 looks at the evaluations done so far and the results obtained on the extent to which associations are preserved in the statistically matched data set. Section 6 has a few concluding comments and conjectures, plus suggestions for future study.

2. Overview of Statistical Matching

Statistical matching involves combining information from two or more data files to construct one file containing information that is not available on any one file by itself. Usually, the input data sources are microdata files that contain information on individuals, families or firms. Statistically matched data sets have been used extensively in microsimulation modeling (e.g., Cohen 1991) to examine the impact of policy changes on population subgroups.

In the standard statistical matching framework (e.g., Moriarity and Scheuren 2000a and b), one has observations from two data sets, File A and File B, on a set of common variables (**X**-variables). File A contains variables (**Y**-variables) not available on File B and File B contains variables (**Z**-variables) not available on File A. Statistical matching involves constructing a new data set (File C) with information on **X**, **Y** and **Z** on each record. Generally, one file (File A) is considered the primary file and is referred to as the Host file. The other file (File B) is referred to as the Donor file.

2.1 Types of matches. There are two distinct types of statistical matching methods: ***unconstrained*** matching and ***constrained*** matching:

(1) In ***unconstrained*** matching (e.g., Rodgers 1984), each record in the Host file (File A) appears in the matched file (File C), but it is not required that all of the records in the Donor file (File B) be used in the match. Limits are usually placed on the number of

times a Donor record can be used from File B. These limits are imposed to ensure that the (weighted) distributions of the Z-variables "brought over" to the Host file in the match are closely aligned with the distributions on the original file. Even so, one of the criticisms of unconstrained matching methods is that the marginal distributions of the Z-variables in the matched file can be quite different than on the original file (Moriarity and Scheuren 2000b).

(2) In ***constrained*** matching (e.g., Barr and Turner, 1979, 1980), all of the records in ***both*** data files are represented in the matched file (File C). To accomplish this, records on ***both*** files may be used more than once. Limits on the number of times records can be used involve making sure that the population weight attached to each record is "used-up" in the match. A necessary condition for performing a constrained match, but one that is difficult to meet in practice, is that both input files have the same weighted population totals. One drawback of constrained matching is that records may end up being matched with an unacceptably large distance between the **X**-variables. When constrained matching is used, the weighted sample means and variances of the **X, Y** and **Z** variables in both input files are preserved as a direct consequence of the constraints imposed on the weights occurring in the final matched data set.

Unconstrained matching has been the most popular method because it is intuitive, relatively simple to implement, cost-effective, easy to replicate and update, and it makes fewer demands on system resources. However, Paass (1985) and Rodgers (1984) favor constrained matching because they believe that the risk of a poor match is lower with constrained matching and this outweighs the higher cost. With the advent of more powerful computers, cheaper memory and faster numerical algorithms, constrained matching has become more common.

2.2 Choosing a "Close" Match. Regardless of whether unconstrained or constrained matching is used, one seeks to match records on the two files that resemble or are, in some sense **close** to each other. Most statistical matches use a distance metric to assess how "close" two records are, and then match the two records separated by the smallest distance as measured by that metric (e.g., Armstrong 1989; Okner 1972). The **X**-variables, those variables common to both files, are involved in the distance function. The set of variables used to construct the distance function has important consequences for the integrity of the matched data set (Paass 1985).

The most commonly used distance metric is the Euclidean distance function. Another nearest-neighbor technique is predictive mean matching (e.g., Little 1989; Rubin 1986). In predictive mean matching, regressions are performed on the Donor file by regressing some of the **X**-variables on one or more **Z**-variables, predicted **Z**-values are obtained for both the Donor and Host files, and records from the Donor file are matched with records from the Host file using these predicted values.

While the **X**-variables in the regression can be thought of as playing the same sort of role as the variables in a classic distance function, predictive mean matching differs from the usual distance metric approach because it also involves **Z**-variables. Predictive mean matching has performed quite well in practice (e.g., Armstrong 1989; O'Hare 1997).

2.3 Partitioning. Partitioning, or blocking, is a technique whereby matches between records in different categories are not allowed. Whether one is using constrained or unconstrained matching, it is clear that matches between certain types of records should be avoided because the characteristics of the individuals are sufficiently dissimilar. In such cases, matches are only allowed for records that agree exactly on certain variables. For example, in the present context where the **Z**-variables include measures of health status and health care utilization, it would probably be unwise to allow a match between a man and a woman.

