Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Regression Analysis With Linked Data

P Lahiri & Michael D Larsen

To cite this article: P Lahiri & Michael D Larsen (2005) Regression Analysis With Linked Data, Journal of the American Statistical Association, 100:469, 222-230, DOI: 10.1198/016214504000001277

To link to this article: https://doi.org/10.1198/016214504000001277

Regression Analysis With Linked Data

P. LAHIRI and Michael D. LARSEN

Record linkage, or exact matching, can be used to join together two files that contain information on the same individuals but lack unique personal identification codes. The possibility of errors in linkage causes problems for estimating the relationships between variables on the two files. The effect is analogous to the impact of measurement error. A model of a linear regression relationship between variables in linked files is proposed. Assuming the probabilities that pairs of records are links are known, an unbiased estimator of the regression coefficients is derived. Methods for estimating the linkage probabilities by using mixture models are discussed. A consistent estimator of the covariance matrix of the proposed estimator is proposed. A bootstrap estimator is used to reflect the impact of the uncertainty in record linkage model parameters on the estimators of the regression parameters. A simulation study compares the performance of the proposed estimator and alternatives.

KEY WORDS: Fellegi–Sunter; File matching; Latent class; Measurement error; Mixture model; Propagation of error; Record linkage.

1. INTRODUCTION

A goal of record linkage is to join together two files that contain information on the same individuals but lack unique personal identification codes. Computerized record linkage (CRL) methods are used in many federal statistical systems (Alvey and Jamerson 1997) and often in medical studies (Newcombe 1988), in which the databases are very large and processing time and accuracy are concerns. Sophisticated software has been developed for large applications by organizations including Statistics Canada (CANLINK software), the U.S. Census Bureau (Winkler 1994, 1995; Jaro 1989, 1995), and the Oxford Medical Record Linkage Study (Gill 1997). Because CRL utilizes already existing databases, it enables new statistical analyses without the substantial time and resources needed to collect new data.

Fellegi and Sunter (1969), formalizing ideas of Newcombe, Kennedy, Axford, and James (1959), proposed a model for record linkage. In the Fellegi–Sunter (1969) model, the two files being compared are called file A and file B. The set of pairs of records $A \times B = \{(a, b), a \in A, b \in B\}$ is composed of two disjoint subsets, the set of true links, M, and the set of true nonlinks, U. Most CRL software attaches weights, similar in nature to weights described by Fellegi and Sunter (1969), reflecting the likelihood that a pair of records, one from each of the two files, corresponds to the same subject.

Mixture models are useful when the population being studied is composed of two or more subpopulations that are not clearly identified (McLachlan and Peel 2000). In the case of record linkage, before clerical review has been completed and in the absence of unique identifying information, the status of pairs as true links and true nonlinks is unknown, but real. Before clerical review is undertaken, mixture models can be applied to measurements of the similarity among pairs of records to estimate probabilities used in calculating record linkage weights. In some applications (Larsen and Rubin 2001; Winkler 1988, 1994, 1995; Jaro 1989, 1995), the mixture classes correspond very closely to the sets of true links and true nonlinks.

If mismatch errors are introduced by CRL, then statistical analyses based on linked data can be adversely affected. Neter, Maynes, and Ramanathan (1965) studied the effect of mismatch errors in finite population sampling and observed that relatively small mismatch error could lead to a substantial bias in estimating the relationship between response errors and true values. Scheuren and Winkler (1993) (henceforth referred to as SW) investigated the effect of mismatch errors on the bias of ordinary least squares estimators of regression coefficients in a standard regression model and proposed a method of adjusting for the bias. SW (1997) advanced the work further with an iterative procedure that modified the regression and matching results for apparent outliers (see also Scheuren and Winkler 1991).

In this article we consider an alternative to the bias correction method of SW (1993). For known linkage probabilities, SW obtained their estimator of regression coefficients by adjusting the bias of the ordinary least squares estimator for the regression model with mismatch errors, whereas our proposed method provides an unbiased estimator directly for a transformed regression model. In Section 2 we describe the record linkage problem and model. In Section 3 we consider the use of mixture models for estimating relevant linkage probabilities and three implementation issues. In Section 4 we review the SW method and then propose a new method of estimating regression coefficients in the presence of mismatch errors. In Section 5 we propose a variance estimator for our regression estimator. We also discuss a bootstrap addition to the variance estimator to account for uncertainty in mixture model parameters. We present simulation results in Section 6. Our method improves on a naive method, a robust method, and the SW (1993) method. We defer technical proofs to the Appendix.

2. RECORD LINKAGE

In record linkage, the records in two files are compared with one another using available information, which typically does not include unique, error-free personal codes. Individuals can be compared on surname, first name, age or date of birth, and other variables. Some of these matching variables carry a lot of information for identifying individuals, whereas others (e.g., race or sex) contain very little. Some comparisons, however, are useful for discriminating between certain people, such as

P. Lahiri is Professor, Joint Program in Survey Methodology, University of Maryland, College Park, MD 20742 (E-mail: *plahiri@survey.umd.edu*). Michael D. Larsen is Assistant Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *larsen@iastate.edu*). Lahiri's work was supported in part by National Science Foundation grant SES-9978145 and a grant from the Gallup Organization. Larsen's work was supported in part by the U.S. Bureau of the Census through Census Contract No. 50-YABC-7-66021 under Census Task Order #46-YABC-8-00004. The authors thank the editor, an anonymous associate editor, and two referees for constructive comments that led to a substantial improvement of an earlier version of the article. Dr. Lahiri would like to thank the Harvard University Department of Statistics for hosting him in May of 1998 to visit Dr. Larsen.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001277

individuals living in the same household. Information can be missing or recorded with typographical or spelling errors.

The comparisons made on available fields of information result in measurements of agreement between the records in the two files. The outcome of agreement versus disagreement or of the level of correspondence measured in some manner (see, e.g., Winkler 1990) on comparison k is stored in comparison variable γ_k. In the simplified case of dichotomous agreement/disagreement outcomes, let $\gamma_k = 1$ if the pair agrees on comparison k and 0 otherwise. The set of K comparisons creates a comparison vector, $\boldsymbol{\gamma} = (\gamma_k, k = 1, \ldots, K)$ for each pair of records. For the special case of three available fields (i.e., $K = 3$), the possible comparison vectors are $(0, 0, 0)$ (i.e., all disagreements), $(0, 0, 1)$, $(0, 1, 0)$, $(0, 1, 1)$, $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$, and $(1, 1, 1)$ (i.e., all agreements).

