Imputation Techniques in Microsimulation*

Hyunsub Kum**

Abstract: Microsimulation has gained attention for its use in analyzing and forecasting the individual impacts of alternative economic and social policy measures. In practice, however, microsimulation cannot be carried out from a single data source, since it requires far more information than any single data source can provide. This paper discusses ways to combine separate data sources when there are no identical key variables, using imputation techniques, to make a large but synthetic data source for microsimulation. A new approach based on propensity score matching is suggested and discussed.

Keywords: microsimulation, imputation, policy analysis, data management

INTRODUCTION

Microsimulation as an instrument for analyzing and forecasting the individual impacts of alternative economic and social policy measures has been increasingly discussed in the economics and policy literature. Microsimulation is the process of imitating the behavior of individuals or groups in a system through a hypothesized model of behavioral responses and institutional relationships. Its aim is to see the effects of planned economic or social policies at the individual or subgroup level. Since the economic and social world usually does not allow experimentation to gauge the effects of policy,[1] simulation via an appropriate model is a natural alternative. It is usually concerned with behavioral responses at the individual level or the distributional consequences of policy changes on particular subgroups (Merz 1994).

Since Orcutt's (1957) seminal work, microsimulation has been used in business,

---

* This article is partly based on an earlier draft, presented with Thomas Masterson at the Levy Institute of Economics, on creating a new measure of economic wellbeing.

** Hyunsub Kum is Assistant Professor in the Graduate School of Public Administration, Seoul National University, Korea. Email: hyunsk@snu.ac.kr.

[1] A few such real-life social experiments have been conducted—for instance, the New Jersey Income Maintenance Project (Rees 1977).

Manuscript received October 10, 2010; out for review October 21, 2010; review completed December 2, 2010; accepted December 8, 2010

The Korean Journal of Policy Studies, Vol. 25, No. 3 (2010), pp. 69-83.

education, and policy research. By simulating, for instance, various aspects of market conditions such as customer's preference, technologies, and supply and demand, a company can set up a new strategy for production, investment, and consumption. Especially in policy research, the impacts of alternative tax and transfer policies can be analyzed and new options can be proposed based on microsimulation results. Related practical examples can be found in many areas.

Microsimulation was originally developed as a conditional forecasting tool to investigate the direct and indirect effects of alternative scenarios in policy analysis by comparing them with the status quo, or with the absence of any policy, at the individual level. Conditional forecasting means that the forecasting is conducted within the scenario based on assumptions about behavioral responses and institutional relationships. In practice, many requirements and components are needed to perform a microsimulation. Since the structure of a microsimulation model is mainly determined by "if-then" relationships expressed in mathematical equations and decision structures on representative data, simulated results are highly sensitive to various settings such as the concepts, micro-macro link, assumptions about individual behavioral responses, algorithm (or simulation module), and micro-level data. Merz's (1994) discussion of microsimulation requirements reflects this kind of difficulty in detail.[2]

This paper tries to shed light on the requirements for data preparation. In microsimulation, since every estimate should be extracted from the results of simulations based on micro-level data, it is critical to collect reasonably representative micro-level data. Due to time limits and costs, however, it is not feasible to use the population data as a whole; a sample survey or administrative data or both are usually employed instead. The problem with this approach is that no single sample survey or administrative data set can cover all the individual behaviors and institutional regulations that are of interest in microsimulation. One alternative is to combine the separated data or impute one with the other by providing extra attributes or expanding the detail of an attribute already available.

The purpose of this short study is to describe the need and options for large-scale imputation and describe its theoretical basis as well as the algorithms and procedures by which it is achieved. In particular, an algorithm based on propensity score matching

---

2. Merz (1994) indicated five requirements for microsimulation: (1) the provision of an appropriate micro database with tools for merging several databases; (2) the construction of micro and macro modules including institutional regulations and behavioral responses that compute certain characteristics of micro-units; (3) the simulation of the micro-model itself in modifying micro and macro data; (4) the adjustment of micro-data to fit given control data; and (5) the evaluation of the executed simulation, whether it is the result of a single simulation or a number of stochastic or deterministic simulations.

that has been popular in evaluation studies is examined in detail. Although the aims of evaluation and imputation studies are different, the logic of propensity score matching in searching for closest and compatible records between two data sets can be employed to construct an imputation algorithm.

