Error Model for Analysis of Computer Linked Files

William E. Winkler, Bureau of the Census, Rm 3000-4, Washington, DC 20233

1. INTRODUCTION

Better information for analyses supporting policy decisions can sometimes be produced by linking records from two computer data bases. An epidemiologist might wish to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death index that contains information about the cause and date of death. An economist might wish to evaluate energy policy decisions by matching a data base containing fuel and commodity information for a set of companies against a data base containing the values and types of goods produced by the companies.

If unique identifiers such as verified social security numbers are not available, then file matching may incorrectly link some records and resultant statistical analyses could be subject to error. Matching is subject to error because information such as company or individual name, address, age, and other descriptive information may not uniquely identify an individual. For instance, 'Smith' and 'Robert' may not uniquely identify an individual. Legitimate and typographical variations such as 'Mrs John', 'Elizabeth', 'Elzboth', and 'Beth' increase the difficulty of correctly identifying links. Use of address information is often subject to error because existing parsing and standardization software do not effectively allow comparison of, say, a house number with a house number and a street name with a street name. The addresses of an individual we wish to match may differ because one is erroneous or because the individual has moved.

Fellegi and Sunter (1969) presented a formal mathematical model and showed the optimality of decision rules in a record linkage strategy. Pairs of records in a file are given a score. Those above a certain score are designated matches, those below a second, lower, score are designated nonmatches, and those with with scores between the higher and lower scores are held for clerical review. The scores, or computer matching weights, are based on a crucial likelihood ratio that is often difficult to estimate (see e.g., Winkler and Thibaudeau 1991, Belin and Rubin 1991).

With files of moderate size, several thousand pairs may need to be clerically reviewed. As such review often involves examining paper forms (if they exist) or use of additional data sources, it is expensive and subject to error. With large files, reviewing hundreds of thousands of pairs is likely to be prohibitively expensive.

This paper highlights methods introduced by Winkler and Scheuren (1991, also Scheuren and Winkler 1991) for adjusting certain analyses for matching error. The main purpose of the adjustment procedure is to reduce or eliminate the need for clerical review. Preliminary investigations have considered simple regression and loglinear models. At a minimum, the goals of the research are to tell us how much accuracy is improved via adjustment, whether estimates are sufficiently accurate for statistical analyses and policy decisions, and how much cost must be incurred (through targetted clerical review) to insure a given benefit in increased accuracy. For the bivariate situations that have so far been considered these goals are often met. The key to the adjustment procedure is estimating accurately the proportions of matches and nonmatches within a set of pairs for all ranges of scores. The method of estimating proportions of matches within weight ranges is due to Belin and Rubin (1991).

The inferential framework that best summarizes how well the adjustment procedures work on the average are Monte Carlo methods similar to Rubin's multiple imputation (see e.g., Rubin 1987, pp. 75-77). The intuitive idea of multiple imputation is that the structure of data relationships and the model under which we impute places restraints on the statistical estimates being considered. For nonresponse (Rubin 1987), the set of data values associated with respondents, the pattern of nonresponse, and the imputation model all effect multiply imputed parameter estimates and their variances. For this paper, what records from one file are matched with what records from another file, the data associated with the matched records, and the model for adjusting for matching error all effect the multiply imputed estimates.

The empirical data bases are constructed from two files for which true matching status of pairs is known. Very extensive review and verification of pairs was done to assure that matching status is accurate. For regression-related analyses, numerical data are constructed using known normal models so that the effects of matching error can be evaluated rather than the effects of messy regression data bases. For loglinear-related analyses, known uniformly generated pseudo-random variates with carefully induced dependency are considered. Different sets of seed numbers produce different samples. The outline for the remainder of the presentation is as follows. In the second section we present some of the theoretical background for adjustments to ordinary regression and the intuitive ideas of the loglinear adjustments. Scheuren and Winkler (1991) present the general theoretical results for regression. In the third section we present highlights of fairly extensive reg ression results and much more preliminary loglinear results. The fourth section consists of discussion and the final section mentions future research.

2. BACKGROUND

2.1. Theoretical Adjustment Model for Regression

This section provides a description of the regression framework and adjustment methodology for the simplest classes of univariate regression. The theory for general regression is given by Winkler and Scheuren (1991).

Let $Y = X \beta + \varepsilon$ be the ordinary univariate regression model for which error terms are independent with constant variance σ^2. If we were working with a single data base, Y would be regressed on X in the usual manner. For $i = 1, \cdots, N$, we wish to use (X_i,Y_i) but use (X_i,Z_i). Z_i is usually Y_i but may take some other value Y_j due to matching error.

