Survey Methodology, December 1997
Vol. 23, No. 2, pp. 157-165
Statistics Canada

Regression Analysis of Data Files that are Computer Matched - Part II

FRITZ SCHEUREN and WILLIAM E. WINKLER[1]

ABSTRACT

Many policy decisions are best made when there is supporting statistical evidence based on analyses of appropriate microdata. Sometimes all the needed data exist but reside in multiple files for which common identifiers (*e.g.*, SIN's, EIN's, or SSN's) are unavailable. This paper demonstrates a methodology for analyzing two such files: (1) when there is common nonunique information subject to significant error and (2) when each source file contains noncommon quantitative data that can be connected with appropriate models. Such a situation might arise with files of businesses only having difficult-to-use name and address information in common, one file with the energy products consumed by the companies, and the other file containing the types and amounts of goods they produce. Another situation might arise with files on individuals in which one file has earnings data, another information about health-related expenses, and a third information about receipts of supplemental payments. The goal of the methodology presented is to produce valid statistical analyses; appropriate microdata files may or may not be produced.

KEY WORDS: Edit; Imputation; Record linkage; Regression analysis.

1. INTRODUCTION

1.1 Application Setting

To model the energy economy properly, an economist might need company-specific microdata on the fuel and feedstocks used by companies that are only available from Agency A and corresponding microdata on the goods produced for companies that is only available from Agency B. To model the health of individuals in society, a demographer or health science policy worker might need individual-specific information on those receiving social benefits from Agencies B1, B2, and B3, corresponding income information from Agency I, and information on health services from Agencies H1 and H2. Such modeling is possible if analysts have access to the microdata and if unique, common identifiers are available (*e.g.*, Oh and Scheuren 1975; Jabine and Scheuren 1986). If the only common identifiers are error-prone or nonunique or both, then probabilistic matching techniques (*e.g.*, Newcombe, Kennedy, Axford and James 1959, Fellegi and Sunter 1969) are needed.

1.2 Relation to Earlier Work

In earlier work (Scheuren and Winkler 1993), we provided theory showing that elementary regression analyses could be accurately adjusted for matching error, employing knowledge of the quality of the matching. In that work we relied heavily on an error-rate estimation procedure of Belin and Rubin (1995). In later research *e.g.*, (Winkler and Scheuren 1995, 1996), we showed that we could make further improvements by using noncommon quantitative data from the two files to improve matching and adjust statistical analyses for matching error. The main requirement – even in heretofore seemingly impossible situations – was that there exist a reasonable model for the relationships among the noncommon quantitative data. In the empirical example of this paper, we use data for which a very small subset of pairs can be accurately matched using name and address information only and for which the noncommon quantitative data is at least moderately correlated. In other situations, researchers might have a small microdata set that accurately represents relationships of noncommon data across a set of large administrative files or they might just have a reasonable guess at what the relationships among the noncommon data are. We are not sure, but conjecture that, with a reasonable starting point, the methods discussed here will succeed often enough to be of general value.

1.3 Basic Approach

The intuitive underpinnings of our methods are based on now well-known probabilistic record linkage (RL) and edit/imputation (EI) technologies. The ideas of modern RL were introduced by Newcombe (Newcombe *et al.* 1959) and mathematically formalized by Fellegi and Sunter (1969). Recent methods are described in Winkler (1994, 1995). EI has traditionally been used to clean up erroneous data in files. The most pertinent methods are based on the EI model of Fellegi and Holt (1976).

To adjust a statistical analysis for matching error, we employ a four-step recursive approach that is very powerful. We begin with an enhanced RL approach (*e.g.*, Winkler 1994, Belin and Rubin 1995) to delineate a subset of pairs of records in which the matching error rate is estimated to be very low. We perform a regression analysis, RA, on the

[1] Fritz Scheuren, Ernst and Young, 1225 Connecticut Avenue, N.W., Washington, DC 20036, U.S.A., Scheuren@aol.com; William E. Winkler, U.S. Bureau of the Census, Washington, DC 20023, U.S.A.

low-error-rate linked records and partially adjust the regression model on the remainder of the pairs by applying previous methods (Scheuren and Winkler 1993). Then, we refine the EI model using traditional outlier-detection methods to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (RA) is done and this time the results are fed back into the linkage step so that the RL step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, these *analytic linking* methods take the form

$$\begin{tikzcd}
& \mathrm{RA} \arrow[dr] & \\
\mathrm{RL} \arrow[ur] & \mathrm{RA} \arrow[l] & \mathrm{EI} \arrow[l]
\end{tikzcd}$$