The pattern of 0's and 1's in a comparison vector $\boldsymbol{\gamma}$ contains information about whether the pair is a true link or true nonlink. True links tend to have more agreements than true nonlinks. If Γ is the space of all comparison vectors $\boldsymbol{\gamma}$ and the probabilities of seeing a vector $\boldsymbol{\gamma}$ among true links and true nonlinks are known, then Fellegi and Sunter's (1969) decision rule for designating pairs as links and nonlinks is based on the ratio

$$R = P(\boldsymbol{\gamma} \in \Gamma | M)/P(\boldsymbol{\gamma} \in \Gamma | U).$$

Intuitively, if R is large, then pairs should be designated as links. On the other hand, if R is small, then pairs should be called nonlinks. Some values of R, however, are moderate and do not clearly suggest link or nonlink. In practice, pairs with moderate values of R can be sent to clerks for review or can be subjected to further comparisons. Fellegi and Sunter (1969) showed that at prespecified error levels for false links and false nonlinks, optimal cutoffs can be determined. The cutoffs are optimal in that they minimize the set of pairs that are sent to clerical review for deciding link status at prespecified error levels. The decision rule can be characterized as follows:

- If $R \geq upper$, then designate the pair as a link.
- If $upper > R > lower$, then postpone the decision pending clerical review.
- If $R \leq lower$, then designate the pair as a nonlink.

Three issues arise in practice. First, not all possible pairs of records are compared. Instead, pairs are compared within blocks of records that are similar in terms of basic characteristics, such as geography or first letter of last name. Forming blocks, or "blocking," greatly reduces the number of pairs compared. If individuals rarely move between blocks, then few true links are lost by implicitly treating all pairs excluded by blocking as nonlinks.

Second, probabilities of comparison vectors by link status, $P(\boldsymbol{\gamma}|M)$ and $P(\boldsymbol{\gamma}|U)$, are not known; they must be estimated under a model using certain assumptions. Given prespecified error rates and estimates of these probabilities, the Fellegi–Sunter (1969) method determines corresponding values of *upper* and *lower*. The performance of the procedure in terms of actual versus specified error rates is sensitive to estimates of probabilities and choice of *upper* and *lower* (Belin 1993; Belin and Rubin 1995). In Section 3 we describe the use of mixture models for estimating these probabilities.

Third, for a record in file A there might be several candidate links within a particular block in file B. We assume in this work that only one of the records in file B is a true link for the record in file A. Given estimated probabilities, in practice, single links for individual records are chosen according to some procedure. Many applications, such as those at the U.S. Census Bureau (e.g., Jaro 1989) use a one-to-one, linear-sum assignment procedure (Burkard and Derigs 1980) to choose individual links. The one-to-one assignment procedure can effectively eliminate many candidate links that have some degree of similarity but actually are nonlinks. On the other hand, forcing one-to-one matching could remove the true link if one member of the record pair has a better matching record in the other file. The possibility of false matches and false nonmatches has serious implications in many record linkage applications, such as counterterrorism (Gomatam and Larsen 2004).

3. MIXTURE MODELS

Let G be the number of subpopulations. In our application, we have $G = 2$ subpopulations, one consisting of links and the other consisting of nonlinks. The comparison vector $\boldsymbol{\gamma}$ is assumed to follow a finite mixture model with probability mass function given by

$$P(\boldsymbol{\gamma}) = \sum_{g=1}^{G} \pi_g P(\boldsymbol{\gamma} | \text{class } g),$$

where π_g is the probability that a pair of records belongs to the mixture class g and $P(\boldsymbol{\gamma}|\text{class } g)$ is the probability mass function of the comparison vector in class g.

The model in each mixture class makes simplifying assumptions about the relationship between fields of comparison. For example, a common assumption suggested by Fellegi and Sunter (1969) is to assume the fields of comparison are independent within a given class. If there are K comparison fields that are conditionally independent, then in class g the probability of observing a comparison vector is

$$P(\boldsymbol{\gamma} | \text{class } g) = \prod_{k=1}^{K} P(\gamma_k | \text{class } g),$$

where $P(\gamma_k|\text{class } g)$ is the probability of outcome γ_k on comparison k in class g. Other modeling assumptions are possible and, in some cases, correspond better to the observed data (Larsen and Rubin 2001; Armstrong and Mayda 1993; Thibaudeau 1993). A few authors in other contexts have used mixture models applied to discrete data with modeling assumptions other than conditional independence (see, e.g., Becker and Yang 1998; Larsen and Rubin 2001).

The parameters of the mixture model can be estimated using the expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) and expectation-conditional maximization (ECM) (Meng and Rubin 1993) algorithms. Several authors, including Larsen and Rubin (2001), have implemented these algorithms for the purposes of record linkage. The estimated probability that the pair that produced the comparison vector $\boldsymbol{\gamma}$ belongs to class g is, by Bayes's theorem,

$$P(\text{class } g | \boldsymbol{\gamma}) = \pi_g P(\boldsymbol{\gamma} | \text{class } g) \Big/ \sum_{h=1}^{G} \pi_h P(\boldsymbol{\gamma} | \text{class } h).$$

As described by Larsen and Rubin (2001), the estimated probabilities can be used to partition the record pairs into designated links and nonlinks and to estimate error rates. Larsen and Rubin (2001) studied model selection and used partial clerical review to improve estimation.

Certain logical inequality constraints can be incorporated into the estimation algorithms. For example, the estimated size of the mixture class should be less than the size of the smaller of the two files, A or B. Additionally, the probability of agreeing on a comparison should be higher among links than among nonlinks. These inequality constraints are implemented as part of the estimation in the simulation. Winkler (1993) has considered more extensive restrictions on parameters and estimated probabilities.

4. ESTIMATION OF REGRESSION COEFFICIENTS

After file A and file B have been linked together, it might be of interest to analyze the relationship between a response variable (y) that was originally in file A and a set of covariates ($\mathbf{x}$) that were originally in file B. Neter et al. (1965), SW (1993, 1997), Larsen (1999, 2001), and Lahiri and Larsen (2000), (henceforth referred to as LL), discussed this problem. Table 1 illustrates the structure of the files. In the illustration, the matching variables v_k and w_k, $k = 1, \ldots, K$, are used for comparison of record pairs. These two files are linked using a CRL technique. Thus the true data pairs $(\mathbf{x}_i, y_i)$ are not observable. Instead, the CRL procedure produces pairs $(\mathbf{x}_i, z_i)$ in which z_i may or may not correspond to y_i.

Consider the following regression model for $\mathbf{y} = (y_1, \ldots, y_n)'$:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is a column vector of p known covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a column vector of p unknown regression coefficients, $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$, and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j, i, j = 1, \ldots, n$. SW (1993) considered the following model for $\mathbf{z} = (z_1, \ldots, z_n)'$ given y:

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i, j = 1, \ldots, n, \end{cases} \tag{2}$$

where $\sum_{j=1}^{n} q_{ij} = 1, i = 1, \ldots, n$. Define $\mathbf{q}_i = (q_{i1}, \ldots, q_{in})'$, $i = 1, \ldots, n$, and $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)'$. The naive least squares estimator of $\boldsymbol{\beta}$, which ignores mismatch errors, is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{N}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is an $n \times p$ matrix. An alternative to this naive estimator would be to use a robust estimator such as an estimator that minimizes the sum of absolute deviations (Press, Teukolsky, Vetterling, and Flannery 1992, p. 698). Robust estimators that decrease the influence of outliers should decrease the impact of erroneously paired predictor and response values.