For $i = 1, \cdots, N$,

$$Z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} \text{ for } j \neq i. \end{cases}$$

$$p_i + \sum_{j \neq i} q_{ij} = 1.$$

The probability p_i may be zero or one. We define $h_i = 1 - p_i$ and divide the set of pairs into N mutually exclusive classes. The classes are determined by records from one of the files. Each class consists of the independent x-variable X_i, the true value of the dependent y-variable, the values of the y-variables from records in the second file to which the record in the first file containing X_i have been paired, and computer matching weights. Some of the N classes may have zero matching weights By <u>paired</u> we mean two records from the two files that have been brought together during the record linkage process but for which no determination of matching status may have been made. Under an assumption of 1-1 matching, for each $i = 1, \cdots, N$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$.

To define regression properly, we need to find μ_z=E(z), σ_z^2, and σ_{zx}. We observe that

$$E(Z) = (1/N) \sum_i E(Z|i) = (1/N) \sum_i (Y_i\, p_i + \sum_{j \neq i} Y_j\, q_{ij})$$

$$= (1/N) \sum_i Y_i + (1/N) \sum_i [Y_i (-h_i) + Y_{\phi(i)}\, h_i] \qquad (1)$$

$$= \bar{Y} + B.$$

Similarly, we can represent σ_{zy} in terms of σ_{xy} and a bias term B_{xy} and σ_z^2 in terms of σ_y^2 and a bias term B_{zz}. We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data.

Different equations yield the adjustments that relate regression coefficients β_{zx} based on observed data with regression coefficients β_{yx} based on true values. Our assumption of 1-1 matching (which is not needed for the general theory) is done for computational tractability to reduce the number of records and amount of information that must be tracked during the matching process.

In implementing the adjustments, we make two crucial assumptions. The first is that, for $i = 1,\cdots, N$, we can accurately estimate the true probabilities of a match p_i. The second is that, for each $i = 1, \cdots, N$, the true value Y_i associated with independent variable X_i is the pair with the highest matching weight and the false value $Y_{\phi(i)}$ is associated with the second highest matching weight.

2.2. Idea for Adjustment Model for Loglinear

As with the regression model, we simplify by using 1-1 matching and consider pairs of the form (A_i,C_i), $i = 1, \cdots, N$, where

$$C_i = \begin{cases} B_i & \text{with probability } p_i \\ B_j & \text{with probability } h_i \text{ for some } j. \end{cases}$$

$$p_i + h_i = 1.$$

The adjustment procedure for loglinear models is similar to that given by (1) for regression models. To get the counts needed for the loglinear analysis, we need to adjust the counts associated with the observed values (A_i,C_i) back to the counts associated with true values (A_i,B_i). We do this by adding h_i to the cell count determined by the value of (A_i,B_i) and subtracting h_i from the cell count determined by the value of $(A_i,B_{\phi(i)}) = (A_i,B_j)$.

As an example, let the observed value of (A,C) be (1,3). If, consistent with our model, (A,C) takes value (1,3) with probability $p_i = 0.75$ and takes value (1,2) with probability $h_i = 0.25$, then we adjust the count in the cell (1,3) up by 0.25 and the count in cell (1,2) down by 0.25.

2.3. Empirical Data Bases

The empirical data bases are created from two files of 10,000 records having known matching status. Basic matching parameters (see e.g., Winkler and Thibaudeau 1991) are estimated that cause the curves of log frequencies versus matching weight for nonmatches and matches to separate (Figure 1). Matching probabilities are estimated using the Belin-Rubin methodology (Table 1). We see that the estimated probabilities agree quite closely in the tails (above weight 4 and below weight 2). For weight 3, the deviation is relatively large because the true proportion of false matches is 0.06 while the estimated one is 0.20

For the regression analysis, each unique record in the merged data files has an independent x-variable that is generated according to a uniform distribution between 1 and 101 and a dependent y-variable that is generated via with a random normal distribution such that the slope is 2 and the R-square value is approximately 0.45. Error arises because the observed (x,y)-pair that is normally used in computation has a y-value from a record to which the record containing the x-value was falsely matched.