The following section discusses the need for data imputation in microsimulation, and the third section introduces a special type of imputation called statistical matching that is used widely in practice. The fourth section explains general algorithms of matching, and the fifth section examines the algorithm based on propensity score and its distinctions compared with others. Some limitations and suggestions for future study are discussed in the conclusion.

IMPUTATION

Imputation has been widely discussed and used to fill in the missings in data analysis. It is well known that incomplete information can distort the overall analysis. Simply dropping the missings would taint the sampling frame that was originally designed to represent the population. But keeping them does not help the quality of analysis at all. Imputing the missings as an alternative is recommended by some scholars, but still heavily criticized by others because imputed values are artificial. Therefore, there is wide interest in the arguments surrounding imputation, its procedures, and the quality of their results.

If a researcher is contemplating imputation, three possibilities should be taken into account. The first possibility is that one data set has huge amounts of information on households with some missings due to errors such as nonresponse, editing, or unquestioned. If missings occur relatively seldom, several imputation techniques—such as hot-deck, cold-deck, or predicted mean value imputation based on regression models—can be employed. This kind of imputation is only reasonable when the missings occur in relatively small frequency and there is an acceptable internal reference. Internal reference here means that data are provided within the data set by the same individual at different times (through aging the variables) or by different but closely similar individuals (through replacement or matching). Of course, care should be taken to preserve marginal and joint distribution during the imputation. But when it is done, the imputed cases are treated as ordinary non-missing cases.

Second, missing information in one data set may be available from external reference or another data set (based on the same population), in which case the former can be replaced with the latter. If two data sets share the same identity (social security number or personal identification number for example) with different variables, exact matching (or record linkage) based on the same identity can be made by transferring the non-missing attributes from the latter to the former with missings. In this case, the quality of key variable such as identification number such as SSN or PIN is critical. But it is often hard to get a reasonable key variable due to the protection of privacy in the survey or administrative data, and it may be allowed only under certain strict conditions.

Third, if there are more missings than usual and no reliable internal or external reference can be sought within the data frame, the usual imputation techniques cannot be applied. Furthermore, if missings occur not in a relative sense but in an absolute sense—if important variables are entirely lacking, rather than just some missings within the variables—only two alternatives are available. One is to run a new survey to cover both existing and nonexisting variables together, which is usually impractical; the other is to combine the variables from two or more data sets with reasonable assumptions. The latter is called statistical matching. It is relatively new to academic studies, but is commonly used in practice because it can be an efficient way to work with fragmented data sets.

STATISTICAL MATCHING

Statistical matching is a technique for linking two records in separate data sets by means of some common (not the same) variables to both data sets and generating a new synthetic data set (D'Orazio, Di Zio, and Scanu 2006). Its aim is to allow more flexible analysis than could be carried out on either data set alone—in particular, when the associations between variables are never jointly observed but in need. Thus, it is a tool to produce a complete but synthetic data set.

In Europe, statistical matching was originally pursued by media planners who were interested in consumer behaviors. They wanted to examine, for instance, the empirical association between individual television viewing and purchasing behavior by linking the viewing information available from a television measurement panel with the purchasing data available from an existing large market-tracking panel through the means of some variables common to both data files (Rassler 2002).

In the United States and Canada, federal agencies rather than businessmen started to merge data from different sources in 1970s. Due to the growing interest in social problems at the microeconomic level, and the lack of a consistent and comprehensive set of household income data for instance, the demand for large amounts of disaggregated economic and demographic information encouraged the construction of a new micro-analytic database through simulation. Furthermore legal restrictions on the use of administrative data in US federal agencies made it hard or almost impossible to do exact matching or records linkage. The so-called 1966 MERGE File was conducted by the US Office of Economic Opportunity combining the 1967 Survey of Economic Opportunity and the 1966 tax file through statistical matching without specific individual identification information (Okner 1972; Budd 1972; Ruggles and Ruggles 1974). Similar matching was carried out for the Canadian Survey of Consumer Finances and the Family Expenditure Survey in Canada (Alter 1974).

