“Secure” Log-Linear and Logistic Regression Analysis of Distributed Databases

Stephen E. Fienberg,[12] William J. Fulp,[1] Aleksandra B. Slavkovic,[3] and Tracey A. Wrobel[3]

[1] Department of Statistics, Carnegie Mellon University
[2] Cylab and Machine Learning Department, Carnegie Mellon University
[3] Department of Statistics, Pennsylvania State University

Abstract. The machine learning community has focused on confidentiality problems associated with statistical analyses that “integrate” data stored in multiple, distributed databases where there are barriers to simply integrating the databases. This paper discusses various techniques which can be used to perform statistical analysis for categorical data, especially in the form of log-linear analysis and logistic regression over partitioned databases, while limiting confidentiality concerns. We show how ideas from the current literature that focus on “secure” summations and secure regression analysis can be adapted or generalized to the categorical data setting.

1 Introduction

There are many scientific or business settings which require statistical analysis that “integrate” data stored in multiple, distributed databases. Unfortunately, there can be barriers to simply integrating the databases. In many cases, the owners of the distributed databases are bound by confidentiality to their data subjects, and cannot allow outsiders access. This paper discusses various techniques which can be used to perform statistical analysis for categorical data, especially in the form of log-linear analysis and logistic regression over partitioned databases, while limiting confidentiality concerns. The technique used depends on how the database is partitioned, either horizontally (with the same variables but different cases) or vertically (with the same cases but different variables), and also whether log-linear or logistic regression analysis is the goal. This paper will focus primarily on horizontally partitioned databases, and especially on the fully categorical data situation in which case the minimal sufficient statistics are marginal totals and logistic regression is effectively equivalent to log-linear model analysis (e.g., see [1,3,7]).

Much of the literature on privacy-preserving data mining and secure computation has focused on regression problems. A subset of the technical issues relevant to those problems are of interest in this paper. In the vertically partitioned case the concern remains the same, that is specifying a full model based on all of the variables. But in the the horizontally partitioned case there is a new element, whether any single owner actually has enough data to get maximum likelihood estimates (MLEs)! For regression problems this is primarily an issue of identification and we usually require that the sample size n is greater that the number of variables p, although as n increases we get greater accuracy for our regression coefficients and our inferences. But for categorical data problems we will often need to deal with a different form of degeneracy due to sparse data—that associated with patterns of zero counts which yield MLEs on the boundary of the parameter space and thus "do not exist" (for details on existence see especially [6,9,23]). Thus a very important reason for entering into arrangements to do secure computation is that pooled sufficient statistics and tables may well produce existence when no single party has sufficient data to assure the same.

While this paper will focus primarily on log-linear modeling and logistic regression for horizontally partitioned databases, there has been a lot of recent work on broader literature related to partitioned databases. The National Institute of Statistical Sciences (NISS) has produced much work for securely combining a horizontally partitioned database and on performing linear regression analysis on a horizontally partitioned database without actually integrating the data (e.g., see [15,16,21]). Theory regarding performing linear regression on vertically partitioned databases has also been devoloped (e.g., see [14]). There has also been work exploring some broader issues of the privacy impact of data mining methods and their work is related to the literature on secure multi-party computation (e.g., see [27]). Specifically, Kantarcioglu and Clifton [13] discuss mining of association rules on horizontally partitioned database, while the work of Vaidya et al. [25] relates to mining for association rules on vertically partitioned database.

The paper is organized in the following manner. In the next section we present a formulation of the general problem. Then, in Section 3, we turn to the problem of secure computation for log-linear models (and logit models) over horizontally partitioned databases and we relate some of the ideas to the literature on disclosure limitation for single databases involving such data. In Section 4 we present a technique for dealing with logistic regression over horizontally partitioned databases and we contrast it with the approach from section 3 in the case of categorical predictors. We conclude with a discussion of distributed database techniques and other ongoing work.

2 Problem Formulation

Consider a "global" database that is partitioned among a number of parties or "owners." These owners could be thought of as companies or people who have distinct parts of the global database. In a statistical context, these owners are referred to as agencies. These agencies may want to perform log-linear or logistic analysis on the global database, but are unable or unwilling to combine the databases for confidentiality or other proprietary reasons. The goal is to share the statistical analysis as if the global database existed, without actually creating it in a form that any of the owners can identify and utilize.