Table 1. Probabilities and Counts of Matches and Nonmatches in Weight Ranges

	Count		Probability	
weight	Mat	NM	true	est
11	6950	0	.00	.00
10	785	0	.00	.00
9	610	0	.00	.00
8	439	3	.00	.00
7	250	4	.00	.01
6	265	9	.03	.03
5	167	8	.05	.06
4	89	6	.06	.11
3	84	5	.06	.20
2	38	7	.16	.31
1	33	34	.51	.46
0	13	19	.59	.61
-1	7	20	.74	.74
-2	3	11	.79	.84
-3	4	19	.83	.89
-4	0	15	.99	.94
-5	0	15	.99	.96
-6	0	27	.99	.98
-7	0	107	.99	.99

1/ In the first column, weight 10 means weight range from 10 to 11. Weight ranges 11 and above and -7 and below are added together separately. Mat is match and NM is nonmatch.

For the analysis of regression-related adjustments we consider only those pairs having matching weights between 0 and 10 because all pairs above weight 10 are true matches. Pairs between 0 and 10 contain both true and false matches. We do this to determine how much the adjustment improves the accuracy of the regression analyses in situations for which there is significant matching error. If we include pairs above weight 10, then it is more difficult to judge the adjustment process because ordinary regression estimates based on observed data and adjusted regression estimates will both be relatively more accurate.

For the adjustment related to the loglinear analysis, we use a uniform generator to create random variable A that takes values 1, 2, and 3 with probabilities 0.20, 0.60, and 0.20, resp. Random variable B is similarly created so that with probability 0.05 it takes the same values as A and with probability 0.95 it independently takes values 1, 2, and 3 in the same proportions that A assumed the values 1, 2, and 3. We, thus, induce slight dependence. Only thoses pairs having matching weights between 0 and 7 are included in the analysis.

In the remainder of the paper, whenever we use true, we will mean estimates based on the true values. Similarly, when we use observed, we mean estimates based on observed data. Adjusted will always refer to estimates obtained via the adjustment methods of this paper.

3. RESULTS

3.1. Regression

The results of using the adjustment process are illustrated in Figure 2. Figure 2a provides a comparison of the relative coefficients of variation of the adjusted procedure versus the nonadjusted procedure. To get the plotted points, the coefficients of variations (cvs) computed via either procedure are divided by the true cv for weight class 8. The results show that both adjusted and nonadjusted procedures yield approximately the same cv estimates and that cvs decrease as sample size increases. The relative bias of the cvs for the adjusted procedure is substantially lower than the relative bias for the nonadjusted procedure (Figure 2b). The nonadjusted procedure uses ordinary linear regression on the observed data pairs.

Multiply imputed estimates for 25 samples (Table 2) show the relative cv estimates for both adjusted and nonadjusted procedures are about the same while the higher bias of the nonadjusted procedure yields higher quasi root mean square errors (qmrse). The term qrmse is used because we use an estimate of the variance component of root mean square error rather than the true value. We observe that for higher weight ranges, say between 6 and 10, both the adjusted procedure and nonadjusted procedure produce about the same qmrses, 0.056 and 0.058, resp. As weight ranges having more erroneous data are included, say between 0 and 10, qrmse under the adjusted procedure, 0.048, is substantially lower than under the nonadjusted procedure, 0.081.

Table 2. Comparison of Estimates Averaged over 25 Samples Coefficient Estimates

wgt class	size	true	est	obs
8	442	2.020	2.018	2.004
cv		0.082	0.082	0.082
qrmse			0.082	0.082
6	970	2.015	2.002	1.976
cv		0.053	0.056	0.056
qrmse			0.056	0.058
4	1240	2.010	2.006	1.956
cv		0.046	0.048	0.049
qrmse			0.048	0.055
2	1374	2.005	2.025	1.940
cv		0.044	0.047	0.047
qrmse			0.049	0.056
0	1473	2.007	1.976	1.870
cv		0.042	0.046	0.046
qrmse			0.048	0.081

Note: Weight class 2 means those pairs having weight between 2 and 10.

3.2. Loglinear

The counts and fitted values from the pairs used in the loginear analysis are given in Tables 3a, 3b, and 3c for the true, observed, and adjusted cases, resp. The Pearson Chi-Square and Likelihood Ratio Chi-Square Values for the true case are 0.006 and 0.008, resp.; for the observed case, 0.053 and 0.059, resp.; and for the adjusted case, 0.018 and 0.020, resp. We would not reject the null hypothesis that A and B are independent at the 95 percent confidence level for the observed data. We would reject with the true and adjusted values.