Since then, statistical matching has been developed and used in many empirical studies-including medical and statistical literature (Little and Rubin 2000; Rubin and Thomas 1992, 1996; Rosenbaum and Rubin 1983; Rubin 1974), economics (Radner 1981; Greenwood 1983, 1987; Kennickell 2001; Sutherland, Taylor and Gomulka 2001), and sociology (Keister 2000, 2004). Furthermore, most microsimulation studies, such as the Urban-Brookings Tax Microsimulation (Rohaly, Carasso, and Saleem 2005) and the Congressional Budget Office report on the income tax burden (CBO 2003), employed this technique to prepare the relevant micro-data sets.

In the standard statistical matching framework, two data files, A and B, exist 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. Both have variables Z, but they are not usually from the same identity (person, firm, or organization). In this situation, one wants to construct a new data set with information on X, Y, and Z together, but that kind of data set is not available in reality.

Assume that file A (which exclusively has X) is the recipient data set and file B (which exclusively has Y) is the donor data set, because we want to transfer the information about Y in file B to file A. Usually, combining two data sets is done by identifying the key information about the same identity such as SSN or PIN; but in statistical matching, two data sets are combined using a particular distance function to assess the similarity between the records in both data sets, since there is no key information for identification.

Instead of searching for the same identity, certain distance functions are constructed by their common variables (Z) with various forms and used to search for a donor record that is the most similar to a recipient record. This record is called the nearest neighbor. Then the nearest neighbor record of the donor file (Y) is added to the recipient file, which leads to a new and complete but artificially matched data set containing X, Y, and Z variables. Once the data sets are matched and merged, the analysis proceeds as if the artificial matched data sets were a real sample representative of the true population of interest (Radner 1981; Keister 2000; Rassler 2002; Wolff and Zacharias 2007).

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 to both files) are multivariate random variables with a joint probability or density function fxyz, while usually no single data set has information on X, Y, and Z together. Also we assume that the records of each sample as well as the samples themselves are drawn randomly and independently of each other from the same population. In other words, each sample to be matched is regarded as a single-source random sample from the underlying population but with different variables. Thus, combining the two data sets is acceptable if and only if the specific variables Y and X are conditionally independent given the common variables Z=z (Sims 1972).

Another way to look at this conditionally independent assumption is to check the covariance before and after matching.

$$
\begin{array}{ll}
(\textit{before}) & \mathrm{cov}(X,Y) = \mathrm{E}(\mathrm{cov}(X,Y \mid Z)) + \mathrm{cov}(E(X \mid Z), E(Y \mid Z)) \\
(\textit{after}) & \hat{\mathrm{cov}}(X,Y) = \mathrm{cov}(E(X \mid Z), E(Y \mid Z)) \\
 & E(\mathrm{cov}(X,Y \mid Z)) = \mathrm{cov}(X,Y) - \hat{\mathrm{cov}}(X,Y)
\end{array}
$$

Therefore, statistical matching can be regarded as good only if the variables X and Y are on the average conditionally uncorrelated—that is, E(cov(X,Y | Z))=0.

Two types of statistical matching have been developed: unconstrained and constrained (Rodgers 1984). Both are performed by transferring the nearest possible records from the donor to the recipient by comparing the similarity between records. The difference exists in the guidance and constraints of the similarity—similarity in records is emphasized in unconstrained matching, and similarity in marginal distribution is stressed in constrained matching.

In unconstrained matching, every record in the donor data set is compared with every record in the recipient data set based on the predefined criteria of similarity, and then any record for which the criteria are met is selected. This procedure allows multiple selections or no selection of donor records for a recipient record. As a result, it is possible that some donor or recipient records will remain unmatched, due to the strictness of predefined criteria of similarity. That would cause not likely to satisfy the requirement of quality criteria to be assessed, which means preserving the empirical marginal distribution of Y or empirical conditional distribution of Y given Z in the matched file compared with those in the original donor file.