First, note that under the model described by (1) and (2),

$$\mathrm{E}(z_i) = \mathbf{w}_i'\boldsymbol{\beta},$$

where $\mathbf{w}_i = \mathbf{q}_i'\mathbf{X} = \sum_{j=1}^{n} q_{ij}\mathbf{x}_j'$, $i = 1, \ldots, n$, is a $p \times 1$ column vector (see Thm. A.1). Thus bias of $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ is given by

$$\begin{aligned} \mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) &= \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta}] \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W} - \mathbf{I}]\boldsymbol{\beta} \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{X} - \mathbf{I}]\boldsymbol{\beta}, \end{aligned}$$

where $\mathbf{I}$ is an identity matrix of dimension p and $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)'$. Thus in general, $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ is not an unbiased estimator of $\boldsymbol{\beta}$ and the magnitude of the bias depends on $\mathbf{W}$ and $\boldsymbol{\beta}$. Evidently, in the special case of no mismatch error, that is, when $q_{ii} = 1, i = 1, \ldots, n$, $\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}) = 0$.

In order to reduce the bias of $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$, SW (1993) first investigated the bias of $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ conditional on the values of $\mathbf{y}$. It can be seen from the calculations of SW that

$$\mathrm{bias}(\hat{\boldsymbol{\beta}}_{\mathrm{N}}|\mathbf{y}) = \mathrm{E}[(\hat{\boldsymbol{\beta}}_{\mathrm{N}} - \boldsymbol{\beta})|\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}, \tag{3}$$

where $\mathbf{B} = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij}y_j = \mathbf{q}_i'\mathbf{y} - y_i$. Thus B_i is the difference between a weighted average of responses from all observations and the actual response y_i. If an estimator of $\mathbf{B}$, say $\hat{\mathbf{B}}$, is available, then the SW estimator is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{SW}} = \hat{\boldsymbol{\beta}}_{\mathrm{N}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{B}}. \tag{4}$$

Let q_{ij_1} and q_{ij_2} denote the highest and the second-highest elements of the vector $\mathbf{q}_i$, and let z_{j_1} and z_{j_2} denote the corresponding elements in the vector $\mathbf{z}$. Then the estimator of B_i given by SW (1993) is $\hat{B}_i^{\mathrm{TR}} = (q_{ij_1} - 1)z_{j_1} + q_{ij_2}z_{j_2}$, where TR denotes "truncation." Note that $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$ is not an unbiased estimator of $\boldsymbol{\beta}$ in general. However, if the probability is high that the best candidate link is the true link, then the truncation might produce a very small bias. We can suggest a further modification that would use as many candidate links as necessary ($2 \leq s \leq n$) in the calculation of B_i so that $\sum_{r=1}^{s} q_{ij_r} > p$, where $p = .90$ or similar cumulative probability.

Because $\mathrm{E}(\mathbf{z}) = \mathbf{W}\boldsymbol{\beta}$, we can propose an exactly unbiased estimator of $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{U}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{z}.$$

Table 1. File Structure for Regression of Linked Files

	File A				File B				
	Matching variables					Matching variables			
	v_1	$\cdots$	v_K	X	Y	w_1	$\cdots$	w_K	
Record a									Record b

As with the SW estimator, it is possible to "truncate" our estimator to use only the best candidate links for a record. Instead of w_i, such a procedure could use $w_i^{\mathrm{TR}} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$.

Note that $\mathbf{Q}$ (or, equivalently $\mathbf{W}$) is a function of the parameters of the mixture distribution defined in Section 3; that is, $\mathbf{Q} = \mathbf{Q}(\psi)$ or $\mathbf{W} = \mathbf{W}(\psi)$, where $\psi = \{(P(\boldsymbol{\gamma}|M), \boldsymbol{\gamma} \in \Gamma), (P(\boldsymbol{\gamma}|U), \boldsymbol{\gamma} \in \Gamma), \pi_M\}$. Thus we can write $\hat{\boldsymbol{\beta}}_{\mathrm{SW}} = \hat{\boldsymbol{\beta}}_{\mathrm{SW}}(\psi)$ and $\hat{\boldsymbol{\beta}}_{\mathrm{U}} = \hat{\boldsymbol{\beta}}_{\mathrm{U}}(\psi)$. In practice, ψ is unknown, and a reasonable estimator $\hat{\psi}$ (e.g., the maximum likelihood estimator) of ψ is used. In this case we obtain the SW and our estimator as $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}(\hat{\psi})$ and $\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})$. Interestingly, $\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})$ is an unbiased estimator of $\boldsymbol{\beta}$ whenever $\hat{\psi}$ can be assumed to be independent of $\mathbf{z}$. Such a situation is expected in most applications, because the distribution of the matching variables (e.g., last name, phone number) which determines the distribution of $\hat{\psi}$ is usually independent of the response variable $\mathbf{y}$ (e.g., income) and hence of $\mathbf{z}$.

5. VARIANCE ESTIMATION

5.1 Estimation of var($\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$)

SW (1993) suggested estimating the variance of their estimator [i.e., $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}(\hat{\psi})$] by modifying the usual least squares regression variance estimators. The usual variance estimator of $\hat{\boldsymbol{\beta}}_{\mathbf{yx}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is $\hat{\sigma}_0^2(\mathbf{X}'\mathbf{X})^{-1}$, where $(n-p)\hat{\sigma}_0^2$ is given by

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{xy}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{xy}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'_{\mathbf{xy}}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'_{\mathbf{xy}}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{xy}}. \quad (5)$$

Because $\mathbf{z} \approx \mathbf{y} + \mathbf{B}(\mathrm{E}(\mathbf{z}|\mathbf{y}) = \mathbf{y} + \mathbf{B})$, it is readily seen that $\mathbf{y}'\mathbf{y} \approx \mathbf{z}'\mathbf{z} - 2\mathbf{B}'\mathbf{z} + \mathbf{B}'\mathbf{B}$. The SW (1993) estimator of σ^2 is taken to be

$$\hat{\sigma}^2_{\mathrm{SW}} = (\mathbf{z}'\mathbf{z} - 2\mathbf{B}'\mathbf{z} + \mathbf{B}'\mathbf{B} - \hat{\boldsymbol{\beta}}'_{\mathrm{SW}}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{SW}})/(n-p).$$

The corresponding estimator of var($\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$) is taken to be $\hat{\sigma}^2_{\mathrm{SW}}(\mathbf{X}'\mathbf{X})^{-1}$. Clearly, the efficiency of this variance estimator depends on the degree of agreement between the vectors $\mathbf{z}$ and $\mathbf{y}$, reflected by the elements of the matrix $\mathbf{Q}$. When it is difficult to match two files, this variance estimator is likely to perform poorly. The estimator of the truncated version of the SW estimator will use $\hat{B}_i^{\mathrm{TR}}$ in place of $\hat{B}_i$ and $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}^{\mathrm{TR}}$ in place of $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$.