Table 3a. Actual Counts and Fitted Estimates for True Case

	1	2	3	
1	46 32.8	89 103.5	39 37.7	174
2	74 81.5	275 257.0	83 93.5	432
3	23 28.7	87 90.5	42 32.8	152
	143	451	164	758

Table 3b. Actual Counts and Fitted Estimates for Observed Case

	1	2	3	
1	42 33.0	91 102.0	41 39.0	174
2	78 82.1	268 253.0	86 96.9	432
3	24 28.9	85 89.0	43 34.1	152
	144	444	170	758

Table 3c. Actual Counts and Fitted Estimates for Adjusted Case

	1	2	3	
1	41.9 32.9	91.9 102.7	40.2 38.4	174
2	79.0 81.5	270.5 255.0	82.5 95.5	432
3	22.2 28.7	85.0 89.7	44.8 33.6	152
	143.1	447.4	167.5	758

4. DISCUSSION

4.1. Information Tracked During Matching

Although the adjustment procedures for regression and loglinear models are quite different, the information from the matching process that is needed for the adjustment is always the same. To implement the procedures of this paper, for each record in one file, the two records to which it is matched with the two highest matching weights and their associated matching weights must be tracked. All other information needed for statistical analyses (for either the simple settings of this paper or more involved multivariate settings) can then be pulled from the corresponding records in the two files.

If the information needed for adjustments is not tracked at a record-specific level, then the adjustments of this paper cannot be performed. In those situations, at best, the linker can state that unknown, possibly large, biases may limit the usefulness of the linked data for statistical analysis and policy decisions.

We do note, however, that applying the matching program and Belin software for estimating probabilities is straightforward. Special software only needs be written for the particular statistical analyses being performed. If the quality of the matching is sufficiently high (e.g., Scheuren and Winkler 1991), then the usual statistical analysis procedures can be applied. No adjustments are needed.

4.2. Average Improvment in Regression

With a few samples, using data created via the adjustment procedures of this paper was worse than using the observed data. As the multiple imputation procedures showed, on the average the adjustment procedures of this paper yielded improvements. At present, there has been little investigation of whether it is possible to determine better if the adjustments are causing improvements in individual cases.

4.3. Limitations for Regression

Based on close to 1000 simulations (Winkler and Scheuren 1991) for simple regression models, the regression adjustments appear to improve accuracy on the average. Due to the difficulty of the programming involved, no exploration of regression models having two or more independent variables or for which contamination (such as outliers and mild colinearity) have yet been investigated.

4.4. Limitations on Loglinear

At this point, it is too early to determine whether the adjustment procedure in loglinear models is effective. While the adjustment appeared to work well for a variety of simple situations (not presented) similar to the one presented in this paper, extensions of empirical results to higher dimensional contingency tables has been erratic. While the programming of adjustment procedures for higher dimensional loglinear models is much easier than for regression models, the generation of three or more dimensional data sets having carefully controlled departures from independence (such as those requiring two-way-interaction models) is difficult.

For very simple three-dimensional models, the adjustment procedures worked well in some cases and, in others, performed poorly. Due to space limitations, the three-dimensional models are not presented.

5. FUTURE WORK

Considerably more empirical work is needed to determine in what situations, if any, the adjustment procedures of this paper can be used effectively. A formal theory of the adjustment procedures for loglinear models is being investigated.

This paper reflects views of the author and not necessarily those of the Census Bureau.

REFERENCES

Belin, T. and Rubin, D. (1991) "Recent Developments in Calibrating Error Rates for Computer Matching," 1991 Census Bureau Annual Research Conference, to appear.

Fellegi, I. and Sunter, A. (1969) "A Theory of Record Linkage," J. Amer. Stat. Assn. 1183-1210.

Rubin, D. B. (1987) Multiple Imputation for Nonresponse in Surveys, New York: J. Wiley.

Scheuren, F. and Winkler, W. E. (1991) "Regression Analysis of Data Files that are Computer Matched," 1991 Census Bureau Annual Research Conference

Winkler, W. E. and Scheuren, F. (1991) "How Matching Error Effects Regression Analysis -- Exploratory and Confirmatory Results," technical report.

Winkler, W. E. and Thibaudeau, Y. (1991) "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," Bureau of the Census, Statistical Research Division Technical Report 91/09.