2.1 Partitioned Database Types

There are two types of partitioned databases discussed in this paper, horizontally and vertically partitioned databases. We are going to assume there are K agencies with $K \geq 2$, but note that a case with $K = 2$ is often trivial for security purposes. Horizontally partitioned data is the case such that agencies share the same fields but not the same individuals, or subjects. Assume the data consist of vectors $\mathbf{X}$ and $\mathbf{Y}$, such that:

$$\mathbf{X'} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \cdots, \mathbf{X^{(k)}}] \text{ and } \mathbf{Y'} = [\mathbf{Y^{(1)}}, \mathbf{Y^{(2)}}, \cdots, \mathbf{Y^{(k)}}], \tag{1}$$

and $\mathbf{X^{(k)}}$ is the matrix of independent variables, $\mathbf{Y^{(k)}}$ is the vector of responses, and $n^{(k)}$ is the number of individuals, all that belong to agency k, $k = 1, \ldots, K$. Let $N = \sum_{k=1}^{K} n^{(k)}$. Each $\mathbf{X^{(k)}}$ is an $n^{(k)} \times p$ matrix and we will assume that the first column of each $\mathbf{X^{(k)}}$ matrix is a column of 1's. We will refer to $\mathbf{X}$ and $\mathbf{Y}$ as the "global" predictor matrix and the "global" response vector respectively ([22]). For horizontally partitioned databases it is assumed that agencies all have the same variables, and that no agencies share observations. Also, the attributes need to be in the same order.

In vertically partitioned data, agencies all have the same subjects, but different attributes. Assume the data looks like the following:

$$[\mathbf{YX}] = \left[\mathbf{Y}\ \mathbf{X^{(1)}} \ldots \mathbf{X^{(k-1)}}\right], \tag{2}$$

where $\mathbf{X^{(k)}}$ is the matrix of a distinct number of independent variables on all N subjects, $\mathbf{Y}$ is the vector of responses, and $p^{(k)}$ is the number of variables for agency k, $k = 1, \ldots, K$. Note that each $\mathbf{X^{(k)}}$ is an $N \times p^{(k)}$ matrix and we will assume that the first column of the $\mathbf{X^{(1)}}$ matrix is a column of 1's. For vertically partitioned database it is assumed that agencies all have the same observations, and that no agencies share variables. In order to match up a vertically partitioned database, all agencies must have a global identifier, such as social security number. We are currently working on the problem of vertically partitioned data in the categorical data setting but do not report on any results here.

There is a third possible kind of partitioning which goes well beyond the two special cases and corresponds more closely to real-world settings, namely horizontally and vertically overlapping data, perhaps with measurement error. Kohnen et al. [19] treat a special case of this in the form of vertically partitioned, partially overlapping as an incomplete data regression problem, and use the EM algorithm to estimate values of the "missing" data.

3 Secure Computation for Horizontally Partitioned Categorical Databases

Karr et al. [16] outline an approach that allows for secure maximum likelihood estimation for a density belonging to an exponential family. This technique can be used for log-linear model analysis in fully categorical data situations, where the minimal sufficient statistics are sets of marginal totals. This secure maximum likelihood technique uses a process called secure summation, which we describe first and then point out how this fits with the exponential family formulation. We then discuss the implementation for log-linear models as well as a possible way to simply combine the tables securely.

Secure Summation Consider agencies which all have a single number, and would like to know the sum of all their numbers. However, the agencies do not want to reveal their individual number to any other agency. Secure summation is a process where the sum of all the agencies can be securely computed. The basic idea is that one agency adds a random number R to their number v_1 and then reports to the next agency in line $R + v_1$. The second agency adds their number v_2 to the number received and sends $R + v_1 + v_2$ to the third agency. The pattern continues until agency k has computed $R + v_1 + \ldots + v_k$ and gives the number to agency 1. Agency 1 then subtracts R from the total, and shares the number with all of the other agencies. As long as multiple agencies are not colluding, secure summation is a very secure process. For a more detailed description of this process, consult Karr et al. [16]. There are other techniques that have been suggested to eliminate collusion but we do not consider them here.