Constrained matching is designed to overcome this shortcoming. In this approach, all of the records in both files are preserved in the matched file in terms of weighted total population. The (simple) constrained match can be written by means of the weights wij, and the objective function of the problem is

$$(1)\ \sum_{j=1}^{nR}\sum_{i=1}^{nD} d_{ij} w_{ij}$$

where $w_{ij}>=0$ and d_{ij} is distance between donor and recipient observations, which has to be minimized according to the restrictions of weights for donor and recipient. In other words, the empirical multivariate distribution and the marginal distribution of the variables in the donor file are replicated in the statistically matched data set by splitting or adding up the weights attached to each record. Since weights attached to a record indicate how many people the record represents, the preservation of weighted sum means the preservation of marginal distribution.

For instance, when weights for each donor record are split up, only some of the weights of the donor record are transferred to a recipient record, up to the same amount that the recipient record has, and the leftover weights of the donor record are transferred to another nearest neighbor recipient record sequentially. A weight-split or duplication procedure across donor and recipient records is kept continuously during the matching procedure until the weights in both files are exhausted. For this reason, the weighted population totals should be equalized between donor and recipient files beforehand and can be preserved afterward.

One noticeable feature of constrained matching is that matching is conducted according to the rank based on a predefined similarity index rather than absolute values of itself or common variables Z (Rodgers 1984). This is why constrained matching is frequently called an imputation based on rank, and many linear programming approaches have been employed in earlier practices. Due to the nature of rank order matching, however, some matches may end up with a large distance that is not acceptable to researchers which is the main disadvantage of constrained matching.

To reduce the impact of this disadvantage, stratification or segmentation, grouping some records according to a certain criteria, is carried out before matching. This leads some matches between certain types of records to be avoided or required. For example, males should be matched with males, not females, and teenagers should be matched with teenagers, not people in their sixties, if sex and age variables are critical in matching. Stratification could in fact narrow the distance or similarity between records in donor and recipient data sets and lead to a tighter match. For this, both donor and recipient data sets are first stratified into N matching cells, which are identically defined in a mutually exclusive and exhaustive way, and then matching is allowed only within the same cell. This is similar to the use of strata variables in survey sampling, and it ensures perfect matches across files within it.

Of course, a single variable as well as combinations of multiple variables can be used in selecting the strata variables. Strata variables are usually selected from common variables based on subjective considerations or their explanatory power in a certain model. When strata variables are defined and stratification is done accordingly, the distance functions can be applied separately and matching is performed separately at the matching cell level. Sometimes, of course, different distance functions with respect to variables and their subjective weights can be constructed for different cells.

MATCHING ALGORITHMS

In the construction of a matching algorithm to find a closest record, it is important to decide how to define the similarity. Two broad approaches are employed in the previous literature. Under the assumption stated above, the main task of statistical matching is to search for a donor record Y_j whose observed values of the common variables Z_j are identical or closest to those Z_i of the recipient record X_i. It can be shown simply as follows:

$$\text{(2) file A: } (X_i, Z_i) + \text{file B: } (Z_j, Y_j)$$

where Z_i equals or closest to $Z_j \rightarrow$ file C: (X_i, Z_i, Z_j, Y_j).

Usually, this searching process is carried out using an algorithm based on nearest neighbor matching by calculating a distance function. The distance function can be constructed in various forms. If we assume Euclidean distance, for instance, the distance function is

$$\text{(3) } d(z_i, z_j) = \sqrt{\sum_{k=1}^{p} \mathrm{g}(z_k)(z_{ki} - z_{kj})^2}$$

where $\mathrm{g}(z_k)$ is an individual weight for each distance, z_{ki} is k^{th} common variable in the recipient file, and z_{kj} is k^{th} common variable in the donor file. For every record of the recipient file, distance measures using all donor records are calculated and the donor record that has the smallest distance is selected.[3]

But the procedure is not as straightforward as it appears. Several caveats should be taken into account in constructing the algorithm. First, it should be cross-checked whether donor and recipient files differ significantly by means of the common variables (Kum and Masterson 2008). This is because matching is based on the assumption that both are randomly drawn from the same population, and there should be at least some close similarity in marginal distributions of the common variables even though they are not identical. Furthermore, to account for the different scales of the common variables—for instance, age, education, and occupation—it is better to stan-

---

3. Due to this feature, many earlier applications used linear programming techniques.

dardize continuous (even ordinal) variables with zero mean and one standard error.

