NISS

Secure Regression for Vertically Partitioned, Partially Overlapping Data

Jerome P. Reiter, Christine N. Kohnen, Alan F. Karr
Xiaodong Lin, and Ashish P. Sanil

Technical Report Number 146
October 2004

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Secure Regression for Vertically Partitioned, Partially Overlapping Data

Jerome P. Reiter[1], Christine N. Kohnen[1], Alan F. Karr[2], Xiaodong Lin[2], Ashish P. Sanil[2]

Abstract: We consider the setting where multiple parties with different variables and units seek to combine their data to fit regressions but are not willing or not allowed to share their data values. We present a general strategy to tackle such problems by treating them as missing data problems, and we estimate regression coefficients using secure EM algorithms. We present secure EM algorithms for linear and log-linear regressions, based on the multivariate normal and multinomial distributions. The parties compute and share the sufficient statistics required for the EM algorithms via secure matrix product protocols, which avoid sharing individual data values.

Key Words: Confidentiality, Data Integration, Disclosure, EM algorithm, Regression

1 Introduction

When multiple parties collect data on different but related variables, they may seek to combine their information to fit regression models. The parties' combined data can be described as vertically partitioned and partially overlapping: not all parties have the same variables, and some records are common to multiple parties. Horizontally partitioned data, i.e. all parties have the same variables but different records, has been considered by Karr *et al.* (2004b). Purely vertically partitioned data, i.e. all parties have the same units but different variables, has been considered by Sanil *et al.* (2004); Karr *et al.* (2004a). The vertically partitioned, partially overlapping setting includes these other two settings as special cases.

The vertically partitioned, partially overlapping data setting can be conceived as a missing data problem. That is, conceptually the data from the multiple parties could be concatenated to form a rectangular data set with missing values for those records not common to all parties. If a common model is reasonable for all units, i.e. if they can be considered exchangeable, regression inferences from this incomplete data can be obtained using missing data methods, for example the EM algorithm or Bayesian posterior simulation.

Standard missing data methods require that the parties share their data values; however, the parties may be unwilling, or not allowed legally, to do so because of concerns over data confidentiality. In this article, we describe EM algorithms that do not require the sharing of data values and hence may satisfy disclosure limitation constraints. These algorithms assume the data, suitably transformed, can be reasonably described by either a multivariate normal or multinomial distribution. These assumptions allow us to capitalize on the fact that, for exponential families, the EM algorithm requires sharing only sufficient statistics, which for the multivariate normal are sums and inner-products of the observed data values and for the multinomial model are counts in relevant tables. These quantities can then be shared using secure summation and secure inner-product protocols, which allow parties to compute sums and inner-products without sharing data values.

The remainder of the paper is organized as follows. Section 2 reviews the EM algorithm for incomplete, multivariate normal and multinomial data. Section 3 describes secure EM algorithms for the vertically partitioned, partially overlapping setting for these data types. Section 4 describes the confidentiality protection and data utility of these algorithms. Section 5 concludes with a discussion of extensions to the approach.

2 EM algorithm for incomplete data

2.1 Multivariate normal data

Let Y be an $n \times p$ matrix of n independent and identically distributed draws from a p-variate normal distribution with mean vector μ and covariance matrix Σ. Following the notation of Little and Rubin (2002), we write $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} represents all observed values and Y_{mis} represents all missing values. We further write $Y_{obs} = (y_{obs,1}, y_{obs,2}, \ldots, y_{obs,n})$, where $y_{obs,i}$ represents the variables observed for unit

[1]Institute of Statistics and Decision Sciences, Duke University
[2]National Institute of Statistical Sciences

i, for $i = 1, \ldots, n$. It is assumed the data are missing at random (Rubin, 1976).

The coefficients and standard errors of any regression involving Y can be obtained from functions of the maximum likelihood estimates (MLEs) of μ and Σ, given Y_{obs}. Unfortunately, the MLE equations for μ and Σ do not have closed form solutions when the data are incomplete. Instead the MLEs are found using iterative algorithms, such as the Newton-Raphson algorithm or the EM algorithm (Dempster *et al.*, 1977). The EM algorithm has desirable properties: it is relatively simple to code and debug; it is guaranteed to converge to a mode in the likelihood function of μ and Σ, given Y_{obs}; and, unlike Newton-Raphson algorithms, it does not require inverting matrices of second derivatives (Schafer, 2000; Little and Rubin, 2002). Convergence of the EM algorithm is linear in the amount of missing information.