3.1 Secure Maximum Likelihood Estimation for Exponential Families

Consider a global database $\{x_i\}$ modeled as independent samples from an unknown density $f(\theta, \cdot)$ belonging to an exponential family:

$$\log f(\theta, x) = \sum_{\ell=1}^{L} c_\ell(x) d_\ell(\theta). \tag{3}$$

Here the $\{d_\ell(\theta)\}$ are known as canonical parameters and the $\{c_\ell(x)\}$ are the corresponding minimal sufficient statistics (MSSs). Then under the assumption of independence of L rows, the global log-likelihood function is

$$\log L(\theta, x) = \sum_{\ell=1}^{L} d_\ell(\theta) \left[\sum_{k=1}^{K} \sum_{x_i \in D_k} c_\ell(x_i) \right], \tag{4}$$

where D_k is the database of owner k.

If the database owners can agree in advance on the model (3), e.g., the log-linear model with no second order interaction, they can use secure summation to compute each of the L terms in (4). Then each agency can maximize the likelihood function however they choose. There remains serious potential confidentiality problems once $L \geq 2$ since the MSSs are not independent of one another and they jointly contain information about the full table. We thus need to check the extent to which this information is sufficient to seriously compromise the confidentiality of any individual in the database — i.e., if one party can identify with sufficiently high probability an individual in another party's database. In what follows we exploit the fact that log-linear models have a discrete exponential family structure.

3.2 Secure Maximum Likelihood for Log-linear Models

The secure maximal likelihood technique can be used for fitting a log-linear model. Consider a three-dimensional model coming from simple multinomial sampling. We are therefore assuming that the total sample size n is fixed. In this situation, the p.d.f. for the multinomial distribution of $\{n_{ijk}\}$ is

$$\frac{n!}{\prod\limits_{i,j,k} n_{ijk}!} \prod_{i,j,k} \left(\frac{m_{ijk}}{n}\right)^{n_{ijk}}, \tag{5}$$

where $\{m_{ijk}\}$ are the expected cell counts. The log-likelihood of the multinomial is readily obtained from the p.d.f. (5) as

$$\text{constant} + \sum_{i,j,k} n_{ijk} \log(m_{ijk}) - n \log(n). \tag{6}$$

Since the first and third term do not depend on the expected cell counts m_{ijk}, we need only to consider the remaining middle term, the kernel of this function. The saturated log-linear model for the expected cell count m_{ijk} is

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}. \tag{7}$$

Substituting for m_{ijk} from (7) into (6), we obtain for the kernel

$$\begin{aligned}\sum_{i,j,k} n_{ijk} \log(m_{ijk}) = Nu &+ \sum_{i} n_{i++} u_{1(i)} + \sum_{j} n_{+j+} u_{2(j)} + \sum_{k} n_{++k} u_{3(k)} \\ &+ \sum_{i,j} n_{ij+} u_{12(ij)} + \sum_{i,k} n_{i+k} u_{13(ik)} \\ &+ \sum_{jk} n_{+jk} u_{23(jk)} + \sum_{i,j,k} n_{ijk} u_{123(ijk)}.\end{aligned} \tag{8}$$

Since the multinomial distribution belongs to the class of discrete exponential family densities, the minimal sufficient statistics (MSSs) are the observed count n-terms adjacent to the unknown parameters, the u-terms. If we consider an unsaturated model the n_{ijk} terms fall out of expression (8), and those terms that remain give the MSSs. These marginal tables can then be used to estimate the cell expectations $\{\hat{m}_{ijk}\}$ under the model. In fact, it is in general multi-way tables that the MSSs correspond to the highest order u-terms in the model and the likelihood equations are found by setting them equal to their expectations (e.g., see [1,3,8,12]). Further, since the multinomial distribution is in the exponential family, working with log-linear models allows us to use the general secure maximum likelihood equation (4). Similarly, if the sampling model was "Poisson" or product-multinomial, the MSSs are essentially the same once we add in any margin fixed by the sampling scheme, and so the same secure computation idea works. In the product-multinomial situation, the log-linear model can be re-expressed as a logit model and this provides a way for dealing with the secure logistic regression computation problem in the fully categorical data case.