1.4 Structure of What Follows

Beginning with this introduction, the paper is divided into five sections. In the second section, we undertake a short review of Edit/Imputation (EI) and Record Linkage (RL) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (RA) is so well known, our treatment of it is covered only in the particular simulated application (Section 3). The intent of these simulations is to use matching scenarios that are more difficult than what most linkers typically encounter. Simultaneously, we employ quantitative data that is both easy to understand but hard to use in matching. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

2. EI AND RL METHODS REVIEWED

2.1 Edit/Imputation

Methods of editing microdata have traditionally dealt with logical inconsistencies in data bases. Software consisted of if-then-else rules that were data-base-specific and very difficult to maintain or modify, so as to keep current. Imputation methods were part of the set of if-then-else rules and could yield revised records that still failed edits. In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976) introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values, so that the revised record satisfies all edits. An additional advantage of Fellegi and Holt (1976) systems is that their edit methods tie directly with current methods of imputing microdata (*e.g.*, Little and Rubin 1987).

Although we will only consider continuous data in this paper, EI techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$c_1X < Y < c_2X$$

In words,

> Y can be expected to be greater than c_1X and less than c_2X; hence, if Y less than c_1X and greater than c_2X, then the data record should be reviewed (with resource and other practical considerations determining the actual bounds used).

Here Y may be total wages, X the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$. When an (X, Y) pair associated with a record fails an edit, we may replace, say, Y with an estimate (or prediction).

2.2 Record Linkage

A record linkage process attempts to classify pairs in a product space $A \times B$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.* 1959; Newcombe, Fair and Lalonde 1992), Fellegi and Sunter (1969) considered ratios R of probabilities of the form

$$R = \Pr((\gamma\in\Gamma \mid M)/\Pr((\gamma\in\Gamma \mid U)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called *matching variables*. The decision rule is given by

> If $R > Upper$, then designate pair as a link.
>
> If $Lower \le R \le Upper$, then designate pair as a possible link and hold for clerical review.
>
> If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (typically a logarithm) a *matching weight* or *total agreement weight.*

With the availability of inexpensive computing power, there has been an outpouring of new work on record linkage techniques (*e.g.*, Jaro 1989, Newcombe, *et al.* 1992, Winkler 1994, 1995). The new computer-intensive methods reduce, or even sometimes eliminate, the need for clerical review when name, address, and other information used in matching is of reasonable quality. The proceedings from a recently concluded international conference on record linkage showcase these ideas and might be the best single reference (Alvey and Jamerson 1997).

3. SIMULATION SETTING

3.1 Matching Scenarios

For our simulations, we considered a scenario in which matches are virtually indistinguishable from nonmatches. In our earlier work (Scheuren and Winkler 1993), we considered three matching scenarios in which matches are more easily distinguished from nonmatches than in the scenario of the present paper.

In both papers, the basic idea is to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. Because the methods of this paper work better than what we did earlier, we only consider a matching scenario that we label "Second Poor," because it is more difficult than the poor (most difficult) scenario we considered previously.

We started here with two population files (sizes 12,000 and 15,000), each having good matching information and for which true match status was known. Three settings were examined: high, medium and low – depending on the extent to which the smaller file had cases also included in the larger file. In the high file inclusion situation, about 10,000 cases are on both files for a file inclusion or intersection rate on the smaller or base file of about 83%. In the medium file intersection situation, we took a sample of one file so that the intersection of the two files being matched was approximately 25%. In the low file intersection situation, we took samples of both files so that the intersection of the files being matched was approximately 5%. The number of intersecting cases, obviously, bounds the number of true matches that can be found.

We then generated quantitative data with known distributional properties and adjoined the data to the files. These variations are described below and displayed in Figure 1 where we show the poor scenario (labeled "first poor") of our previous 1993 paper and the "second poor" scenario used in this paper. In the figure, the match weight, the logarithm of R, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o).

3.2 "First Poor Scenario" (Figure 1a)

The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names in one of the files. Moderately severe typographical errors were made independently in one fourth of the addresses of the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was for the links to be made in a manner that a practitioner might choose after gaining only a little experience. The situation is analogous to that of using administrative lists of individuals where information used in matching is of poor quality. The true mismatch rate here was 10.1%.

3.3 "Second Poor" Scenario (Figure 1b)

The second poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names in one of the files. Severe typographical errors were made in one fourth of the addresses in the same file. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. Name information – a key identifying characteristic – is often very difficult to compare effectively with business lists. The true mismatch rate was 14.6%.