Second, the algorithm of a distance function to find a relevant donor record for a recipient record may use all or some of the common variables. The main criterion is whose distance is minimal. But there is no explicit restriction on how to compose a distance function. Therefore, a subjective weight for each common variable can be used in the distance function to incorporate the subjective importance of that variable to the researcher.

Third, constrained matching requires that some common variables should match perfectly. Therefore, stratification and restriction on matching to be bounded within it can be imposed in the matching procedure. Also, one donor record may be used for different recipient records and vice versa, based on the size of the attached weights. To limit the number of times a donor is taken for a recipient, however, some penalty scheme can be placed on donor records through the distance function. This can be realized with an additional weighting factor. But this restriction may lead to a loss in variability or sample size.

Fourth, there are many ways of constructing a distance function to compare the similarity between donor and recipient records. One popular method is predicted mean matching, in which searching for the closest record is performed through regression estimation rather than other distance functions. The procedure is as follows: first, one or some of the Y variables are regressed on the common variables Z in the donor data set, and the predicted values of Y ($\hat{Y}_d$) are obtained. Second, the predicted values of Y ($\hat{Y}_r$) in the recipient data set are obtained by applying the parameter estimates from the donor data set to the same sets of Z variables in the recipient data set. Since the Y variable does not exist in the recipient data set, this will produce the simulated values of Y in the recipient data set, but based on the formulated relationship in the donor data set. Third, matching is conducted by selecting the nearest neighbor records from each data set based on the predicted values of Y ($\hat{Y}_d$ in the donor set and $\hat{Y}_r$ in the recipient set). The closest record can be defined in terms of the predicted value of Y or its rank within the predefined cell. In the latter case, therefore, stratification with balanced weights is critical.[4]

---

[4] This algorithm is quite popular and was adopted by the Urban-Brookings microsimulation model.

PROPENSITY SCORE MATCHING

As discussed above, the construction of the distance function can be based on various theoretical underpinnings. Not only Euclidean but also other geometric distances, such as hyperbolic or elliptic, are possible. Predicted mean based on regression is another case. In recent years, a new approach to the construction of the distance function has been suggested in the evaluation literature. Propensity score matching originally referred to a multivariate method to construct control groups that have similar distributions on many covariates compared with those of treated groups in evaluation studies when a randomized experiment is not available.[5]

It is usually assumed in propensity score matching that the treated and control groups are drawn randomly and independently from the same underlying population. If the variables observed only in one group are conditionally independent from the assignment rule (T) given the covariates Z=z, then it is regarded that the assignment of the records to each group (T) is strongly ignorable given the covariates Z=z. Randomization in experimental design is based on this assumption and can be expressed as

$$(4)\ f_{X|T,Z} = f_{X|Z} \text{ and } f_{Y|T,Z} = f_{Y|Z}$$

If the assignment (T) is strongly ignorable given Z=z, as Rosenbaum and Rubin (1983) proved, then it is also strongly ignorable given any function of the observed covariates Z. In this regard, the function of Z is called a balancing score b(z). In other words, if the balancing hypothesis is satisfied, observations with the same propensity score must have the same distribution of observable (and unobservable) characteristics independently of treatment status, and therefore treated and control units should be on average observationally identical.

One significant feature of the propensity score matching method is that the propensity score reduces the dimensionality problem of matching treated and control units on the basis of the multidimensional vector X. This is a very important leverage point in statistical matching because (1) the large number of common variables with different weights should be considered at the same time in the search for the nearest neighbor or closest records; and (2) donor and recipient data sets may show different empirical distributions of the common variables due to different sampling designs or over- or under-sampling of specific population groups even from the same population. Thus, dimensionality reduction through balancing score can be an attractive alternative for

---

[5] Rosenbaum and Rubin (1983) defined the propensity score as the conditional probability of receiving a treatment given pre-treatment characteristics.

these reasons.