Now consider a horizontally partitioned categorical database. Since (4) is satisfied for log-linear models, it is possible to use secure summation to find the global sufficient statistics, which are marginals that correspond to the highest order u-terms in the model. The agencies will use multiple secure summation processes to create the global marginal statistics. The first agency adds a random number to each marginal value agreed to be summed, and then passes the values to the next agency. The second agency adds their numbers to the marginals, and passes them along. Once the first agency receives these it removes the random values and shares the marginals with all the agencies. If only necessary marginals for a specific model are computed through secure summation, the downside of this process is limited model comparison. If we wish to assess the fit of the model, then we can compare it to a larger log-linear model with additional u-terms. Thus we need to compute additional marginal tables in order to estimate the expected values under the larger model. The two models could be compared to see whether the more parsimonious provides an adequate fit to the data.

As we noted above, the MSSs, i.e., the marginal tables, carry information about the full table. This can come in the form of bounds for cell counts, or actual distributions over possible tables, for example see [5,8,10,11,24]. Computing and thus revealing additional combined marginal totals increases the information known about the individual cell in the overall combined table, possibly to an unacceptable level. Thus to protect individual level confidentiality in this setting we need to go beyond secure computation to incorporate methods from the more traditional disclosure limitation literature. There is also related literature on association rule mining, e.g., see [13,27], but it either focuses on the release of a single marginal or the form of the rule without the relevant data which turns out to be marginal totals [11]. Since using the association rule requires data to allow one to make predictions, releasing just the rule is rarely "useful."

3.3 Secure Contingency Table Analysis

Depending on the level of confidentiality, agencies may be willing to create a global contingency table, as long as the sources of data elements remain protected. Once a global contingency table is created, statistical analysis can be performed normally on the full database. A secure contingency table of counts or sums can be created using multiple secure summations. The general process is as described earlier in the paper, but instead of the first agency creating just one random number, the agency will create a random number for each cell in the table. Then the secure summation pattern applied to every cell in the table continues until the first agency gets the table back, removes all of the random cell values, and reports the full contingency table to the other agencies.

Often a categorical database is too large and sparse for this secure summation process to be efficient enough to use. If that is the case, then a secure data integration process can be used to get a list of cells which have non-zero cell counts, c.f., see discussion in [4] on issues with large sparse contingency tables. This general process is summarized later in this paper. The only adjustment for secure contingency table analysis is that the "data" being inserted into the growing database is really a list of non-zero cell counts. Once a list of non-zero cell counts is created, the multiple secure summation process can be used to get the complete table. This way, the agencies only need to use secure summation for a possibly very small subset of cells in a given table.

The secure contingency table process is only effective if the data elements themselves do not reveal from which party they come. This problem of the data revealing their source is one faced by other methodologies on secure data integration, e.g., see [17,27].

Secure Data Integration Secure data integration is the process of securely combining observations of horizontally distributed databases into one data set. The basic secure data integration process consists of agencies incrementally contributing data into a growing database until the full database is complete. The goal of SDI is to combine these databases in a way so that the agencies will not be able to tell which agency a particular observation came from, except of course for the agency which originally had that observation. Karr et al. [16] lay out the secure data integration process in a reasonably complete fashion.

The growing database is passed from agency to agency in a round robin order, but in an order unknown to the agencies. Therefore, a trusted third party must be used, but the data can be encrypted so that only agencies can read the data. As the growing database is passed around, the agencies input a random number of observations into the database. This pattern continues until all the observations are into the growing database. Using this secure data integration process, a database can be securely combined.

4 Logistic Regression Over Horizontally Partitioned Data

In this setting, logistic regression over a horizontally partitioned database is desired. We first explain that logistic regression can be considered as a specific form of the log-linear modeling. Later in the section we explain a specific technique for performing logistic regression over a horizontally partitioned database, which is not related to log-linear modeling.