Let $\mu^{(t)}$ and $\Sigma^{(t)}$ be the parameter estimates at iteration t. Reasonable starting values for $\mu^{(1)}$ and $\Sigma^{(1)}$ are the sample moments for records with all variables observed. Generally, it is wise to run EM multiple times using different starting values. This explores the likelihood function for multiple modes. The algorithm is judged to converge when the changes in parameter values from iteration (t) to iteration $(t+1)$ do not exceed small tolerance values.

The EM algorithm cycles between two steps, the E-step and the M-step. The E-step of the algorithm involves calculating the following expectations:

$$E(\sum_{i=1}^{n} y_{ij} | Y_{obs}, \mu^{(t)}, \Sigma^{(t)}), \ j = 1, \ldots, p, \qquad (1)$$

$$E(\sum_{i=1}^{n} y_{ij} y_{ik} | Y_{obs}, \mu^{(t)}, \Sigma^{(t)}), \ j, k = 1, \ldots, p. \qquad (2)$$

For any j, the expectation in (1) can be written as $\sum_{i=1}^{n} y_{ij}^{(t)}$, where

$$\begin{aligned} y_{ij}^{(t)} &= y_{ij} \quad \text{when } y_{ij} \text{ observed} \qquad (3) \\ &= \hat{y}_{ij} \quad \text{when } y_{ij} \text{ missing} \qquad (4) \end{aligned}$$

and $\hat{y}_{ij} = E(y_{ij} | Y_{obs}, \mu^{(t)}, \Sigma^{(t)})$. To calculate $\hat{y}_{ij}$, we require the following notation. Let $\mu_j^{(t)}$ be the element of $\mu^{(t)}$ for variable j, and let $\mu_{Bi}^{(t)}$ be the vector of elements of μ for all variables in $y_{obs,i}$. Let $\Sigma_{j,Bi}^{(t)}$ be the elements of $\Sigma^{(t)}$ corresponding to the covariances between variable j and all variables in $y_{obs,i}$. Let $\Sigma_{Bi,Bi}^{(t)}$ be the elements of $\Sigma^{(t)}$ corresponding to the covariance matrix of all variables in $y_{obs,i}$. Then, the $\hat{y}_{ij}$ is obtained from

$$\hat{y}_{ij} = \mu_j^{(t)} + \Sigma_{j,Bi}^{(t)} (\Sigma_{Bi,Bi}^{(t)})^{-1} (y_{obs,i} - \mu_{Bi}^{(t)}). \qquad (5)$$

which is the predicted value in the regression of y_{ij} on y_{Bi}.

For any j, k, the expectation in (2) can be written as

$$\sum_{i=1}^{n} y_{ij}^{(t)} y_{ik}^{(t)} + c_{jki}^{(t)}, \qquad (6)$$

where $c_{jki}^{(t)} = 0$ when y_{ij} or y_{ik} is observed, and $c_{jki}^{(t)} = Cov(y_{ij}, y_{ik} | Y_{obs}, \mu^{(t)}, \Sigma^{(t)})$ when y_{ij} and y_{ik} are missing. To calculate the covariance in (6), we need the following notation. Let $\Sigma_{jk}^{(t)}$ correspond to the covariance submatrix of $\Sigma^{(t)}$ for variables j and k. Let $\Sigma_{jk,Bi}^{(t)}$ be the submatrix of $\Sigma^{(t)}$ corresponding to the covariances between (y_{ij}, y_{ik}) and the variables in $y_{obs,i}$. The covariance in (6) is the off-diagonal element in the 2×2 matrix:

$$\Sigma_{jk}^{(t)} - \Sigma_{jk,Bi}^{(t)} (\Sigma_{Bi,Bi}^{(t)})^{-1} \Sigma_{jk,Bi}^{(t)}. \qquad (7)$$