Figure 1. Log of Frequency vs Weight Matches & Nonmatches

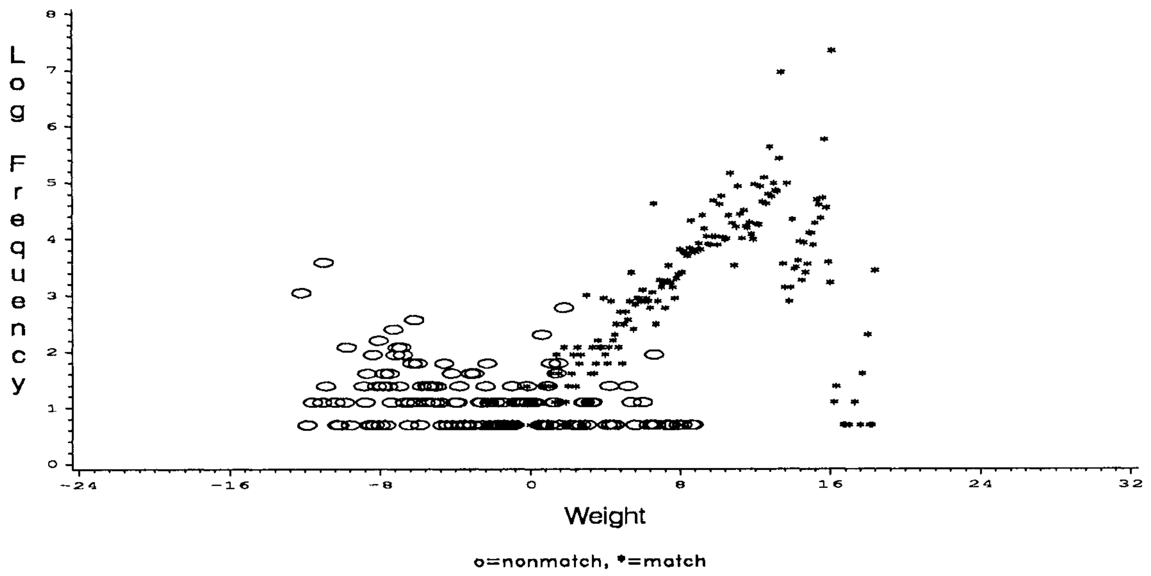
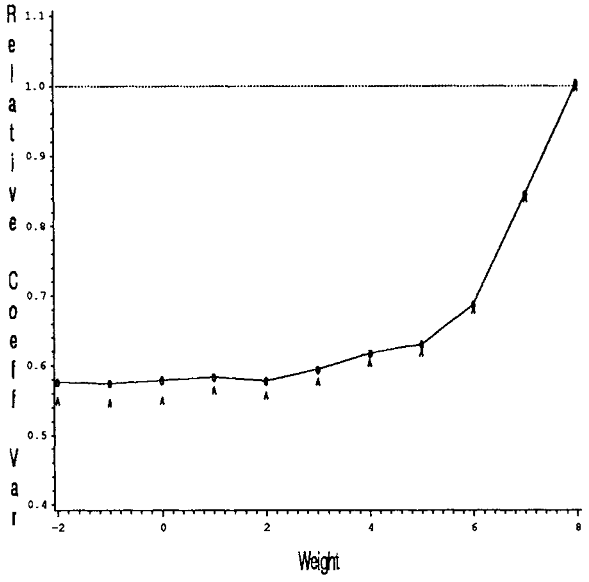
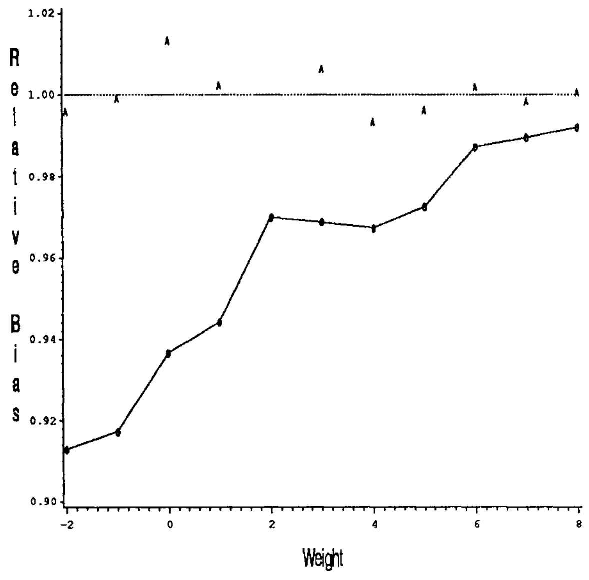


Figure 2a. Relative Coeff Var vs Weight, Estimated Probabilities R-square=0.45



A=adjusted, o=observed
range of true cvs (0.029 , 0.054)

Figure 2b. Relative Bias vs Weight, Estimated Probabilities R-square=0.45



A=adjusted, o=observed
range of true coeff estimates (1.86 , 1.95)