4.1 Logistic Regression From Log-Linear Modeling

It is possible to use the approach above for log-linear models to do secure logistic analysis if all the explanatory variables are categorical. Consider simple logistic regression case with a single binary response variable. We can represent the data in contingency table form. The linear logistic regression model for this problem is essentially identical to the *logit* model found by differencing the log expectations for the two levels of the response variable, c.f. [3,7].

We illustrate for a logistic regression model with two binary explanatory variables (variables 1 and 2) and a binary response variable (variable 3). The data form a $2 \times 2 \times 2$ table. We work with the no second-order interaction model and construct the logit, i.e.,

$$\begin{aligned}\text{logit}_{ij} = \log\left(\frac{m_{ij1}}{m_{ij2}}\right) &= \log(m_{ij1}) - \log(m_{ij2}) \\ &= \left[u + u_{1(i)} + u_{2(j)} + u_{3(1)} + u_{12(ij)} + u_{13(i1)} + u_{23(j1)}\right] \\ &\quad - \left[u + u_{1(i)} + u_{2(j)} + u_{3(2)} + u_{12(ij)} + u_{13(i2)} + u_{23(j2)}\right] \\ &= (u_{3(1)} - u_{3(2)}) + (u_{13(i1)} - u_{13(i2)}) + (u_{23(j1)} - u_{23(j2)}). \end{aligned} \tag{9}$$

Since we may place zero-sum constraints over the k index, the logit model in equation (9) simplifies to

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right) = 2u_{3(1)} + 2u_{13(i1)} + 2u_{23(j1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{10}$$

where $x_1 = 1$ for $\log(m_{1j1}/m_{1j2})$ and $x_1 = 0$ for $\log(m_{2j1}/m_{2j2})$ and similarly for x_2. Therefore, performing logistic regression over a horizontally partitioned database can be acheived through the techniques discussed in Section 3.

4.2 Secure Logistic Regression Approach

We now turn to a more general approach for logistic regression over a horizontally partitioned databases using ideas from secure regression (e.g. see [15],[16],[22]). In ordinary linear regression, the estimate of the vector of coefficients is

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}. \tag{11}$$

To find the global $\hat{\beta}$ vector, agency k calculates their own $((\mathbf{X}^{(k)})^t\mathbf{X}^{(k)})$ and $(\mathbf{X}^{(k)})^t\mathbf{Y}^{(k)}$ matrices. The sum of these respective matrices are the global $\mathbf{X}^t\mathbf{X}$ and $\mathbf{X}^t\mathbf{Y}$ matrices. Since the direct sharing of these matrices results in a full disclosure, the agencies need to employ some other method such as secure summation described earlier in the paper. In this secure summation process, the first agency adds a random matrix to its data matrix. The remaining agencies add their raw data to the updated matrix until in the last step the first agency subtracts off their added random values and shares the global matrices. Reiter [22] discusses some possibilities of a disclosure with this method.

We are suggesting to use the developed secure matrix sharing techniques and apply them to the logistic regression setting. We wish to fit a logistic regression

$$\log(\frac{\pi}{1-\pi}) = \mathbf{X}\beta \tag{12}$$

model to the global data, $\mathbf{X}$ and $\mathbf{Y}$. In logistic regression, the vector of coefficients, or β, is of interest, but since the estimate of β cannot be found in closed form, we use Newton-Raphson or a related iterative method. At each iteration of Newton-Raphson, we calculate the new estimate of $\hat{\beta}$ by

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + (\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y} - \mu^{(s)}) \tag{13}$$

where $\mathbf{W}^{(s)} = Diag(n_j\pi_j^{(s)}(1-\pi_j^{(s)}))$, $\mu^{(s)} = n_j\pi_j^{(s)}$ and $\pi_j^{(s)}$ is the probability of a "success" for the j^{th} observation in the iteration s, $j = 1, \cdots, N$. The algorithm stops when the estimate converges. Note that we require an initial estimate of $\hat{\beta}$ (e.g., see [1] for more details).