5.2 Estimation of var($\hat{\boldsymbol{\beta}}_{\mathrm{U}}$)

First, consider the case when ψ is known. Note that

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{U}}) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}, \quad (6)$$

where $\mathrm{var}(\mathbf{z}) = \boldsymbol{\Sigma} = ((\sigma_{ij}))$ and the expressions for σ_{ij} are given in Theorem A.1. We stress that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta}, \sigma^2, \psi)$; that is, it depends on both the parameters of the regression model (1), $\boldsymbol{\beta}$ and σ^2, and those of the mixture model, ψ. Because in this case ψ is known, we simply replace $\boldsymbol{\beta}$ and σ^2 by their estimators to obtain a variance estimator in the known ψ case. For example, we can use the unbiased estimator $\hat{\boldsymbol{\beta}}_{\mathrm{U}}$ to estimate $\boldsymbol{\beta}$.

We now consider an alternate estimator of σ^2. A naive estimator of mean squared error (MSE) based on $\mathbf{z}$ is given by

$$\mathrm{MSE} = \frac{1}{n-p}\mathbf{z}'\big(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\mathbf{z}.$$

Note that MSE is not unbiased for σ^2 under the model described by (1) and (2); it would be unbiased for σ^2 if the real data were $(\mathbf{X}, \mathbf{z})$. To obtain an alternate estimator of σ^2, consider

$$S^2 = \mathbf{z}'\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\mathbf{z},$$

where $\mathbf{I}$ is an n-dimensional identity matrix. According to Theorem A.2,

$$\mathrm{E}(S^2) = (n-p)\sigma^2 + \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\mathbf{H}\big], \quad (7)$$

where $\mathbf{H} = ((h_{ij}))$ with $h_{ii} = \boldsymbol{\beta}'\mathbf{A}_i\boldsymbol{\beta}$ and $h_{ij} = \boldsymbol{\beta}'\mathbf{A}_{ij}\boldsymbol{\beta}$.

Equation (7) motivates us to consider the following estimator of σ^2:

$$\hat{\sigma}^2 = \max\left(0, \frac{S^2 - \mathrm{tr}[(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')\hat{\mathbf{H}}]}{n-p}\right),$$

where $\hat{\mathbf{H}} = ((\hat{h}_{ij}))$ with $\hat{h}_{ii} = \hat{\boldsymbol{\beta}}'_{\mathrm{U}}\mathbf{A}_i\hat{\boldsymbol{\beta}}_{\mathrm{U}}$ and $\hat{h}_{ij} = \hat{\boldsymbol{\beta}}'_{\mathrm{U}}\mathbf{A}_{ij}\hat{\boldsymbol{\beta}}_{\mathrm{U}}$. It can be shown that $\hat{\sigma}^2$ is consistent for σ^2 under the model described by (1) and (2) and mild regularity conditions (see Thm. A.3).

Now consider the most practical situation when ψ is unknown. In this case, one may naively use var($\hat{\boldsymbol{\beta}}_{\mathrm{U}}$) with estimated $\boldsymbol{\beta}, \sigma^2$, and ψ as a variance estimator of $\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})$. However, this variance estimator fails to incorporate the uncertainties due to the estimation of ψ and thus underestimates the true variability of $\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})$. The same comment applies to the variance estimator of $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}(\hat{\psi})$ given in the previous section.

Parametric bootstrap methods have been quite effective in providing accurate variance estimators in many complex settings (e.g., see Lahiri 2003 for a review of parametric bootstrap methods for complex multilevel models). We now develop a parametric bootstrap method in our context to obtain a reliable variance estimator that captures the additional variability due to the estimation of ψ. Under such a method, we draw B bootstrap samples from the mixture distribution with ψ replaced by $\hat{\psi}$. A single sample is a table of size 2^K of counts. Let $\hat{\psi}^\star$ denote the estimator of ψ obtained from the procedure used for $\hat{\psi}$ but based on the bootstrap sample instead of the original sample. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\psi})$ denote any arbitrary estimator of $\boldsymbol{\beta}$ that depends on $\hat{\psi}$. Then we propose our bootstrap variance estimator as

$$v_{\mathrm{boot}} = E_\star\big[\mathrm{var}(\hat{\boldsymbol{\beta}}(\hat{\psi}^\star))\big] + V_\star\big[\hat{\boldsymbol{\beta}}(\hat{\psi}^\star)\big], \quad (8)$$

where $E_\star$ and $V_\star$ denote the expectation and the variance with respect to the bootstrap distribution and var$[\hat{\boldsymbol{\beta}}(\hat{\psi}^\star)]$ is an estimator of var($\hat{\boldsymbol{\beta}}$) with $\hat{\psi}^\star$ substituted for ψ. In practice, we propose using the Monte Carlo method to approximate the bootstrap expectation and variance. Thus

$$E_\star\big[\mathrm{var}(\hat{\boldsymbol{\beta}}(\hat{\psi}^\star))\big] \approx \frac{1}{B}\sum_{b=1}^{B}\mathrm{var}\big[\hat{\boldsymbol{\beta}}(\hat{\psi}^{\star b})\big]$$

and

$$V_\star\big[\hat{\boldsymbol{\beta}}(\hat{\psi}^\star)\big] \approx \frac{1}{B}\sum_{b=1}^{B}\big[\hat{\boldsymbol{\beta}}(\hat{\psi}^{\star b}) - \hat{\boldsymbol{\beta}}(\hat{\psi})\big]\big[\hat{\boldsymbol{\beta}}(\hat{\psi}^{\star b}) - \hat{\boldsymbol{\beta}}(\hat{\psi})\big]',$$

where $\hat{\psi}^{\star b}$ is the estimate of ψ from the bth bootstrap sample, $b = 1, \ldots, B$.

6. A MONTE CARLO SIMULATION

In this section we use a Monte Carlo simulation to investigate the performances of different estimators of a regression coefficient and the associated variance estimators for a simple linear regression model in the presence of mismatch errors. Our simulation study includes the naive estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$, the SW estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}(\hat{\psi})$; our proposed estimator, $\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})$; and a robust estimator mentioned in Section 4. We first describe the simulation conditions and then present results.

6.1 Simulation Conditions

We performed 400 replications under each of two sets of conditions. Table 2 describes the main conditions. In both sets of conditions, the sizes of the files vary between 2,000 and 10,000 records but are the same for files A and B. The regression slope β varies between .20 and .80 with the simulated data generated based on a regression model having error variance σ^2 equal to $1 - \beta^2$. The X variable is univariate normal with mean 0 and variance 1.

In case 1, files A and B have eight to twelve matching fields, whereas in case 2 they have six to ten. Agreements on the fields of information are independent of one another. The probability of agreement among matches varies between .55 and .95 in case 1 and between .55 and .85 in case 2. The probability of agreement among nonmatches varies between .10 and .50 in case 1 and between .20 and .50 in case 2. The size of the blocks affects how many potential links there are between the two files. Blocks are assumed to be linked together correctly, as they would be if they corresponded to geographical areas. Pairs from different blocks are nonlinks and are not used to estimate probabilities. Block sizes in case 1 range from 10 to 40 records (100–1,600 potential links per block), whereas in case 2 they range from 20 to 40 records. Thus case 2 yields more nonmatches than case 1, allowing us to understand the effect of nonmatches on different estimation methods.