3.4 Summary of Matching Scenarios

Clearly, depending on the scenario, our ability to distinguish between true links and true nonlinks differs significantly. With the first poor scenario, the overlap, shown visually between the log-frequency-versus-weight curves, is substantial (Figure 1a); and, with the second poor scheme, the overlap of the log-frequency-versus-weight curves is almost total (Figure 1b). In the earlier work, we showed that our theoretical adjustment procedure worked well using the known true match rates in our data sets. For situations where the curves of true links and true nonlinks were reasonably well separated, we accurately estimated error rates via a procedure of Belin and Rubin (1995) and our procedure could be used in practice. In the poor matching scenario of that paper (first poor scenario of this paper), the Belin-Rubin procedure was unable to provide accurate estimates of error rates but our theoretical adjustment procedure still worked well. This indicated that we either had to find an enhancement to the Belin-Rubin procedures or to develop methods that used more of the available data. (That conclusion, incidentally, from our earlier workled, after some false starts, to the present approach.)

3.5 Quantitative Scenarios

Having specified the above linkage situations, we used SAS to generate ordinary least squares data under the model $Y = 6X + \varepsilon$. The X values were chosen to be uniformly distributed between 1 and 101. The error terms, are normal and homoscedastic with variances 13,000, 36,000, and 125,000, respectively. The resulting regressions of Y on X have R^2 values in the true matched population of 70%, 47%, and 20%, respectively. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. To make modeling and analysis even more difficult in the high file overlap scenario, we used all false matches and only 5% of the true matches; in the medium file overlap scenario, we used all false matches and only 25% of true matches. (Note: Here to heighten the visual effect, we have introduced another random sampling step, so the reader can "see"