After obtaining the E-steps for all variables k, the algorithm proceeds to the M-step. The M-step involves maximizing the likelihood after completing the data with the expectations from the E-step, which means finding updated iterates $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$ as follows for each j, k:

$$\mu_j^{(t+1)} = (1/n) \sum_{i=1}^{n} y_{ij}^{(t)} \qquad (8)$$

$$\begin{aligned} \sigma_{jk}^{(t+1)} &= (1/n) \sum_{i=1}^{n} (y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) \\ &+ (1/n) \sum_{i=1}^{n} c_{ijk}^{(t)} \qquad (9) \end{aligned}$$

Estimates of the regression coefficients and residual variance involving any subset of Y can be obtained from the relevant entries in the MLEs of μ and Σ. The result in (5) provides the relevant intercept and slope coefficients, and the result in (9) provides the relevant residual variance. Standard errors for the regression coefficients can be approximated using the approach of Beale and Little (1975) and Little (1979). Or, they can be determined using the supplemented EM approach of (Meng and Rubin, 1991). Both approaches are described in Little and Rubin (2002, p. 240).

2.2 Multinomial data

Let $Y = (Y_1, \ldots, Y_p)$ be an $n \times p$ matrix of categorical data, where each Y_j has d_j categories. As before,

let $Y = (Y_{obs}, Y_{mis})$. We assume the agencies seek to estimate the parameters of a saturated log-linear model. This corresponds to estimating cell probabilities in the full contingency table with $D = \prod_{j=1}^{p} d_j$ cells, some of which may be structural zeros.

For any cell c in the full table, let x_c be the total number of sampled units in that cell, and let θ_c be the probability of sampling a unit belonging to cell c. If $Y = Y_{obs}$, the table has a multinomial distribution with index $n = \sum x_c$ and parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$. The MLE of θ for this table is the cell percentages, and the x_c are sufficient statistics for θ. When $Y \neq Y_{obs}$, it is not possible to compute the x_c directly, since we observe lower dimensional tables of marginal counts. These sub-tables are used in the EM algorithm to find the MLE of θ.

Each observation can be classified into some missing data pattern, which we index by superscript $m = 1, \ldots, M$. Let $x_c^m = x_{obs,c}^m + x_{mis,c}^m$ be the count in cell c for those units with missing data pattern m, where $x_{obs,c}^m$ is the observed contribution to that count and $x_{mis,c}^m$ is the unobserved contribution to that count (due to missing data). Let z_c^m be the count of units whose pattern of missing data m qualifies them to be potentially in cell c.

Let $\beta^{m(t)} = \sum_c \theta_c^{(t)}$, where the summation is over all cells c that the units with missing data pattern m are qualified to belong to. The E-step of the EM algorithm for any cell c is then

$$\begin{aligned} E(x_c|Y_{obs}, \theta^{(t)}) &= \sum_{m=1}^{M} E(x_c^m|Y_{obs}, \theta^{(t)}) \\ &= \sum_{m=1}^{M} x_{obs,c}^m + z_c^m \theta_c^{(t)} / \beta^{m(t)}. \end{aligned} \tag{10}$$

Once the E-steps are completed for all cells c, $\theta^{(t)}$ is updated with a new value $\theta^{(t+1)}$, given by

$$\theta_c^{(t+1)} = E(x_c|Y_{obs}, \theta^{(t)})/n \tag{11}$$

for all $c \in \mathcal{C}$. The algorithm is iterated until convergence of θ to a mode.

3 Secure EM algorithms

Regression with vertically partitioned, partially overlapping data is essentially a missing data problem. In this section, we describe how secure matrix product protocols can be used to perform secure EM.

We make several simplifying assumptions. First, we assume that the parties share the unique identifiers of the units in their data sets. This is needed to allow parties to identify the units that are common to multiple parties' data sets. Second, we assume that matching on these unique identifiers can be done without errors. Third, we assume that each party has distinct variables except for the unique identifiers, i.e. there are no overlapping variables. We describe how to relax some of these assumptions in Section 4.