Partitioning has the effect of narrowing the distance between records and allows for a "tighter" fit across the two data sets. Hypothetically, partitioning can be achieved using a distance function that assigns an "infinite" distance, but the usual way to achieve partitioning is to divide the data files into mutually exclusive and exhaustive cells or equivalence classes and perform the match only within these cells. The variables used for the partitioning must, of course, come from the set that is common to both files (the **X**-variables).

2.4 Criticisms and more Comprehensive Views. A criticism of statistical matching is that it relies on strong assumptions about the relationship between the **Y**- and **Z**-variables (e.g., Kadane 1978). In particular, statistical matching assumes that the **Y**- and **Z**-variables are independent (or uncorrelated if normality is assumed), given an observation of **X**-variables.

The conditional independence assumption is often violated in practice and has caused researchers to investigate alternative methods of combining data sets (e.g., Armstrong 1989; Rubin 1986; Singh et. al. 1993). Beyond theoretical considerations, experience suggests that statistical matching works best when the input data sets are similar with respect to sample size, the population of interest, the time period over which the surveys are

taken, sample stratification, weighting and the type of questions that are asked.

An early but still comprehensive review of theoretical and applied approaches to statistical matching is contained in Sims (1978). Several papers provide good descriptions of how statistical matching performs in applied work, notably Armstrong (1989), Rodgers (1984) and Singh et. al. (1993). Cohen (1991) provides a good survey of the issues and techniques, as related to the important role statistical matching plays in microsimulation modeling. Paass (1985) emphasizes how the use of auxiliary information can improve the predictions from microsimulation models. The National Academy of Sciences report, ***Combining Information***, published by the American Statistical Association (Draper et. al. 1992) demonstrates how statistical matching relates to the broader topic of combining information across numerous data sources to assist in better decision-making.

An important strand of research treats statistical matching in the framework of the missing data problem and relies on multiple imputation to create synthetic data sets for use in policy analysis. This approach was put forth by Rubin (1978, 1986) as an alternative to unconstrained and constrained statistical matching methods with their somewhat restrictive conditional independence assumptions. Kadane (1978) offered improvements similar in spirit to those of Rubin. Both are working from a Bayesian context. Moriarity and Scheuren (2000a and 2000b) elaborate on the work of Kadane and Rubin, updating it, among other things.

Some of the weaknesses found by the validation results described later are quite predictable given the literature on this topic. What is new is the attempt to put the failure of the conditional independence assumption into the context of other forms of microsimulation modeling error. Because of time and space restrictions, this latter goal has only been touched on very briefly.

3. Statistical Match Being Evaluated

Under contract to the Department of Health and Human Services, the Urban Institute statistically matched the 1996 March Income Supplement to the CPS and the 1995 NHIS. Two matches were performed, one with the CPS playing the role of Host and the other with the NHIS as Host. This satisfied users' different needs and served as a check on the validity of the matching methodology.

A fully-constrained predictive mean match with partitioning was used to perform the statistical match. The match was performed in two stages. In the first stage, missing data on health insurance coverage on the NHIS were imputed; in the second stage the actual match was performed. One benefit of imputing the missing health insurance data was that this item could then be used as a blocking variable. This was desirable because health insurance coverage is strongly associated with health status and health care utilization.

The missing NHIS health insurance data was imputed by performing a fully-constrained statistical match, using partitioning and predictive mean matching, with the NHIS as Host and the CPS as Donor. In this match, health insurance coverage was not used as a blocking variable in the partitioning. The missing health insurance data on the NHIS records were replaced with the health insurance data from the matching CPS Donor records. The final matches were performed using this "imputed" version of the NHIS with health insurance coverage available as a blocking variable.

The steps that were followed in both the "imputation"and final matches are described briefly in the subsections which follow.

3.1 Partitioning. An extensive and flexible partitioning scheme was imposed on the two data files to create cells (equivalence classes). Matches were only allowed within these cells. The criteria used to choose the blocking variables were: (1) The variables selected had to be important determinants of health status and health care utilization in the NHIS and of income variables in the CPS. (2) The variables selected had to be defined similarly in both surveys.