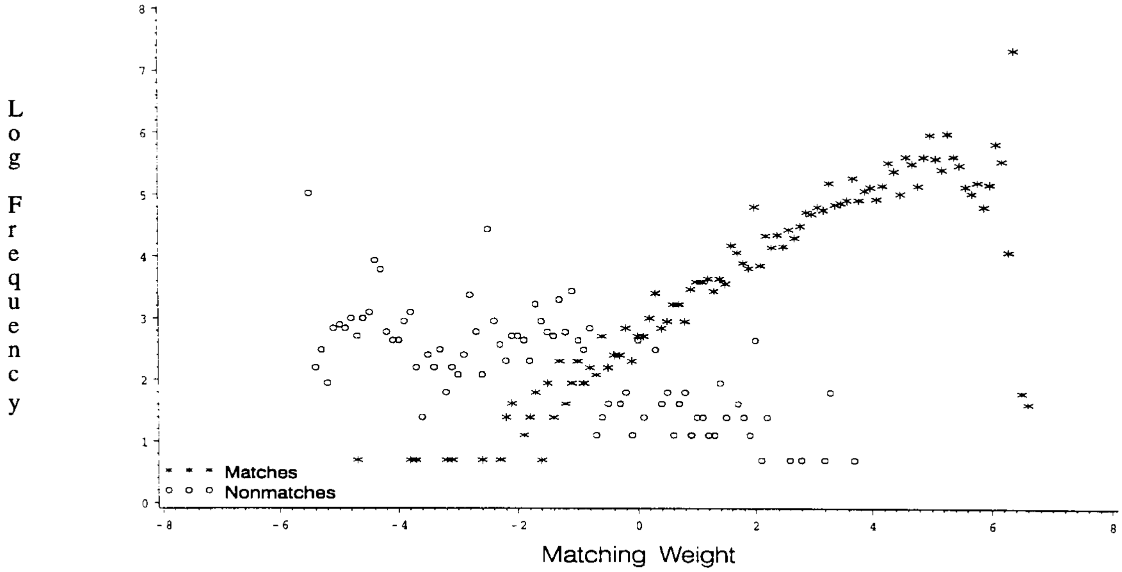

Figure 1a. 1st Poor Matching Scenario

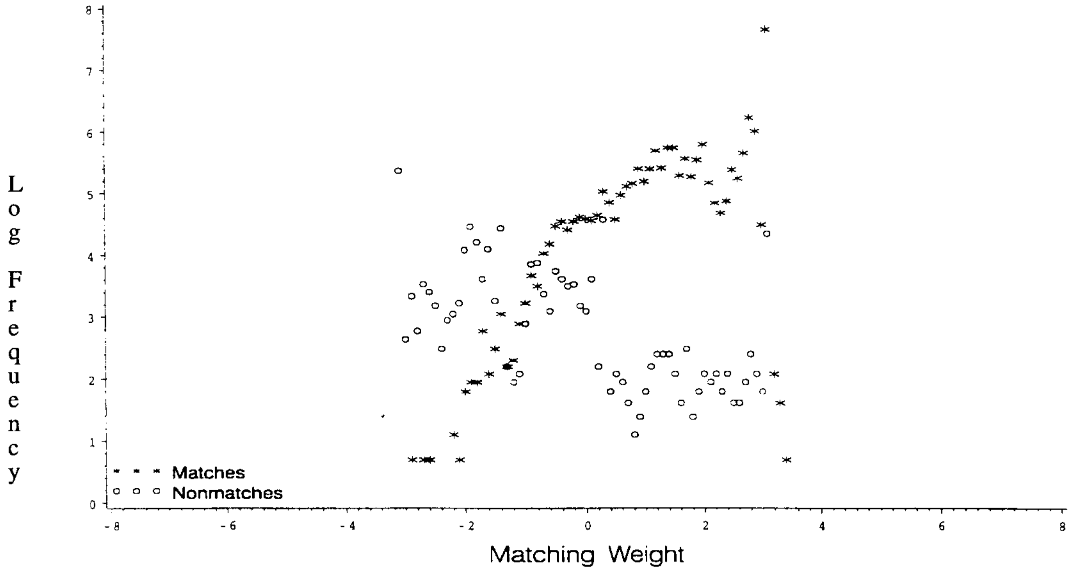

Figure 1b. 2nd Poor Matching Scenario

better in the figures the effect of bad matching. This sample depends on the match status of the case and is confined only to those cases that were matched, whether correctly or falsely.)

A crucial practical assumption for the work of this paper is that analysts are able to produce a reasonable model (guesstimate) for the relationships between the noncommon quantitative items. For the initial modeling in the empirical example of this paper, we use the subset of pairs for which matching weight is high and the error-rate is low. Thus, the number of false matches in the subset is kept to a minimum. Although neither the procedure of Belin and Rubin (1995) nor an alternative procedure of Winkler (1994), that requires an *ad hoc* intervention, could be used to estimate error rates, we believe it is possible for an experienced matcher to pick out a low-error-rate set of pairs even in the second poor scenario.

4. SIMULATION RESULTS

Most of this Section is devoted to presenting graphs and results of the overall process for the second poor scenario, where the R^2 value is moderate, and the intersection between the two files is high. These results best illustrate the procedures of this paper. At the end of the Section (in subsection 4.8), we summarize results over all R^2 situations and all overlaps. To make the modeling more difficult and show the power of the analytic linking methods, we use all false matches and a random sample of only 5% of the true matches. We only consider pairs having matching weight above a lower bound that we determine based on analytic considerations and experience. For the pairs of our analysis, the restriction causes the number of false matches to significantly exceed the number of true matches. (Again, this is done to heighten the visual effect of matching failures and to make the problem even more difficult.)

To illustrate the data situation and the modeling approach, we provide triples of plots. The first plot in the triple shows the true data situation as if each record in one file was linked with its true corresponding record in the other file. The quantitative data pairs correspond to the truth. In the second plot, we show the observed data. Where many of the pairs are in error because they correspond to false matches. To get to the third plot in the triple, we model using a small number of pairs (approximately 100) and then replace outliers with pairs in which the observed Y-value is replaced with a predicted Y-value.

4.1 Initial True Regression Relationship

In Figure 2a, the actual true regression relationship and related scatterplot are shown, for one of our simulations, as they would appear if there were no matching errors. In this figure and the remaining ones, the true regression line is always given for reference. Finally, the true population slope or *beta* coefficient (at 5.85) and the R^2 value (at 43%) are provided for the data (sample of pairs) being displayed.

4.2 Regression After Initial RL→RA Step

In Figure 2b, we are looking at the regression on the actual observed links – not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between Y and X. The observed slope or *beta* coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected – falling to 7% from 43%.

4.3 Regression After First Combined RL→RA→EI→RA Step

Figure 2c completes our display of the first cycle of the iterative process we are employing. Here we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00 or larger, an attempt was made to improve the poor results given in Figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 460 or more were removed and the regression refit on the remaining pairs. This new equation, used in Figure 2c, was essentially $Y = 4.78X + \varepsilon$, with a variance of 40,000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the estimated *beta* coefficient from 4.78 to 5.4. If a pair of matched records yielded an outlier, then predicted values (not shown) using the equation $Y = 5.4X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.4 Second True Reference Regression

Figure 3a displays a scatterplot of X and Y as they would appear if they could be true matches based on a second RL step. Note here that we have a somewhat different set of linked pairs this time from earlier, because we have used the regression results to help in the linkage. In particular, the second RL step employed the predicted Y values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second RL step. Since a considerably better link was obtained, there were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1,104 in Figures 2a through 2c to 650 for Figures 3a through 3c. In this second iteration, the true slope or *beta* coefficient and the R^2 values remained, though, virtually identical for the estimated slope (5.85 v. 5.91) and fit (43% v. 48%).