Our strategy for performing secure EM follows three steps. First, the parties must cooperatively group units by patterns of missing data and share these groupings. Second, the parties securely share the observed-data sufficient statistics needed for the E-steps using a secure inner-product protocol (Karr *et al.*, 2004a). Third, each party independently runs EM based on the shared sufficient statistics. Only the sufficient statistics are shared, not the actual data values. Parties pass around information only once, i.e. the initial sharing of the summary statistics. We now discuss these steps in more detail for the multivariate normal model and the multinomial model.

3.1 Multivariate normal data

When the data are multivariate normal, the parties may want to run the EM algorithm of Section 2.1 to determine inferences for a particular regression. However, restrictions on data sharing would make it impossible for any party to compute the $y_{ij}^{(t)}$ for missing y_{ij}. A version of EM is needed that gets around this difficulty.

In the first step, the parties group units by patterns of incomplete data, which is possible since the parties share unique identifiers. For $i = 1, \ldots, n$, let $y_{mis,i}$ represent the variables missing for unit i, and let Y_{Mi} be the data for all units with the same pattern of missing values as $y_{mis,i}$. All parties know the records' missing data patterns, although parties only know the values of the $y_{obs,i}$ for their own data. Let M be the number of patterns of missing data.

In the second step, the parties calculate and share two tables of summary statistics that are used in the EM. The first table has dimension $M \times p$, with M rows corresponding to the missing data patterns and p columns corresponding to the variables in the data set. The entry in the table for row m and column j is the sum of the observed y_{ij} for those units with the missing data pattern of row m. Because we assume no overlapping variables, each sum is computed by only one party.

The second table has dimension $M \times p(p+1)/2$, with M rows corresponding to the missing data patterns and $p(p+1)/2$ columns corresponding to the inner-products of all p variables in the data set, including the $\sum_{i \in Mi} y_{ij}^2$. The entry in the table for row m and the column associated with variables (j, k) is

the $\sum y_{ij}y_{ik}$ for those units with the missing data pattern of row m. Because we assume no overlapping variables, each entry in the table is derived from a single inner-product between two parties. The table has many structural zeros, because there are no inner-products between the missing and observed data. The parties compute the inner-products using a secure inner-product protocol (Karr *et al.*, 2004a), which allows parties to perform inner-products without sharing values of the variables.

In the third step, each party runs EM independently using the two tables of summary statistics. Let $\mu^{(t)}$ and $\Sigma^{(t)}$ be the parameter estimates at iteration t, and let $y_{ij}^{(t)}$ be defined as in (3) and (4). For the E-step for variable j in (1), we require the sum of the observed y_{ij} and the sum of the predicted values, $\sum_i \hat{y}_{ij}^{(t)}$, from the appropriate regressions. This latter sum can be written as a sum over the patterns of missing data. From (5), this is

$$\begin{aligned}\sum_{Mi}\sum_{i\in Mi} \hat{y}_{ij}^{(t)} &= \sum_{Mi}\sum_{i\in Mi}(\mu_j^{(t)} - \Sigma_{j,Mi}^{(t)}(\Sigma_{Mi,Mi}^{(t)})^{-1}\mu_{Mi}^{(t)}) \\ &\quad + \sum_{Mi}\sum_{i\in Mi}\Sigma_{j,Mi}^{(t)}(\Sigma_{Mi,Mi}^{(t)})^{-1}y_{obs,i} \\ &= \sum_{Mi} N_{Mi}\alpha_{Mi,j}^{(t)} + \sum_{Mi}\sum_{j\in Mi}\beta_{Mi,j}^{(t)}(\sum_{i=1}^{N_{Mi}} y_{ij}). \quad (12)\end{aligned}$$

The first sum in (12) is a function of $\mu^{(t)}$, $\Sigma^{(t)}$, and the number of records in each pattern of incomplete data, N_{Mi}. The second sum in (12) is a function of $\mu^{(t)}$, $\Sigma^{(t)}$, and the shared sums of other variables. Hence, it is easy to calculate these quantities from the tables of summary statistics.