Now we can apply the secure summation approach to our logistic regression analysis. We can choose an initial estimate for the Newton-Raphson procedure in two ways: (*i*) the parties can discuss and share an initial estimate of the coefficients, or (*ii*) we can compute initial estimates using ordinary linear regression of the responses and predictors using secure regression computations. In order to update β, we need the parts shown in (13). We can break the last term on the right-hand side up into two parts: the $(\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X})^{-1}$ matrix and the $\mathbf{X}^t(\mathbf{Y}-\mu^{(s)})$ matrix. At each iteration of Newton-Raphson, we update the π vector, and thus update the $\mathbf{W}$ matrix and the vector μ. We can easily show that

$$\begin{aligned}\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X} = (\mathbf{X}^{(1)})^t(\mathbf{W}^{(1)})^{(s)}\mathbf{X}^{(1)} + (\mathbf{X}^{(2)})^t(\mathbf{W}^{(2)})^{(s)}\mathbf{X}^{(2)}\\ + \cdots + (\mathbf{X}^{(1)})^t(\mathbf{W}^{(k)})^{(s)}\mathbf{X}^{(k)}\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mathbf{X}^t(\mathbf{Y} - \mu^{(s)}) = \mathbf{X}^{(1)}(\mathbf{Y}^{(1)} - (\mu^{(1)})^{(s)}) + \mathbf{X}^{(2)}(\mathbf{Y}^{(2)} - (\mu^{(2)})^{(s)})\\ + \cdots + \mathbf{X}^{(k)}(\mathbf{Y}^{(k)} - (\mu^{(k)})^{(s)})\end{aligned} \tag{15}$$

where $(\mu^{(k)})^{(s)}$ is the vector of $n_l^{(k)}\hat{\pi}_l^{(k)}$ values and $(\mathbf{W}^{(k)})^{(s)} = Diag(n_l^{(k)}\hat{\pi}_l^{(k)}(1-\hat{\pi}_l^{(k)})$ for agency k, $k = 1, \cdots, K$, $l = 1, \cdots, n^{(k)}$ and for iteration, s. This means that for one iteration of Newton-Raphson, we can find the new estimate of β by using secure summation as suggested by Reiter [22].

One major drawback of this method is that we have to perform secure matrix sharing for every iteration of the algorithm; every time it runs, we have to share the old $\hat{\beta}$ vector with all of the agencies so they may calculate their individual pieces. When all variables are categorical, this method involves more computation than using the log-linear model approach to logistic regression, where only the relevant marginal totals must be shared among the agencies. In the more general setting, we also have no simple way to check on potential disclosure of individual level data and thus we are providing security only for the parties and not necessarily for the individuals in their databases, e.g., see discussion in [22] for the linear regression secure computation problem.

Diagnostics Finding the coefficients of a regression equation is not sufficient; we need to know whether the model has a reasonable fit to the data. One way to assess the fit is to use various forms of model diagnostics such as residuals, but this can potentially increase the risk of disclosure. As with the log-linear model approach we can compare log-likelihood functions of the larger model and the more parsimonious model. The log-likelihood for the logistic regression is:

$$\sum_{j=1}^{N} y_j \{\log(\pi_j) + (1 - y_j) \log(1 - \pi_j)\}. \tag{16}$$

We can rewrite the equation in terms of the K agencies and use secure summation to find this value

$$\sum_{k=1}^{K} \sum_{j=1}^{n^{(k)}} \{y_j^{(k)} \log(\pi_j^{(k)}) + (1 - y_j^{(k)}) \log(1 - \pi_j^{(k)})\}, \tag{17}$$

as well Pearson's χ^2 statistic or the deviance:

$$X^2 = \sum_{k=1}^{K} \sum_{j=1}^{n^{(k)}} \left(\frac{y_j^{(k)} - n_j^{(k)} \pi_j^{(k)}}{\sqrt{n_j^{(k)} \pi_j^{(k)} (1 - \pi_j^{(k)})}} \right)^2 \tag{18}$$

$$G^2 = 2 \sum_{k=1}^{K} \sum_{j=1}^{n^{(k)}} \left\{ y_j^{(k)} \log\left(\frac{y_j^{(k)}}{\hat{\mu}_j^{(k)}} \right) + (n_j^{(k)} - y_j^{(k)}) \log\left(\frac{n_j^{(k)} - y_j^{(k)}}{n_j^{(k)} - \hat{\mu}_j^{(k)}} \right) \right\}. \tag{19}$$

If the change in the likelihood is large with respect to a chi-square statistic with (d.f.) degrees of freedom, we can reject the null hypothesis and conclude that the simpler model provides a better fit to the data.