4.5 Regression After Second RL→RA Step

In Figure 3b, we see a considerable improvement in the relationship between Y and X using the actual observed links after the second RL step. The estimated slope has risen from 2.47 initially to 4.75 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 33%.

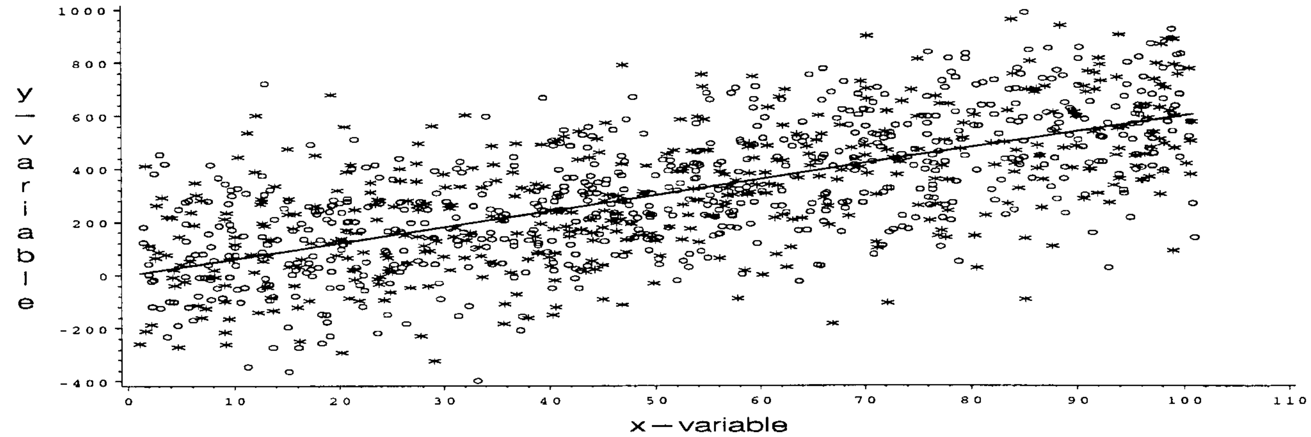

Figure 2a. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, True Data, HighOverlap,
1104 Points, beta = 5.85, R-square = 0.43

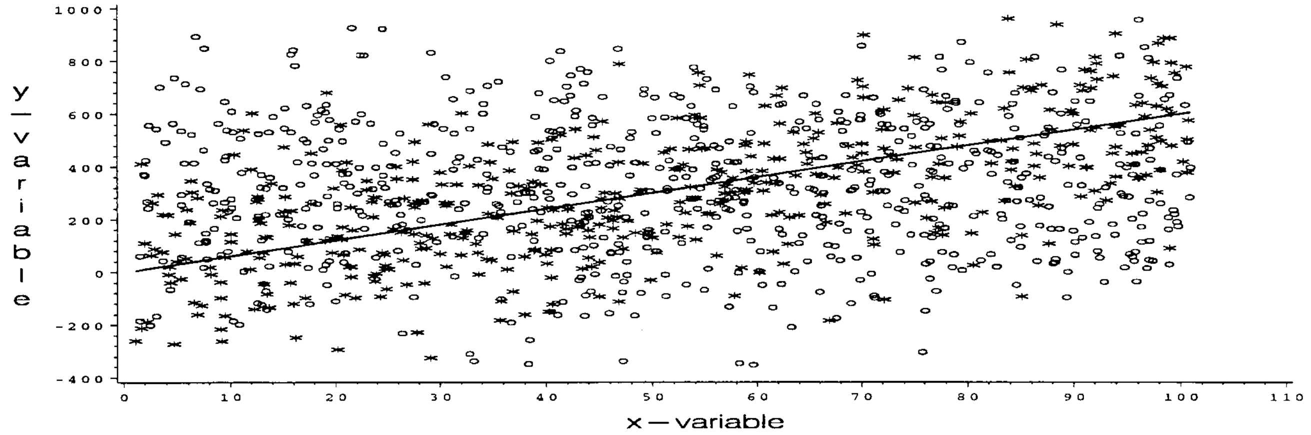

Figure 2b. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Observed Data, HighOverlap,
1104 Points, beta = 2.47, R – square = 0.07