Regressions were employed to select the blocking variables as recommended by Paass(1985). The selected variables were ordered in the partitioning scheme according to their predictive power.

When implementing the partitioning scheme using the selected blocking variables, an attempt was made to keep the unweighted cell sizes between 30 and 1,000 on the CPS and between 20 and 500 on the NHIS. The maximum cell size is important because it affects the computer time required and, if large, may mean that the opportunity for deeper partitioning may not have been realized. The minimum cell size is important as the sampling properties can be adversely affected if cell sizes are too small. Higher minimums should be used if the predictive power of the blocking variables is low, or if serious misalignment of the two files exists. Note that partitioning was not as deep for some groups (e.g., blacks) as for others because of cell size minimums.

When implementing the partitioning, the weighted cell counts for each of the cells in the two files were compared to see if they were unbalanced. Two cells were considered to be unbalanced if the difference between the weighted cell counts was large. When an imbalance was discovered, either (1) the order of the partitioning was changed or (2) further partitioning of the cell was prevented.

3.2 Predictive mean matching. A weighted least squares model was fit to two variables on the Donor file. Predicted values for these two variables were calculated on both the Donor and the Host files using the parameter estimates obtained from this model. To ensure that all records of both files were used in the match, the weighted cell counts of the Donor file were scaled so that they equaled those of the corresponding Host cells. Next, the records within each cell were sorted on the fitted value of one of the predictive mean matching variables (the Z-variables) so that records with the closest regression-weighted values (presumably the closest values of the X-variables) would have a similar rank order. Each record in the Host File was matched to the closest Donor record(s) based on their rank orders, splitting records on the Donor and Host file if necessary to ensure that all the weight on the Donor and Host records was used up.

After the matches were performed, a set of diagnostics were prepared for evaluation purposes. Generally, the weighted means of the selected variables on the matched files were close to the means on the Donor file. Interestingly, the fit appeared better for variables that have a smaller proportion of nonzero responses. It is unclear why this is the case. Also, variables that are known to be under-reported on the CPS (e.g., interest and dividends) have a slightly poorer fit on the matched file.

4. Data sources, especially the NSAF

By sheer happenstance, the 1997 NSAF was available to evaluate the statistical match of the 1995 NHIS with the March 1996 CPS. As none of these three surveys was carried out with the objective of being statistically matched, our evaluation had challenges, only partly overcome. The primary drawback was that the NSAF collected data for a different year than the NHIS and CPS surveys used in the statistical match. Another drawback is that the NSAF primarily collects telephone interviews, whereas the NHIS and CPS collect in-person interviews. This difference in format may well affect the data collected and associations among variables. Despite this, the NSAF offers a unique opportunity to assess the quality of the matched file because many of the questions in NSAF were taken from the CPS or NHIS.

The CPS and NHIS are well-known and will not be described here. Readers unfamiliar with these surveys can obtain detailed information about them from their respective agency web sites. For the NHIS see http://www.cdc.gov/NCHS>. For the CPS : see <http://bls.census.gov/cps.cpsmain.html>. NSAF, <http://newfederalism.urban.org/nsaf>, on the other hand, may be new to many, and therefore, is described briefly below.

The NSAF is a survey of the economic, health, and social characteristics of children, adults under the age of 65, and their families. NSAF data collection was conducted for the Urban Institute and Child Trends by Westat in 1997 and 1999. Plans are to conduct a third round in 2002. Data were collected on large, nationally representative samples. The 1997 survey included over 44,000 households, yielding information on more than 109,000 persons under the age of 65.

The NSAF sample had two components. The main component consisted of a random-digit dial (RDD) sample of households with telephones. The RDD sample was supplemented with an area probability sample of households without telephones. In both the RDD and area samples, interviewing was conducted in two stages. First, a brief screening interview was conducted to determine household eligibility. Households with only adults age 65 and over were not eligible. Second, the main interview was administered to a subsample of eligible telephone households and to all eligible nontelephone households.

The subsampling rates for telephone households depended on the screening interview response to a single question about household income and on the presence of children in the household. The sampling rate for low-income households with children was 1; the sampling rates for higher-income households with children and all households without children (but with someone under 65) were less than one. In 1997, the extended interview was conducted in 42,973 telephone households.