4.3 Comparison of "Secure" Log-Linear Regression Methods

To demonstrate the difference in computation between the log-linear method for logistic regression and the secure logistic regression method, we will go through a simple example. The example is *not* intended to show how secure the processes are, but *only* to demonstrate the difference between computation in the two methods. Any use of secure summation between just two agencies is useless, because both agencies can simply subtract their number from the final result to find the other agency's data.

The data in Table 1 come from a randomized clinical trial on the effectiveness of an analgesic drug for patients in two different centers and with two different statuses reported on in [18], c.f. Fienberg and Slavkovic [11]. Treatment has 2 levels: Active=1 and Placebo=2. The original response had 3 levels: Poor=1, Moderate=2, and Excellent=3, but for the purposes of this example we combine the last two levels so the response variable is binary: Poor=1 and Not Poor=2.

The data from the first center correspond to Agency 1 and those from the second center to Agency 2 (see Table 1). Consider the possibility that the two centers would like to do statistical analysis over their combined data (see Table 2), but are unwilling to share their individual cell values.

Agency 1 Data				Agency 2 Data			
		Response				Response	
Status	Treatment	1	2	Status	Treatment	1	2
1	1	3	25	1	1	12	12
1	2	11	22	1	2	11	10
2	1	3	26	2	1	3	13
2	2	6	18	2	2	6	12

Table 1. Clinical trial data by Agency.

		Response	
Status	Treatment	1	2
1	1	15	37
1	2	22	32
2	1	6	39
2	2	12	30

Table 2. Combined clinical trial data over the clinical center.

Log-Linear Approach for Logistic Regression We first consider logistic regression from log-linear modeling. We fit the log-linear model with no second order interaction which corresponds to the logistic regression model with no interaction (c.f. Section 4.1, and equation (10)). Note the i index relates to variable Status, the j to Treatment, and the k to Response. The two agencies first use secure summation to compute the 12 marginal totals, i.e., MSSs, n_{ij+}, n_{i+k}, and n_{+jk}. For example, to find n_{11+}, Agency 1 adds some random number to its n_{11+} value of 28, and sends the number to Agency 2. Agency 2 adds their n_{11+} value of 24 and sends the updated value to Agency 1, who subtracts the random number and reveals the total n_{11+} value of 52 (see Table 3 for the relevant marginals.)

ind val	n_{ij+}	n_{i+k}	n_{+jk}
11	52	37	21
12	54	69	76
21	45	18	34
22	42	69	62

Table 3. Relevant marginal values computed through secure summation.

Next, we fit the desired log-linear model in Splus via *loglin* function that uses *iterative proportion fitting* (IPF); it converged in 3 iterations. Table 4 reports 4 relevant log odds values.

Secure Logistic Regression Approach In the secure logistic regression approach, we consider the data in a database form instead of a contingency table. We use the Newton-Raphson algorithm to fit the logistic regression model presented in Equation (12). We used 0s for the initial $\hat{\beta}^{(0)}$ values. Since we know the response variable must be 0 or 1, we would not expect the $\hat{\beta}$ values to be very far from the $(-1, 1)$ interval. The algorithm converged in 4 iterations, and Table 4 reports 4 relevant log odds values.