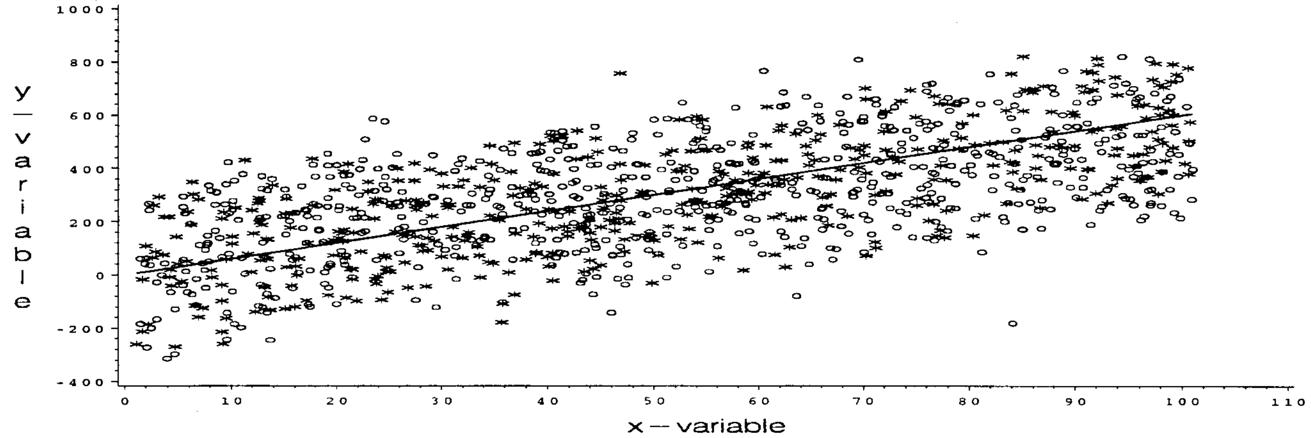

Figure 2c. 2nd Poor Scenario, 1st Pass
All False & 5 % True Matches, Outlier – Adjusted Data
1104 Points, beta = 4.78, R – square = 0.40

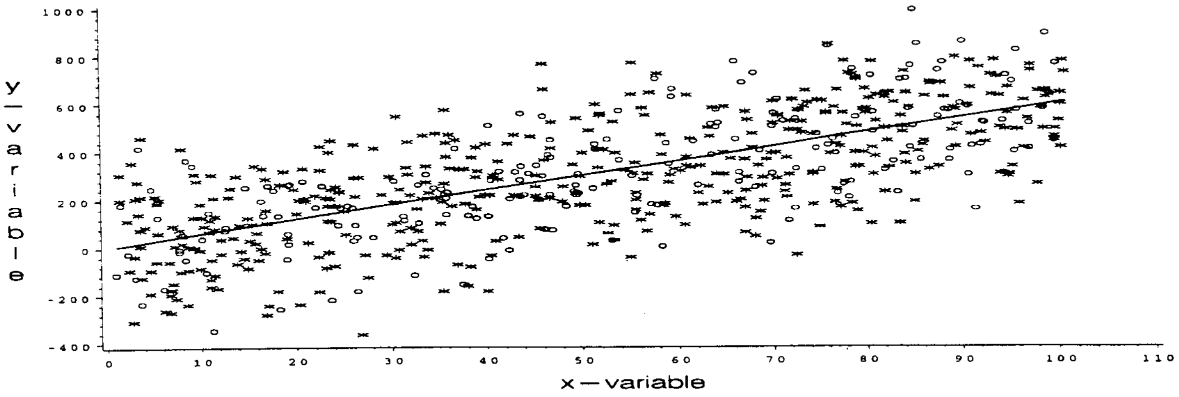

Figure 3a. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, True Data, HighOverlap,
650 Points, beta = 5.91 R – square = 0.48