In the area sample, households within sampled blocks were screened and all nontelephone households with someone under 65 received the extended interview. Because only a small fraction of households do not have a telephone, block groups from the 1990 Census that had a very high percentage of telephone households were eliminated from the area sampling frame. A special coverage adjustment was made during the weighting process to account for excluding persons in nontelephone households in these block groups. In the 1997 NSAF, 1,488 nontelephone households received the extended interview. In all, 44,461 households were in the 1997 survey.

5. Evaluations Done and Results Obtained

The CPS and NHIS include persons of all ages, whereas the NSAF includes only persons under 65. Thus, our evaluation was limited to the population under age 65. In this paper, we present data only for 18-64 year olds. The three surveys cover slightly different years (CPS-March 1996 with 1995 calendar year income, NHIS-1995, and NSAF-1997). For our evaluation, differences across the surveys in the marginal distributions of variables due to the differing survey years

were not of interest. Hence, we eliminated the marginal differences by standardizing on the NSAF marginals.

Our main focus was the extent to which associations between the Y- and Z-variables that are present in the population were preserved in the statistically matched CPS/NHIS data set. We used a chi-square test of independence to assess this. The chi-square test was performed for various Y by Z contingency tables using weighted cell percentages rather than cell counts. For the chi-square tests of the matched file contingency tables, the expected values were computed using the marginals from the corresponding NSAF contingency tables.

We planned to examine the effect of violations of the conditional independence assumption by looking at three levels of Y-Z association (given X): no association, weak/moderate association, and strong association. We conjectured that:

1) If the assumption of conditional independence holds (Y and Z are completely unrelated given X), then the Y-Z association measured in the matched file should be the same as that in the population (represented by the NSAF).

2) If, given X, Y and Z are moderately related, then the Y-Z association measured in the matched file should be detectably different from that in the population (NSAF). However, given the other sources of error in the CPS and NHIS, the difference may not be of any consequence.

3) If, given X, Y and Z are strongly related and the predictive power of the X-variables is not high, then the Y-Z association measured in the matched file may be so different from that of the population (NSAF) as to limit use of the Y and Z variables on the matched file.

In what follows, we present data for only one association, that between the family income-to-needs ratio and number of doctor visits, two variables thought to be strongly related. On the statistically matched data set, the income-to-needs ratio was calculated using variables from the CPS; the number of doctor visits variable was from the NHIS. The income-to-needs ratio is calculated by summing the income of all family members and dividing this sum by the federal poverty line for a family of that size and composition.

For this analysis, the income-to-needs ratio was categorized into six categories: under 50%, 50% to under 100%, 100% to under 150%, 150% to under 200%, 200% to under 300%, and 300% or more. Number of doctors visits was grouped into five categories: 0 visits, 1 visit, 2 visits, 3-5 visits, and 6 or more visits. Two 6x5 tables for the cross-classification of income-to-needs and doctor visits were constructed, one from NSAF and the other from the CPS/NHIS statistically matched data set. These tables are shown below, with the entries expressed as weighted percents.

Table 1. Income-to -needs ratio by number of doctor visits: Adults 18-64. CPS/NHIS matched file

Income-to-needs ratio	Number of doctor visits				
	0	1	2	3-5	6+
	Weighted percent				
<50%	1.8	1.1	0.6	0.9	0.7
50% to <100%	2.6	1.4	1.0	1.2	1.1
100% to <150%	2.8	1.7	1.1	1.3	1.1
150% to <200%-	3.1	1.8	1.3	1.4	1.2
200% to 300%	5.8	3.7	2.6	3.0	2.0
>=300%	15.7	12.6	9.0	10.4	6.0

Table 2. Income-to -needs ratio by number of doctor visits: Adults 18-64. NSAF

Income-to-needs ratio	Number of doctor visits				
	0	1	2	3-5	6+
	Weighted percent				
<50%	2.1	0.8	0.6	0.9	0.7
50% to <100%	3.1	1.1	0.9	1.1	1.1
100%to <150%	3.4	1.4	0.9	1.3	1.0
150%to <200%-	3.3	1.8	1.2	1.4	1.0
200%to 300%	5.6	4.0	2.5	3.0	2.0
>=300%	14.4	13.1	9.4	10.5	6.2

5.1 Results. The chi-square tests from the NSAF and CPS/NHIS matched file were quite different (2.27 and .71, respectively, with a ratio of .32). The ratio of the two chi-squares shows that only about a third of the information about the association between the income-to-needs ratio and number of doctor's visits was captured in the statistically matched CPS/NHIS data set. This was disappointing, albeit predictable given the criticisms of statistical matching regarding violations of the conditional independence assumption. The association was most affected (as measured by the size of the relative residuals) for low income and poverty families and for persons reporting 0 or 1 doctor visit. However, while the percentages were statistically different, for some purposes they may not be substantively different.