We generated the files A and B and calculated comparison vectors. We used the EM algorithm (Dempster et al. 1977) to fit a two-class conditional independence mixture model to the comparison vectors to estimate probabilities for the Fellegi–Sunter (1969) algorithm. One product of the EM algorithm in this case are weights that represent the likelihood that a pair of records is a match. Estimated Fellegi–Sunter weights for links and nonlinks overlap more in case 2 than in case 1. The inequality constraints were used in the estimation, but one-to-one assignment was not enforced. It is not entirely clear how to force one-to-one matches and consider probabilities of matching in which two records in one file have a nonzero probability of matching a record in the second file. We will study the use of one-to-one assignment in the analysis of linked files in future work.

6.2 Simulation Results

We compute four estimates of the slope for each of the 400 simulation runs and compare with the true slope in terms of the absolute and squared deviations. We then compute average absolute deviation (AAD) and average squared deviation (ASD) for each of the four estimators, the average being taken over the 400 simulation runs. Our proposed estimator outperforms all the rival estimators in all cases. To summarize our results, we define the percent improvement with respect to AAD of our proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{U}}$ over a rival estimator $\hat{\boldsymbol{\beta}}$ as

$$100 \times \frac{\mathrm{AAD}_{\hat{\boldsymbol{\beta}}} - \mathrm{AAD}_{\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})}}{\mathrm{AAD}_{\hat{\boldsymbol{\beta}}_{\mathrm{U}}(\hat{\psi})}};$$

we define the percent improvement with respect to ASD similarly.

Table 3 displays the percent relative improvements with respect to both AAD and ASD. The performance of our estimator is impressive. The naive estimator has the worst performance followed by the robust and the Scheuren–Winkler estimators. Because the second set of conditions had less powerful matching information and more difficult settings (e.g., larger blocks), matching was less successful. As expected, the performances of all the three rival estimators relative to our proposed estimator worsen in this situation.

The coverage, reported in Table 4, is the percentage of times out of 400 that the form of a nominal 95% confidence interval,

$$\text{estimate} \pm 2\text{SE},$$

covers the true regression slope. The naive and the robust confidence intervals have the worst coverages. The confidence interval based on the SW method improves the coverage with respect

Table 2. Random Simulation Conditions for Two Cases

Conditions		*Lower limit*	*Upper limit*
Cases 1 and 2			
m	Number of replications	400	
n	Size of files A and B	2,000	10,000
β	Regression slope	.20	.80
σ^2	Regression variance	$1 - \beta^2$	
Case 1			
k	Number of comparison fields	8	12
p_k	Probability of agreement on a field for a match	.55	.95
q_k	Probability of agreement on a field for a nonmatch	.10	.50
	Size of blocks	10	40
Case 2			
k	Number of comparison fields	6	10
p_k	Probability of agreement on a field for a match	.55	.85
q_k	Probability of agreement on a field for a nonmatch	.20	.50
	Size of blocks	20	40

Table 3. The Percent Relative Improvement of the Proposed Estimator Over Rival Estimators

Method	*AAD*	*ASD*
Simulation Case 1		
Naive	84	170
Robust	51	86
Scheuren–Winkler	33	72
Simulation Case 2		
Naive	293	960
Robust	216	590
Scheuren–Winkler	109	327

to both the naive and the robust methods, but it is considerably worse than our proposed method. All the three rival methods are sensitive to the simulation condition. In contrast, our method is very stable under different simulation conditions.

For each dataset in the simulation, we generated 400 bootstrap comparison vector sets. For each of these, we found the maximum likelihood estimates of the mixture model parameters. Based on the matching probabilities determined by these mixture model parameters, we recomputed the regression estimates. When the bootstrap procedure is used, the coverage of the proposed method improves. The other estimators are hardly affected. The naive and robust estimators, as implemented here, do not use the estimated probabilities determined by the mixture models directly in either estimation or variance estimation. They rely simply on the x- and y-values associated with the best matches. Although the probability estimates change slightly with each bootstrap, the best matches are rarely changed. It would be possible to compute the actual variance of these estimators under the model (1) and (2). If one were relying on a naive estimator, however, one would not do so in practice. These estimators also are affected by severe bias.

The SW estimator uses the weights, but only in an estimate of the bias. The SW estimator of variance used here is the one suggested by SW (1993). The actual variance under the model of (1) and (2) would be different. As such, the variance estimate is largely determined by the naive variance. It is possible that the SW estimator would have better coverage if a bootstrap of the entire dataset, including the x- and y-values in addition to the comparison vector counts, were attempted. Although inclusion of the bootstrap variance estimate has improved the coverage of our estimator, it is still somewhat below the nominal 95% level, and further work is needed to produce additional improvements.

Table 4. Percent Coverage of 95% Confidence Intervals With and Without Bootstrap Adjustment of Standard Errors

	Coverage before bootstrap	*Coverage after bootstrap*
Simulation Case 1		
Naive	34	34
Robust	50	50
Scheuren–Winkler	59	60
Lahiri–Larsen	83	88
Simulation Case 2		
Naive	4	4
Robust	8	8
Scheuren–Winkler	40	41
Lahiri–Larsen	85	89

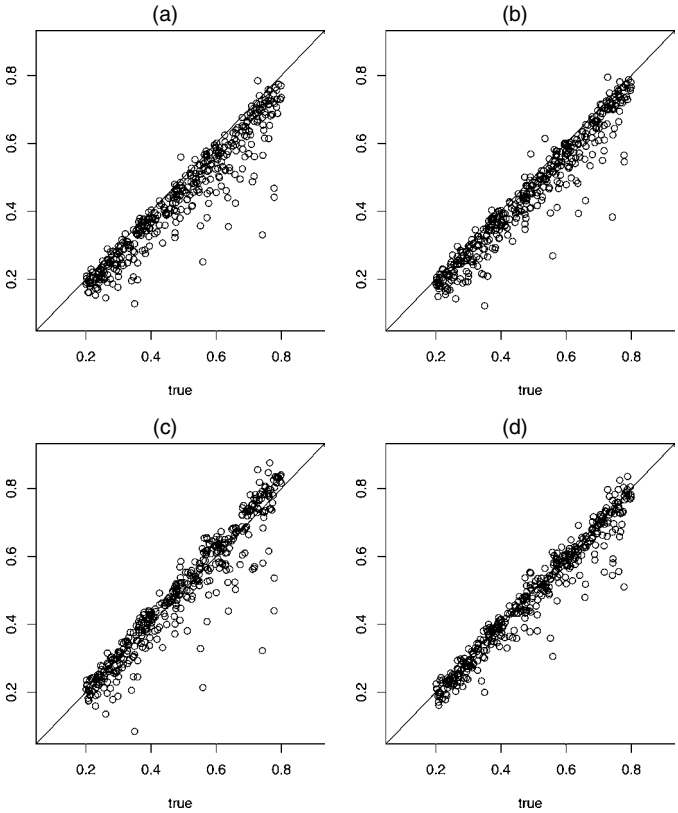

Figure 1. Comparison of Four Estimators on 400 Datasets, First Set of Simulation Conditions. Plots of the (a) naive, (b) robust, (c) Scheuren–Winkler, and (d) Lahiri–Larsen estimators versus the truth. Diagonal lines have slope 1.