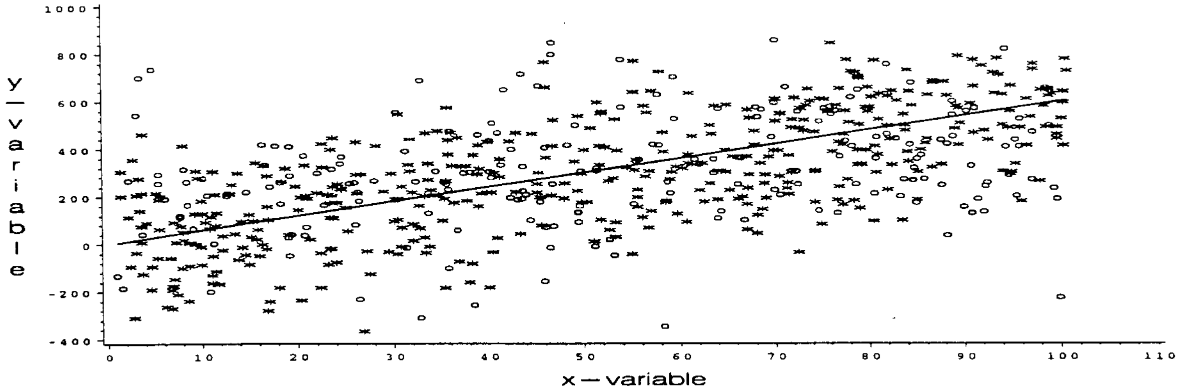

Figure 3b. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Observed Data, HighOverlap
650 Points, beta = 4.75, R – square = 0.33

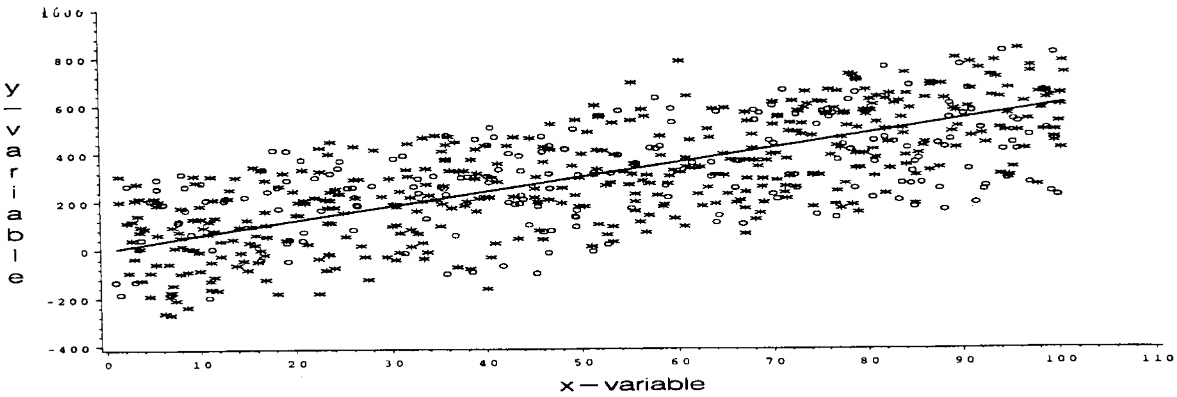

Figure 3c. 2nd Poor Scenario, 2nd Pass
All False & 5 % True Matches, Outlier – Adjusted Data
650 Points, beta = 5.26, R – square = 0.47

4.6 Regression After Second Combined RL→RA→EI→RA Step

Figure 3c completes the display of the second cycle of our iterative process. Here we have edited the data as follows. Using the fit (from subsection 4.5), another set of predicted values was obtained for all the matched cases (as in subsection 4.3). This new equation was essentially $Y = 5.26X + \varepsilon$, with a variance of about 35,000. If a pair of matched records yields an outlier, then predicted values using the equation $Y = 5.3X$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.7 Additional Iterations

While we did not show it in this paper, we did iterate through a third matching pass. The *beta* coefficient, after adjustment, did not change much. We do not conclude from this that asymptotic unbiasedness exists; rather that the method, as it has evolved so far, has a positive benefit and that this benefit may be quickly reached.

4.8 Further Results

Our further results are of two kinds. We looked first at what happened in the medium R^2 scenario (*i.e.*, R^2 equal to .47) for the medium- and low- file intersection situations. We further looked at the cases when R^2 was higher (at .70) or lower (at .20). For the medium R^2 scenario and low intersection case the matching was somewhat easier. This occurs because there were significantly fewer false-match candidates and we could more easily separate true matches from false matches. For the high R^2 scenarios, the modeling and matching were also more straightforward than they were for the medium R^2 scenario. Hence, there were no new issues there either.

On the other hand, for the low R^2 scenario, no matter what degree of file intersection existed, we were unable to distinguish true matches from false matches, even with the improved methods we are using. The reason for this, we believe, is that there are many outliers associated with the true matches. We can no longer assume, therefore, that a moderately higher percentage of the outliers in the regression model are due to false matches. In fact, with each true match that is associated with an outlier Y-value, there may be many false matches that have Y-values that are closer to the predicted Y-value than the true match.