5.2 Confidence Interval. There is, of course, sampling error around this result and a need for thoroughness

makes it necessary to create a confidence interval around the ratio. Calculating variances for statistically matched data is difficult. To do the job properly, we would have to perform multiple statistical matches using independent or balanced replicates of the main data sets e.g.,Atrostic 1994,1995; Zaslawsky and Thurston, 1994, 1995; Thurston and Zaslawsky, 1996). Statistical matching can be considered a form of imputation and the work of Rao and Shao(1992) can be used to decide how to set up the variance calculations in a way that gives enough stability, without introducing bias. Such a complicated procedure was not feasible here.

For this exploratory analysis, we obtained approximate confidence limits using the following approach. The existing statistically matched file was divided into 16 "replicates" by CPS rotation group (months-in-sample 1 and 8; 2 and 7; 3 and 6; 4 and 5) and NHIS calendar interview quarter (first, second, third and fourth). Income-to-needs by doctor visits tables, like table 1, were constructed for each of the 16 subgroups and chi-squares obtained for each table as was done for table 1. The upper limit of the range of these chi-square statistics fell far short of the NSAF estimate, suggesting the difference observed was real.

6. Conclusions, Conjectures and Next Steps

Our analysis showed that the association between the income-to-needs ratio and number of doctors visits was not preserved well in the statistically matched CPS/NHIS data file. Our plans are to repeat the analysis for additional strongly-related variables and for variables with moderate or no association. We will also perform some internal checks (within NHIS or CPS) when possible. We continue to conjecture that, in the case of moderate or weak relationships, the effect of violations of the conditional independence assumption may not matter statistically or substantively either. The irony could be that only the associations between strongly related variables are affected by violations of the conditional independence assumption and that these are the only associations of interest.

6.1 Recommendations First, there are some rudimentary measures that should always be taken when carrying out a statistical match to ensure its quality. For example, the **X-** variables used in the matching should be defined the same way on the surveys being matched and they should be well measured or used in a way that minimizes the impact of measurement error. In addition, we recommend that information about associations between **Y-** and **Z-** variables obtained from prior surveys, when possible, be used when designing the match. This is a variant of what Paass (1985) has recommended. Forth-coming work by Moriarity that involves incorporating the Y-Z associations in the match hold promise for improving statistical matches.

We have two recommendations regarding modifications to surveys that would improve statistical matches involving them. First, employ matrix sampling, a technique whereby subsamples respond to different subsets of the Y and Z items so that all Y-Z associations can be measured without undue respondent burden. This would make information about Y-Z associations in that data file available for use in the match design. Second, add **X-** variables to the two surveys that will improve prediction of the **Y-** and **Z-** variables from the **X-** variables (e.g.,Doyle 2000).

6.2 Next steps. We plan additional work on the effect of violations of the conditional independence assumption on the quality of statistical matches using the CPS/NHIS matched files, as already noted. Our hunch is that the most important step towards improving our ability to capture Y-Z associations in statistically matched data sets is to incorporate reasonable measures of these associations into the matching procedure.

One approach we may explore is to posit plausible ranges for the associations of the most important variables. Three matches would then be done, one under conditional independence and two using the end points of the association range. Then the end user would have a range of estimates to use.

Afterword

It is rather surprising that statistical matching, now about 40 years old and still widely used, should be so unfinished in its operating technology and underlying theory. Still, practical people have in many cases no good alternative, so fail-safe uses are needed. The above recommendations may be helpful but we remain unsure as to whether they will be tried generally.

Acknowledgments and References

H. Lock Oh did the contingency table analyses, only a small portion of which were presented. Rana Atie and Lucy Chung also provided important assistance.

References

References cited in the text are available upon request.