Figure 1 plots the 400 regression estimates using the four methods versus the true simulation values under the first simulation conditions. If all of the dots are close to the line with slope 1, then estimators are doing very well. The naive estimates underestimate the true slope most of the time. The robust estimates improve on the naive estimates slightly but still underestimate the true slope. The SW and our estimates appear to be centered in the correct location around the line. Our estimates seem to have less spread about the line. Figure 2 shows the decreased performance of all estimates. Our estimates seem to be the least affected.

7. CONCLUSION

CRL can introduce errors into the composite file when errors are made in matching records. The mismatch errors can cause problems for analyses of variables brought together from different source files. In the presence of matching errors, naive estimators of linear regression coefficients are biased toward 0, because the errors attenuate the relationship between the predictors and response. In simulations, least median regression was not sufficient to guard against matching errors, whereas the method of Scheuren and Winkler (1993) as applied here made a useful adjustment. Our unbiased method seemed to perform very well across a range of situations. The bootstrap procedure that we have described is useful for reflecting uncertainty due to matching for our estimation procedure.

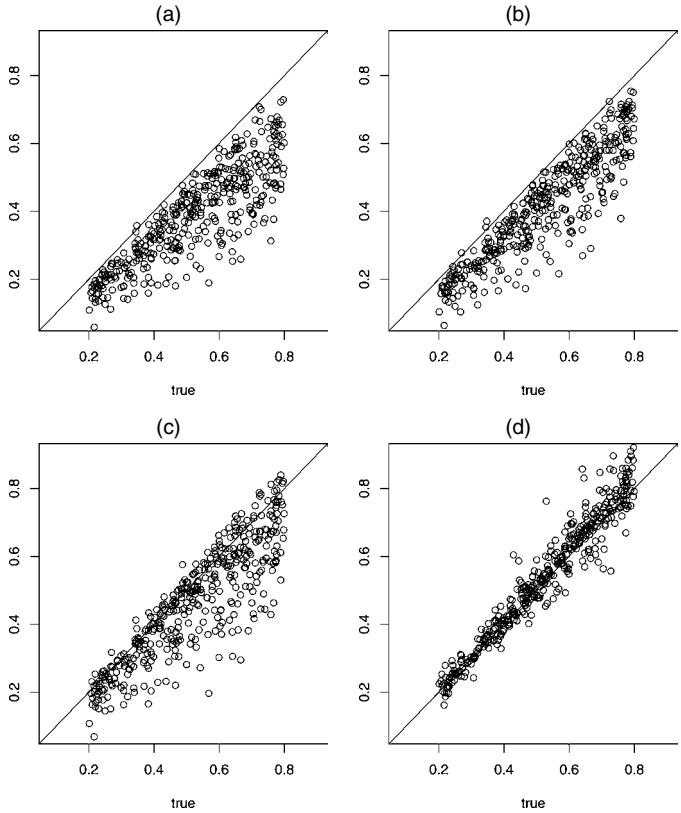

Figure 2. Comparison of Four Estimators on 400 Datasets, Second Set of Simulation Conditions. Plots of the (a) naive, (b) robust, (c) Scheuren–Winkler, and (d) Lahiri–Larsen estimators versus the truth. Diagonal lines have slope 1.

Future work will involve comparing our method with the SW iterative method, which we have not implemented, and incorporating iterative clerical review as done by Larsen and Rubin (2001). We also plan to investigate alternative bootstrap, jackknife, and multiple imputation options for propagation of error in matching through analyses.

APPENDIX: THEOREMS AND PROOFS

Theorem A.1. Under the model described by (1) and (2), we have, for $i, j = 1, \ldots, n$ $(i \neq j)$,

(a) $\mathrm{E}(z_i) = \mathbf{w}_i'\boldsymbol{\beta}$,

(b) $\mathrm{var}(z_i) = \sigma^2 + \boldsymbol{\beta}'\mathbf{A}_i\boldsymbol{\beta}$, where $\mathbf{A}_i = \sum_{j=1}^n q_{ij}\mathbf{d}_{ij}\mathbf{d}_{ij}'$ and $\mathbf{d}_{ij} = \mathbf{x}_j - \mathbf{w}_i$,

(c) $\mathrm{cov}(z_i, z_j) = \boldsymbol{\beta}'\mathbf{A}_{ij}\boldsymbol{\beta}$, where $\mathbf{A}_{ij} = \sum_{u=1}^n \sum_{v\neq u}^n q_{iu}q_{jv} \times \mathbf{d}_{iu}\mathbf{d}_{jv}'$.

Proof. Part (a) follows by noting that for $i = 1, \ldots, n$, $\mathrm{E}(z_i) = \mathrm{E}[\mathrm{E}(z_i|\mathbf{y})]$, $\mathrm{E}(z_i|\mathbf{y}) = q_{ii}y_i + \sum_{l\neq i} q_{il}y_l = \sum_{l=1}^n q_{il}y_l$, and $\mathrm{E}(y_i) = \mathbf{x}_i'\boldsymbol{\beta}$.

To prove part (b), we use the fact that

$$\mathrm{var}(z_i) = \mathrm{E}[\mathrm{var}(z_i|\mathbf{y})] + \mathrm{var}[\mathrm{E}(z_i|\mathbf{y})]. \tag{A.1}$$

Using the intermediate step in (a),

$$\mathrm{var}[\mathrm{E}(z_i|\mathbf{y})] = \mathrm{var}\left(\sum_{l=1}^n q_{il}y_l\right) = \sum_{l=1}^n q_{il}^2\sigma^2. \tag{A.2}$$

By the definition of variance,

$$\begin{aligned}\mathrm{var}(z_i|\mathbf{y}) &= \left(y_i - \sum_{l=1}^n q_{il}y_l\right)^2 q_{ii} + \sum_{j\neq i}\left(y_j - \sum_{l=1}^n q_{il}y_l\right)^2 q_{ij} \\ &= \sum_{j=1}^n \left(y_j - \sum_{l=1}^n q_{il}y_l\right)^2 q_{ij}. \end{aligned}\tag{A.3}$$

Now we compute the expectation for each j,

$$\begin{aligned}&\mathrm{E}\left(y_j - \sum_{l=1}^n q_{il}y_l\right)^2 \\ &\quad= \left\{\mathrm{var}\left(y_j - \sum_{l=1}^n q_{il}y_l\right) + \mathrm{E}\left(y_j - \sum_{l=1}^n q_{il}y_l\right)\right\}^2 \\ &\quad= \sigma^2\left(1 - 2q_{ij} + \sum_{l=1}^n q_{il}^2\right) + (\mathbf{d}_{ij}'\boldsymbol{\beta})^2, \end{aligned}\tag{A.4}$$

because

$$\begin{aligned}\mathrm{var}\left(y_j - \sum_{l=1}^n q_{il}y_l\right) &= \mathrm{var}\left(y_j(1-q_{ij}) - \sum_{l\neq j} q_{il}y_l\right) \\ &= \sigma^2(1-q_{ij})^2 + \sum_{l\neq j} q_{il}^2\sigma^2 \\ &= \sigma^2\left(1 - 2q_{ij} + \sum_{l=1}^n q_{il}^2\right)\end{aligned}$$

and

$$\mathrm{E}\left(y_i - \sum_{l=1}^n q_{il}y_l\right) = \mathbf{x}_i'\boldsymbol{\beta} - \sum_{l=1}^n q_{il}\mathbf{x}_l'\boldsymbol{\beta} = \mathbf{d}_{ij}'\boldsymbol{\beta}.$$

Part (b) follows by (A.4) into (A.3), resultant (A.3) and (A.2) into (A.1), and simplifying.