For the E-step in (2) for any pair of variables (j,k), the expectation of their cross-product, i.e. the sum in (6), can be split into three parts. For the units with both y_{ij} and y_{ik} observed, parties use the inner-product from the shared summary statistics. For the units with exactly one of y_{ij} or y_{ik} missing, the secure inner-product takes several steps. Without loss of generality, assume that y_{ij} is missing and y_{ik} is observed. Thus, for these particular units, we need to compute

$$\sum_{Mi}\sum_{i\in Mi}\hat{y}_{ij}^{(t)}y_{ik} = \sum_{Mi}\sum_{i\in Mi}\left(\alpha_{Mi}^{(t)} + \sum_{u\in Mi}\beta_{Mi,u}^{(t)}y_{iu}\right)y_{ik}$$

where the Mi includes only the patterns of observations where y_{ij} is missing and y_{ik} is observed. For any particular pattern Mi, the needed sums and inner-products can be pulled from the shared summary statistics.

For the units with both y_{ij} and y_{ik} missing, we need to compute the $\sum_i c_{jki}^{(t)} + \sum_i \hat{y}_{ij}^{(t)}\hat{y}_{ik}^{(t)}$. The first component depends only on $\Sigma^{(t)}$. Let $\delta_{Mi}^{(t)}$ and $\gamma_{Mi}^{(t)}$ be the intercept and vector of regression slopes from the regression of y_{ik} on $y_{obs,i}$. The second component is

$$\begin{aligned}\sum_{Mi}\sum_{i\in Mi}\hat{y}_{ij}^{(t)}\hat{y}_{ik}^{(t)} &= \sum_{Mi} N_{Mi}\alpha_{Mi}^{(t)}\delta_{Mi}^{(t)} \\ &\quad + \sum_{Mi}\alpha_{Mi}^{(t)}\sum_{v\in Mi}\gamma_{Mi,v}^{(t)}\left(\sum_{i\in Mi} y_{iv}\right) \\ &\quad + \sum_{Mi}\delta_{Mi}^{(t)}\sum_{u\in Mi}\beta_{Mi,u}^{(t)}\left(\sum_{i\in Mi} y_{iu}\right) \\ &\quad + \sum_{Mi}\sum_{u,v\in Mi}\beta_{Mi,u}^{(t)}\gamma_{Mi,v}^{(t)}\left(\sum_{i\in Mi} y_{iu}y_{iv}\right) \quad (13)\end{aligned}$$

where the Mi include only those units with both y_{ij} and y_{ik} missing. This can be determined using the shared summary statistics.

Once a party has completed its E-steps, it performs the M-step. This is straightforward. For any $\mu_j^{(t+1)}$, the party simply feeds the sums obtained from (12) into (10). For any $\sigma_{jk}^{(t+1)}$, the party takes the quantity $(1/n)E(\sum_{i=1}^n y_{ij}y_{ik}|Y_{obs},\mu^{(t)},\Sigma^{(t)})$ calculated from the E-step, and subtracts $\mu_j^{(t+1)}\mu_k^{(t+1)}$. Once new parameter estimates are obtained, the party uses the values of $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$ for the next iteration, continuing until convergence.

Given the MLEs of μ and Σ, the party can find the regression coefficients for any regression involving Y, as described in Section 2.1. Standard errors can be determined by the SEM algorithm, which requires running the E-steps more times as described by Meng and Rubin (1991). No modifications to the SEM algorithm are needed to run it securely, as it relies solely on the MLEs and the shared summary statistics. Standard errors also can be determined from the MLEs and shared summary statistics using the approximations of Beale and Little (1975).

3.2 Multinomial data

To perform the EM algorithm of Section 2.2 securely, all relevant counts must be shared without sharing individual units' data. This is accomplished as follows. First, the agencies determine the missing data patterns for the units in their combined data, done by sharing identifiers. Then, each agency transforms

all of its Y_j into d_j vectors of indicators,

$$I(Y_j = y) = \left\{ \begin{array}{ll} 1 & \text{if } Y_j = y \\ 0 & \text{otherwise} \end{array} \right\}$$

where the values of y correspond to the values of Y_j. Hence, there are a total of $D = \prod_{j=1}^{p} d_j$ columns of indicators among the agencies.