To further the logic of propensity score matching that can be incorporated in statistical matching, if we follow the above and regard the treated and control groups as donor and recipient data sets, logic indicates that the distributions of the covariates for recipient and donor data sets can be identical if the balancing scores b(z) in both data sets are identical.

$$(5)\ f_{X|T,b(z)} = f_{X|Z} \text{ and } f_{Y|T,b(z)} = f_{Y|Z}$$

Here the specification of b(z) can be variously constructed. For instance, exact matching (or record linkage) based on identical common variables (identification number, for instance) can be regarded as a special case of propensity score matching when identical Z is used as a balancing score. Propensity score based on predicted value of binary logistic or probit regression is another one (Rassler 2002). When this is the case, the propensity scores are defined as the conditional probability of a record i to belong to a certain group (T=1) given the common covariates (Z=z), and is expressed as

$$(6)\ e(z_i) = P(T = 1|\mathrm{Z} = z_i) = g(z'_i\beta)$$

where dependent variable (T=1 or 0) is membership status (donor or recipient) and independent variables are the selected common variables (Z).

$$(7)\ \hat{e}(\mathrm{z}_i) = g(z'_i\hat{\beta})\ \frac{1}{1+e^{-z'_i\overset{6}{\beta}}}\ \text{(logistic model)}$$

$$(8)\ \hat{e}(\mathrm{z}_i) = g(z'_i\hat{\beta}) = \Phi(z'_i\,\hat{\beta})\ \text{(probit model)}$$

When the propensity scores are estimated, the matching procedure is straightforward. The records with closest propensity scores from the two data sets are regarded as records that have the closest similar identities to be matched and transferred.

Two cautions should be noted based on several empirical experiments. First, selection of the specific common variables in the model (logistic or probit) to estimate propensity scores should be made carefully to maximize the explanatory power(Kum and Masterson 2008). This is because the predicted propensity score is supposed to work as a predictor for each record's membership (T=1 or 0), and the quality of a predictor heavily hinges on the power of the common variables in the model. Although individual coefficient estimates in the model are not the main concern here, a reasonable model-fit seems good for prediction in matching.

Second, after estimating the propensity scores, it is better for all records for each data set to be sorted by estimated propensity scores $\hat{e}(z_i)$ in ascending order and weights size in descending order. Under this twisted sorting scheme, a record with the bigger weight in one file would be split and matched to a record and several subsequent records in the other file until all of its weight is used up. A certain level of flexibility, such as collapsing across records due to the imbalance of weight size, seems inevitable (Kum and Masterson 2008).

CONCLUSION

Microsimulation is a useful alternative for policy analysis, but its goal can be accomplished only with a proper set of components. Without proper data preparation (one of the most neglected components), no microsimulation can yield fruitful results. Usually, microsimulation requires representative and large amounts of micro-level data. Largeness can be defined in vertical (observations) as well as horizontal (attributes) terms. According to Zaidi and Scott (2001, p.5), “Sample size is related to representativeness. Thus, the base data must have a large enough sample size to allow reliable disaggregation of simulations across subgroups of interest.” Also, there must be enough attributes (variables) to cover the various dimensions of behavioral relationships. Without sufficiently detailed information (sample size and attributes), microsimulation cannot achieve the full breadth of its potential for analysis and forecasting.

Unfortunately, there is no single data that meets all of these requirements. However, it is possible to construct a compatible but synthetic data set by imputation in large scale under the assumption of conditional independence. This study examined one of those possibilities, statistical matching, especially with propensity score matching, an idea adopted from the evaluation literature.

Although empirical examples have not been presented here, the method is already employed in economic research and has some potential for the use of microsimulation.[6] If the matching is well carried out, the synthetic data set can be used to capture the relationship between the variables of interest that are not jointly observed in any of the previously available data sets and can provide more detailed information for constructive policy design.

A number of limitations remain, and this approach needs further theoretical and

---

[6] A new microsimulation model for social policy in Korea is now under construction, and this study was prepared as a pilot trial.

empirical experiments to explore its practical applications. Among them, checking the quality of the match is critical. Under the constrained matching scheme, all marginal distributions should be preserved before and after matching, and only the joint distributions of variables not jointly observed in initial data sets may be different. A statistical matching procedure is regarded as successful if the marginal and joint empirical distributions of Z in the matched file and those are in the donor file are the same or similar. In this regard, the Lorenz coordinates, Gini coefficients, and deciles values, as well as the weighted mean and median values by strata variable, should be computed and compared between the donor and matched files.