Turning to part (c), we use the fact that, for $i \neq j$,

$$\mathrm{cov}(z_i, z_j) = \mathrm{E}[\mathrm{cov}(z_i, z_j|\mathbf{y})] + \mathrm{cov}[\mathrm{E}(z_i|\mathbf{y}), \mathrm{E}(z_j|\mathbf{y})].$$

Because $\mathrm{E}(z_i|\mathbf{y}) = \sum_{l=1}^n q_{il}y_l$ and $\mathrm{E}(z_j|\mathbf{y}) = \sum_{l=1}^n q_{jl}y_l$,

$$\mathrm{cov}[\mathrm{E}(z_i|\mathbf{y}), \mathrm{E}(z_j|\mathbf{y})] = \sum_{l=1}^n q_{il}q_{jl}\sigma^2. \tag{A.5}$$

By the definition of covariance,

$$\mathrm{cov}(z_i, z_j|\mathbf{y}) = \sum_{u=1}^n \sum_{v=1}^n q_{iujv}\left(y_u - \sum_{k=1}^n q_{ik}y_k\right)\left(y_v - \sum_{l=1}^n q_{jl}y_l\right),$$

where $q_{iujv} = \Pr(z_i = y_u, z_j = y_v|y)$. Note that $q_{iuju} = q_{ivjv} = 0$. Now, for $u \neq v$,

$$\begin{aligned}&\mathrm{E}\left[\left(y_u - \sum_{k=1}^n q_{ik}y_k\right)\left(y_v - \sum_{l=1}^n q_{jl}y_l\right)\right] \\ &\quad= \mathrm{cov}\left[\left(y_u - \sum_{k=1}^n q_{ik}y_k\right), \left(y_v - \sum_{l=1}^n q_{jl}y_l\right)\right] \\ &\qquad+ \mathrm{E}\left(y_u - \sum_{k=1}^n q_{ik}y_k\right)\mathrm{E}\left(y_v - \sum_{l=1}^n q_{jl}y_l\right) \\ &\quad= 0 - \sigma^2(1-q_{iu})q_{ju} - \sigma^2(1-q_{jv})q_{iv} \\ &\qquad+ \sigma^2\sum_{k\neq u,v}^n q_{ik}q_{jk} + (\mathbf{d}_{iu}'\boldsymbol{\beta})(\mathbf{d}_{jv}'\boldsymbol{\beta}),\end{aligned}$$

because $\mathrm{E}(y_u - \sum_{k=1}^n q_{ik}y_k) = (\mathbf{x}_u' - \mathbf{w}_i')\boldsymbol{\beta} = \mathbf{d}_{iu}'\boldsymbol{\beta}$ and $\mathrm{E}(y_v - \sum_{l=1}^n q_{jl}y_l) = \mathbf{d}_{jv}'\boldsymbol{\beta}$. Thus

$$\begin{aligned}&\mathrm{E}[\mathrm{cov}(z_i, z_j|\mathbf{y})]\\ &= \sum_{u=1}^n\sum_{v\neq u}^n\Bigg[q_{iujv}\sigma^2\Bigg(-(1-q_{iu})q_{ju} - (1-q_{jv})q_{iv} + \sum_{k\neq u,v}^n q_{ik}q_{jk}\Bigg)\\ &\qquad + (\mathbf{d}_{iu}'\boldsymbol{\beta})(\mathbf{d}_{jv}'\boldsymbol{\beta})\Bigg].\end{aligned}$$

Noting that $\sum_{v,v\neq u}^n q_{iujv} = q_{iu}$, $\sum_{u,u\neq v}^n q_{iujv} = q_{jv}$, $\sum_{u=1}^n\sum_{v,v\neq u}^n q_{iujv} = 1$ and

$$\begin{aligned}&\sum_{u=1}^n\sum_{v\neq u}^n q_{iujv}\sum_{k\neq u,v}^n q_{ik}q_{jk}\\ &\quad= \sum_{u=1}^n\sum_{v\neq u}^n q_{iujv}\Bigg(\sum_{k=1}^n q_{ik}q_{jk} - q_{iu}q_{ju} - q_{iv}q_{jv}\Bigg)\\ &\quad= \sum_{k=1}^n q_{ik}q_{jk} - \sum_{u=1}^n q_{iu}^2q_{ju} - \sum_{v=1}^n q_{iv}q_{jv}^2,\end{aligned}$$

we arrive at

$$\begin{aligned}\mathrm{E}[\mathrm{cov}(z_i, z_j|\mathbf{y})] = \sigma^2\Bigg[&-\sum_{u=1}^n q_{iu}q_{ju} + \sum_{u=1}^n q_{iu}^2q_{ju}\\ &- \sum_{v=1}^n q_{iv}q_{jv} + \sum_{v=1}^n q_{iv}q_{jv}^2\\ &+ \sum_{k=1}^n q_{ik}q_{jk} - \sum_{u=1}^n q_{iu}^2q_{ju} - \sum_{v=1}^n q_{iv}q_{jv}^2\Bigg]\\ &+ \sum_{u=1}^n\sum_{v\neq u}^n(\mathbf{d}_{iu}'\boldsymbol{\beta})(\mathbf{d}_{jv}'\boldsymbol{\beta}).\end{aligned} \tag{A.6}$$

Adding the two parts, (A.5) and (A.6), yields

$$\mathrm{cov}(z_i, z_j) = \sum_{u=1}^n\sum_{v\neq u}^n q_{iu}q_{jv}(\mathbf{d}_{iu}'\boldsymbol{\beta})(\mathbf{d}_{jv}'\boldsymbol{\beta}). \tag{A.7}$$

Theorem A.2. Under the model described by (1) and (2), we have

$$\mathrm{E}(S^2) = (n-p)\sigma^2 + \mathrm{tr}[\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']\mathbf{H}, \tag{A.8}$$

where $S^2 = \mathbf{z}'[\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']\mathbf{z}$ and $\mathbf{H} = ((h_{ij}))$ with $h_{ii} = \boldsymbol{\beta}'\mathbf{A}_i\boldsymbol{\beta}$ and $h_{ij} = \boldsymbol{\beta}'\mathbf{A}_{ij}\boldsymbol{\beta}$.