The secure inner-product protocol (Karr *et al.*, 2004a) can be used to obtain counts for any marginal subtable. Hence, it is possible to obtain the z_c^m for all missing data patterns m and cells c. After also sharing the $x_{obs,c}^m$, the agencies have all sufficient statistics needed to run the EM algorithm individually.

4 Implementation issues

In this section, we discuss the confidentiality and utility of these approaches, as well as ways to relax some of the assumptions used in the previous section.

4.1 Confidentiality

Secure integration of vertically partitioned data faces some non-standard confidentiality issues. The parties need to determine the sets of overlapping units, which implies revealing that certain records are in (or not in) various databases. They also need to share the names of the attributes in their respective data sets, which parties may be reluctant to do.

Because the parties share summary statistics and cell counts, but not individual data values, secure EM has the potential to allow parties to combine information without revealing sensitive attribute values. Clearly, however, the secure EM presented here does not protect identities, since unique identifiers are shared to enable matching. It may be possible to protect identities of non-overlapping units through secure matching techniques or by sharing constructed identification codes, for example numerical values assigned to combinations of key variables. Parties would match on these values instead of the unique identifiers. Such matching could introduce error into the matching process, which neither version of secure EM accounts for.

Secure EM may be subject to attribute disclosures in certain data settings or when parties are dishonest. The algorithms are only as safe as the summations and secure inner-product algorithms used in the initial sharing of sufficient statistics. For examples, the shared summations for a missing data pattern with only one record equal that record's values; and, an inner-product between a sensitive attribute and a vector containing $(1, 0, 0, 0, 0, 0)$ reveals the sensitive attribute of the first record. Additionally, the sum of known records' sensitive attributes provides an upper bound for those records' individual values, which may be enough information to be considered a disclosure. Dishonest parties can create false missing data patterns with one or a few records, or provide bogus vectors for inner-products, that reveal values from the shared summations and inner-products. There is no way to protect against such deception. The agencies need to cooperate in a semi-honest manner (Karr *et al.*, 2004b).

In the multivariate normal setting, for any missing data pattern with q variables there are $q + q(q + 1)/2$ equations involving the records in that pattern. When the number of records in that pattern is less than or equal to $q + q(q + 1)/2$, the parties can find solutions for the data values in that pattern. To protect confidentiality the parties may have to exclude missing data patterns with small numbers of records from the EM, although this could bias parameter estimates. An alternative is for one party to impute sensible values for enough of the missing data, based on relationships from the cases with all data observed, so that those records fall into a missing data pattern with sufficient numbers of records.

4.2 Utility

If no missing data patterns are excluded or merged into other categories, the secure EM yields the MLEs given the observed data. Hence, its inferential utility is equivalent to concatenating the data and applying EM. We note there is no utility loss typical of disclosure limitation strategies like global recoding, data swapping, or adding random noise.

Practically, given the summary statistics, the secure EM is easy to code and, given a sufficiently regular distribution, is guaranteed to converge to the MLE when it exists. Each party can run the secure EM on its own; the iterations do not require exchanging any information other than the summary statistics at the initial stage. Secure EM requires careful bookkeeping. The missing data patterns must be cataloged and stored, although this is fairly routine programming. Summary statistics and lower dimensional tables must be securely computed and shared. When there are many missing data patterns, this step will require software to manage inputs to the secure inner-product operations.

For multinomial data, it is possible to extend the secure EM to log-linear models that are not saturated, since iterative proportional fitting algorithms again only need counts.

The most serious limitation of the secure EM algorithms presented here are the "sight-unseen" specification of distributions. Because parties are not sharing data values, it is difficult for them to determine whether multivariate normality, or any other distributional assumption, is reasonable for the data. Parties may be able to transform some variables to approximate normal, marginal distributions, but this of course does not guarantee multivariate normality. Fortunately, for obtaining regression inferences, the EM is more robust to the normality assumption than it may appear. Only the distribution of the errors in the regression need be normally distributed.