5. COMMENTS AND FUTURE STUDY

5.1 Overall Summary

In this paper, we have looked at a very restricted analysis setting: a simple regression of one quantitative dependent variable from one file matched to a single quantitative independent variable from another file. This standard analysis was, however, approached in a very nonstandard setting. The matching scenarios, in fact, were quite challenging. Indeed, just a few years ago, we might have said that the "second poor" matching scenario appeared hopeless.

On the other hand, as discussed below, there are many loose ends. Hence, the demonstration given here can be considered, quite rightly in our view, as a limited accomplishment. But make no mistake about it, we are doing something entirely new. In past record linkage applications, there was a clear separation between the identifying data and the analysis data. Here, we have used a regression analysis to improve the linkage and the improved linkage to improve the analysis and so on.

Earlier, in our 1993 paper, we advocated that there be a unified approach between the linkage and the analysis. At that point, though, we were only ready to propose that the linkage probabilities be used in the analysis to correct for the failures to complete the matching step satisfactorily. This paper is the first to propose a completely unified methodology and to demonstrate how it might be carried out.

5.2 Planned Application

We expect that the first applications of our new methods will be with large business data bases. In such situations, noncommon quantitative data are often moderately or highly correlated and the quantitative variables (both predicted and observed) can have great distinguishing power for linkage, especially when combined with name information and geographic information, such as a postal (*e.g.*, ZIP) code.

A second observation is also worth making about our results. The work done here points strongly to the need to improve some of the now routine practices for protecting public use files from reidentification. In fact, it turns out that in some settings – even after quantitative data have been confidentiality protected (by conventional methods) and without any directly identifying variables present – the methods in this paper can be successful in reidentifying a substantial fraction of records thought to be reasonably secure from this risk (as predicted in Scheuren 1995). For examples, see Winkler (1997).

5.3 Expected Extensions

What happens when our results are generalized to the multiple regression case? We are working on this now and results are starting to emerge which have given us insight into where further research is required. We speculate that the degree of underlying association R^2 will continue to be the dominant element in whether a usable analysis is possible.

There is also the case of multivariate regression. This problem is harder and will be more of a challenge. Simple multivariate extensions of the univariate comparison of Y values in this paper have not worked as well as we would like. For this setting, perhaps, variants and extensions of Little and Rubin (1987, Chapters 6 and 8) will prove to be a good starting point

5.4 "Limited Accomplishment"

Until now an analysis based on the second poor scenario would not have been even remotely sensible. For this reason alone we should be happy with our results. A closer examination, though, shows a number of places where the approach demonstrated is weaker than it needs to be or simply unfinished. For those who want theorems proven, this may be a particularly strong sentiment. For example, a convergence proof is among the important loose ends to be dealt with, even in the simple regression setting. A practical demonstration of our approach with more than two matched files also is necessary, albeit this appears to be more straightforward.

5.5 Guiding Practice

We have no ready advise for those who may attempt what we have done. Our own experience, at this point, is insufficient for us to offer ideas on how to guide practice, except the usual extra caution that goes with any new application. Maybe, after our own efforts and those of others have matured, we can offer more.

REFERENCES

ALVEY, W., and JAMERSON, B. (Eds.) (1997). *Record Linkage Techniques – 1997.* Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, Arlington, VA.

BELIN, T.R., and RUBIN, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

FELLEGI, I., and HOLT, T. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

FELLEGI, I., and SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

JABINE, T.B., and SCHEUREN, F. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 89, 414-420.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis With Missing Data.* New York: John Wiley.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science,* 130, 954-959.

NEWCOMBE, H., FAIR, M., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.

OH, H.L., and SCHEUREN, F. (1975). Fiddling around with mismatches and nonmatches. *Proceedings of the Social Statistics Section, American Statistical Association.*

SCHEUREN, F. (1995). Review of private lives and public policies: Confidentiality and accessibility of government services. *Journal of the American Statistical Association*, 90, 386-387.

SCHEUREN, F., and WINKLER, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

WINKLER, W.E. (1994). Advanced methods of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.

WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox *et al.*). New York: John Wiley, 355-384.

WINKLER, W.E., and SCHEUREN, F. (1995). Linking data to create information. *Proceedings: Symposium 95, From Data to Information-Methods and Systems,* Statistics Canada, 29-37.

WINKLER, W.E., and SCHEUREN, F. (1996). Recursive analysis of linked data files. *Proceedings of the 1996 Annual Research Conference.* U.S. Bureau of the Census.

WINKLER, W.E. (1997). Producing Public-Use Microdata That are Analytically Valid and Confidential. Presented at the 1997 Joint Statistical Meetings, Anaheim, CA.