Proof. First note that $S^2 = \mathbf{z}'[\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']\mathbf{z} = \mathrm{tr}[(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')\mathbf{z}\mathbf{z}']$ and that $\mathrm{E}(\mathbf{z}\mathbf{z}') = \mathrm{var}(\mathbf{z}) + \mathrm{E}(\mathbf{z})\mathrm{E}(\mathbf{z}') = \boldsymbol{\Sigma} + (\mathbf{W}\boldsymbol{\beta})(\mathbf{W}\boldsymbol{\beta})'$, where $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ with $\sigma_{ii} = \mathrm{var}(z_i)$ and $\sigma_{ij} = \mathrm{cov}(z_i, z_j)$, $i \neq j$. So

$$\begin{aligned}\mathrm{E}(S^2) &= \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)(\boldsymbol{\Sigma} + \mathbf{W}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{W}')\big]\\ &= \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\boldsymbol{\Sigma}\big]\\ &= \sigma^2\,\mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\big] + \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\mathbf{H}\big]\\ &= (n-p)\sigma^2 + \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\mathbf{H}\big].\end{aligned} \tag{A.9}$$

The second term on the first line of (A.9) vanishes because $\mathrm{tr}[\mathbf{W}\boldsymbol{\beta} \times \boldsymbol{\beta}'\mathbf{W}' - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{W}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{W}'] = \mathrm{tr}[\mathbf{W}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{W}' - \mathbf{W}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{W}'] = 0$. The conversion to the last line of (A.9) follows from usual regression algebra (e.g., Sen and Srivastava 1990, p. 278).

Theorem A.3. Under the model described by (1) and (2) and the regularity conditions

(a) $\sup_{ij}|x_{ij}| \leq c < \infty$
(b) $\mathbf{W}'\boldsymbol{\Sigma}\mathbf{W} = O(n)$,

the estimator $\hat{\sigma}^2 = \max[0, \{S^2 - \mathrm{tr}(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')\hat{\mathbf{H}}\}/(n-p)]$, where $\hat{\mathbf{H}} = ((\hat{h}_{ij}))$ with $\hat{h}_{ii} = \hat{\boldsymbol{\beta}}_{\mathrm{U}}'\mathbf{A}_i\hat{\boldsymbol{\beta}}_{\mathrm{U}}$ and $\hat{h}_{ij} = \hat{\boldsymbol{\beta}}_{\mathrm{U}}'\mathbf{A}_{ij}\hat{\boldsymbol{\beta}}_{\mathrm{U}}$, is consistent for σ^2 as $n \to \infty$.

Proof. Let

$$\begin{aligned}\tilde{\sigma}^2 = u + \frac{1}{n-p}\big(&\mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\mathbf{H}\big]\\ &- \mathrm{tr}\big[\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\hat{\mathbf{H}}\big]\big),\end{aligned} \tag{A.10}$$

where

$$u = \frac{S^2 - \mathrm{tr}[\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']\mathbf{H}}{n-p}.$$

Because $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{U}}) = \beta$ and $\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{U}}) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1} = O(n^{-1})$, by assumptions (a) and (b), we have $\hat{\boldsymbol{\beta}}_{\mathrm{U}} \xrightarrow{p} \boldsymbol{\beta}$ as $n \to \infty$. Thus the second term in (A.10) tends to 0 as $n \to \infty$.

We show that $u \xrightarrow{p} \sigma^2$ as $n \to \infty$. By Theorem A.2, $\mathrm{E}(u) = \sigma^2$. Now,

$$\mathrm{var}(u) = \frac{\mathrm{var}(S^2)}{(n-p)^2}.$$

Note that $\mathrm{var}(S^2) = \mathrm{var}(\boldsymbol{\eta}'(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}')\boldsymbol{\eta})$, where $\boldsymbol{\eta} = \mathbf{z} - \mathbf{W}\boldsymbol{\beta}$. Using part (d) of lemma C.4 of Lahiri and Rao (1995), we have

$$\mathrm{var}\big(\boldsymbol{\eta}'\big(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\big)\boldsymbol{\eta}\big) = O(n).$$

So $\mathrm{var}(S^2) = O(n)$ and $\mathrm{var}(u) = O(n^{-1})$, and thus the result is established.

[Received June 2002. Revised June 2004.]

REFERENCES

Alvey, W., and Jamerson, B. (1997), *Record Linkage Techniques—1997*, Proceedings of an International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management of the Budget.

Armstrong, J. B., and Mayda, J. E. (1993), "Model-Based Estimation of Record Linkage Error Rates," *Survey Methodology*, 19, 137–147.

Becker, M. P., and Yang, I. (1998), "Latent Class Marginal Models for Cross-Classifications of Counts," *Sociological Methodology*, 28, 293–325.

Belin, T. (1993), "Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment," *Survey Methodology*, 19, 13–29.

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.

Burkard, R. E., and Derigs, U. (1980), *Assignment and Matching Problems: Solution Methods With FORTRAN-Programs*, New York: Springer-Verlag.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with comments), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–37.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Gill, L. E. (1997), "OX–LINK: The Oxford Medical Record Linkage System Demonstration of the PC Version," in *Record Linkage Techniques—1997*, Proceedings of an International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management of the Budget, p. 491.

Gomatam, S., and Larsen, M. D. (2004), "Record Linkage and Counterterrorism," *Chance*, 17, 25–29.

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*, 84, 414–420.

——— (1995), "Probabilistic Linkage of Large Public Health Data Files," *Statistics in Medicine*, 14, 491–498.

Lahiri, P. (2003), "On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation," *Statistical Science*, 18, 199–210.

Lahiri, P., and Larsen, M. D. (2000), "Regression Analysis With Linked Data," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 11–19.

Lahiri, P., and Rao, J. N. K. (1995), "Robust Estimation of Mean Squared Error of Small Area Estimators," *Journal of the American Statistical Association*, 90, 758–766.

Larsen, M. D. (1999), "Multiple Imputation Analysis of Records Linkage Using Mixture Models," in *Proceedings of the Statistical Society of Canada, Survey Methods Section*, pp. 65–71.

——— (2001), "Methods for Model-Based Record Linkage and Analysis of Linked Files," in *Proceedings of the Government Statistics Section*, American Statistical Association.

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32–41.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 60, 1005–1027.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford, U.K.: Oxford University Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954–959.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in Fortran* (2nd ed.), Cambridge, U.K.: Cambridge University Press, pp. 698–700.

Scheuren, F., and Winkler, W. E. (1991), "Regression Analysis of Data Files That Are Computer Matched," in *Proceedings of the Annual Research Conference*, U.S. Census Bureau, pp. 669–687.

——— (1993), "Regression Analysis of Data Files That Are Computer Matched," *Survey Methodology*, 19, 39–58.

——— (1997), "Regression Analysis of Data Files That Are Computer Matched—Part II," *Survey Methodology*, 23, 157–165.

Sen, A., and Srivastava, M. (1990), *Regression Analysis*, New York: Springer-Verlag, pp. 277–278.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, 19, 31–38.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of Survey Research Methods Section*, American Statistical Association, pp. 667–671.

——— (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 354–359.

——— (1993), "Improved Decision Rules in the Fellegi–Sunter Model of Record Linkage," in *Proceedings of Survey Research Methods Section*, American Statistical Association, pp. 274–279.

——— (1994), "Advanced Methods for Record Linkage," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 467–472.

——— (1995), "Matching and Record Linkage," in *Business Survey Methods*, eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: Wiley, pp. 355–384.