A related limitation is the assumption that the non-overlapping units' incomplete data are missing at random. The reasons for missing data may depend on variables used in the parties' sampling designs. For example, the parties may have sampled business establishments with probability proportional to number of employees. For the data to be missing at random in this example, the regression models must include functions of size. The parties may need to rely on external knowledge for specifying these functions. To make the missing at random assumption plausible, the parties should include as many variables as possible in the regression model, and hence in the EM algorithm.

4.3 Model diagnostics for linear regression

Because some data are missing and observed data are not shared, the usual residual analyses are complicated. One approach is to perform diagnostics only using the cases with complete information. This requires the party with the dependent variable to compute residuals for the complete cases. To do so, the parties use a secure summation protocol to determine the predicted values for each record (Karr *et al.*, 2004a). The party with the dependent variable then can examine the residuals of the complete-cases to look for outliers or non-normally distributed residuals. Additionally, the party with the dependent variable can use secure inner-products with other parties to calculate the correlations between the residuals and the independent variables for the complete cases. When these correlations are not close to zero, it suggests the regression assumptions are violated.

It is possible to use synthetic diagnostics (Reiter, 2003b) to generate plots of residuals versus predictors for the complete-cases. The party with the residuals for the complete-cases first standardizes the residuals by dividing them by the residual standard deviation of the regression, then adds random noise to the standardized residuals to limit potential disclosures from the residual values. When using standardized residuals, noise generated from a standard normal distribution should be sufficient for many settings (Reiter, 2003b). It may be necessary to top-code outlying standardized residuals, e.g. report then as "greater than 4," rather than report their exact values. Each party can plot the synthetic residuals against its independent variables for the complete cases to check for non-random patterns. The parties can share the conclusions from their investigations, and if necessary make adjustments to the models.

4.4 Overlapping variables

Sometimes more than one party has units with particular missing data patterns. For example, suppose two parties collect the same variables on non-overlapping records, while a third party collects different variables on records that overlap partially with the other two parties' variables. The summary statistics for overlapping variables are computed by each party for their records using summations and secure inner-products. Then, the parties add their quantities to the appropriate cells of the summary statistics tables. This addition could be done by secure summation if protecting each party's sum is required.

4.5 Handling data missing not by design

When no party has a value of a particular datum (e.g., the survey respondent didn't give an answer), that record can be placed into an appropriate incomplete data pattern. This simply adds new patterns to the secure EM; the computations do not change. Disclosures could occur when such units' missing data groups with one or few records. The regression model needs to be sufficiently detailed to make the missing at random assumption plausible for these values.

5 Concluding Remarks

The secure EM algorithms presented here are a first step towards methods of secure data integration. There are technical and implementation-related issues that need to be addressed before the approach can be adopted in practice.

The underlying assumptions of the secure EM approach suggest directions for future research. In some data sets there may be inexact matching, so that methods for incorporating matching errors need to be developed. The multivariate normal assumption is implausible for some data, so that methods

for non-normal data need to be developed. These could take the form of secure EMs for other models or secure Bayesian posterior simulation. The latter is complicated by the restriction on sharing data values. A related approach is for parties to simulate and share synthetic data (Raghunathan *et al.*, 2003; Reiter, 2003a, 2005). The synthetic data also could serve as public-use data.

Acknowledgements

This research was supported by NSF grant IIS-0131884 to the National Institute of Statistical Sciences.

References

Beale, E. M. L. and Little, R. J. A. (1975). Missing value in multivariate analysis. *Journal of the Royal Statistical Society Series B* **37**, 129–145.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004a). Privacy preserving analysis of vertically partitioned data using secure matrix protocols. Tech. rep., National Institute of Statistical Sciences.

Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2004b). Secure regressions on distributed databases. *Journal of Computational and Graphical Statistics* forthcoming.

Little, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values. *Journal of the Royal Statistical Society Series B* **41**, 76–87.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Meng, X. L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: the sem algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2003a). Inference for partially synthetic, public use microdata sets. *Survey Methodology* 181–189.

Reiter, J. P. (2003b). Model diagnostics for remote access servers. *Statistics and Computing* **13**, 371–380.

Reiter, J. P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* forthcoming.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2004). Privacy preserving regression modelling via distributed computation. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 677–682.

Schafer, J. L. (2000). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.