All these tests, however, are restricted to comparing the marginal distributions of the variables in the donor and matched data sets. As such, they are necessary but insufficient indicators of the quality of the match. In theory, without auxiliary and completely observed data against which to check the validity of the synthetic data set, it is not that straightforward to perform the quality check at a sufficient level. In this regard, further simulation studies, such as Rubin's suggestion (1986) on multiple imputations with quality assessment design, should be seriously considered.

REFERENCES

Alter, H. E. 1974. Creation of a synthetic data set by linking records of the Canadian survey of consumer finances with the family expenditure survey. *Annals of Economic and Social Measurement* 3(2): 75-100

Budd, E.C. 1971. The creation of a microdata file for estimating the distribution of income. *Review of Income and Wealth* 17: 317-333

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical matching: Theory and practice*. Chichester, England: Wiley.

Greenwood, D. T. 1983. An estimation of U.S. family wealth and its distribution from microdata, 1973. *Review of Income and Wealth* 29 (1): 23-44.

Greenwood, D. 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*, ed. E. N. Wolff, 121-40. Oxford: Clarendon Press; New York: Oxford University Press.

Keister. L.A. 2004. Race, Family Structure, and Wealth: The Effect of Childhood Family on Adult Asset Ownership. *Sociological Perspectives* 47: 161-187

Keister, L. A., and S. Moller. 2000. Wealth inequality in the United States. *Annual Review of Sociology* 26: 63-81.

Kennickell, A. B. 2001. Modeling wealth with multiple observations of income:

Redesign of the sample for the 2001 Survey of Consumer Finances. Unpublished manuscript, Board of Governors of the Federal Reserve System, Washington, DC.

Kum, H., and T. Masterson. 2008. Statistical matching using propensity scores: Theory and application to the Levy Institute measures of economic wellbeing. Working paper. Levy Economics Institute

Little, R. J., and D. 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.

Merz, J. 1994. Microsimulation—a survey of methods and applications for analyzing economic and social policy. Discussion paper No. 9, FFB.

Okner, B.A. 1972. Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economics and Social Measurement* 1(3): 325-342

Orcutt, G. 1957. A new type of socio-economic systems, *The Review of Economics and Statistics* 58: 773-797

Radner, D. 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.

Rassler, S. 2002. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.

Rees, A., ed. 1977. *Labor-supply response*. Vol 2. of The New Jersey income maintenance experiment. New York.

Rodgers, W. L. 1984. An evaluation of statistical matching. *Journal of Business & Economic Statistics* 2 (January): 91-102.

Rohaly, J., A. Carasso, and M. A. Saleem. 2005. *The Urban-Brookings Tax Policy Center microsimulation model: Documentation and methodology for Version 0304*. Washington, DC: Tax Policy Center. http://taxpolicycenter.org/publications/template.cfm?PubID=9168.

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

Rubin, D.B. 1974. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* 69: 467-474

Rubin, D.B. 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* 4:87-94

Rubin, D. B., and N. Thomas. 1992. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* 79 (4): 797-809.

Rubin, D. B., and N. Thomas. 1996. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52 (March): 249-64.

Ruggles, N. and Ruggles, R. 1974. A strategy for merging and matching microdata sets. *Annals of Economic and Social Measurement* 1(3): 353-371

Sims, C.A. 1972. Comments on Okner(1972). *Annlas of Economic and Social Measurement* 1(3): 343-345

Sutherland, H., R. Taylor, and J. Gomulka. 2001. Combining household income and expenditure data in policy simulations. Working paper MU0101, Department of Applied Economics, University of Cambridge.

Wolff, E., and A. Zacharias. 2007. The Levy Institute measure of economic wellbeing. *Eastern Economics Journal*.

Zaidi, A., and A. Scott. 2001. Base dataset for the Sage model. SAGE Technical Note 1, ESRC-Sage Research Group, London School of Economics, 9.