Log-Linear Model	Logistic Regression
$\log\left\{\frac{\hat{m}_{111}}{\hat{m}_{112}}\right\} = -0.989228$	$\log\left\{\frac{\hat{\pi}_{11}}{1-\hat{\pi}_{11}}\right\} = -0.989230$
$\log\left\{\frac{\hat{m}_{121}}{\hat{m}_{122}}\right\} = -0.305730$	$\log\left\{\frac{\hat{\pi}_{12}}{1-\hat{\pi}_{12}}\right\} = -0.305717$
$\log\left\{\frac{\hat{m}_{211}}{\hat{m}_{212}}\right\} = -1.707879$	$\log\left\{\frac{\hat{\pi}_{21}}{1-\hat{\pi}_{21}}\right\} = -1.707895$
$\log\left\{\frac{\hat{m}_{221}}{\hat{m}_{222}}\right\} = -1.024381$	$\log\left\{\frac{\hat{\pi}_{22}}{1-\hat{\pi}_{22}}\right\} = -1.024382$

Table 4. The estimated log odd ratios from the two different models.

Comparison of the Two Approaches The results for the two approaches as reported in Table 4 agree as expected, but there is a significant computational difference. In the log-linear approach to logistic regression the agencies only need to perform one round of secure summation during this entire process to compute the relevant marginal values. After the relevant marginals have been revealed, the agencies can perform the analysis with them, and do not need to share any information again, thus reducing computations.

The secure logistic regression approach is computationally more intensive than the log-linear method since the agencies need to do secure summation at each iteration of the Newton-Raphson algorithm. Also, in real life settings, the data are likely to be more complex, meaning more iterations needed. This would make the secure logistic regression approach relatively even slower.

5 Conclusion

We have outlined a pair of approaches to carry out "valid" statistical analysis for log-linear model logistic regression of horizontally partitioned databases that does not require actually integrating the data. This allows parties (e.g., statistical agencies) to perform analyses on the global database while not revealing to one another details of the global database beyond those used for the joint computation. For the fully categorical data case we noted that log-linear models provided an alternative approach to logistic regression and one which also allowed us to respect the confidentiality of the data subjects. We also outlined a possible way to securely create a contingency table for horizontally partitioned categorical databases.

We are still developing ideas for logistic regression and log-linear models for strictly vertically partitioned databases and we would like to move towards problems involving partially overlapping data bases with measurement error.

ACKNOWLEDGMENTS

The research reported here was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences, by NSF Grant SES-0532407 to the Department of statistics, Pennsylvania state University, and by Army contract DAAD19-02-1-3-0389 to CyLab at Carnegie Mellon University.

References

1. Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley, New York.
2. Bertino, E., Igor Nai Fovino, I.N., and Provenza, L.P. (2005). "A framework for evaluating privacy preserving data mining algorithms," *Data Mining and Knowledge Discovery*, 11, 121–154.
3. Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Therory and Practice.* MIT Press, Cambridge, MA.
4. Dobra, A., Fienberg, S.E., Rinaldo, A., and Zhou, Yi. (2006). "Confidentiality Protection and Utility for Contingency Table Data: Algorithms and Links to Statistical Theory," Unpublished manuscript.
5. Dobra, A., Fienberg, S.E., and Trottini, M. (2003). "Assessing the risk of disclosure of confidential categorical data," in J. Bernardo et al., eds., *Bayesian Statistics 7*, Oxford University Press, 125–144.
6. Eriksson, N., Fienberg, S.E., and Rinaldo, A., and Sullivant, S. (2006). "Polyhedral conditions for the non-existence of the MLE for hierarchical log-linear models," *Journal of Symbolic Computation,* 41, 222–233.
7. Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.
8. Fienberg, S.E. (2004). "Datamining and Disclosure Limitation for Categorical Statistical Databases," *Proceedings of Workshop on Privacy and Security Aspects of Data Mining, Fourth IEEE International Conference on Data Mining* (ICDM), Brighton, UK.
9. Fienberg and Rinaldo (2006). "Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation," *Journal of Statistical Planning and Inference*, to appear.
10. Fienberg, S.E. and Slavkovic, A.B. (2004a). "Making the release of confidential data from multi-way tables count," *Chance*, 17(3), 5-10.
11. Fienberg, S.E. and Slavkovic, A.B. (2005) Preserving the Confidentiality of Categorical Statistical Data Bases When Releasing Information for Association Rules, *Data Mining and Knowledge Discovery Journal.* 11(2), 155-180.
12. Haberman, S. J. (1974). *The Analysis of Frequency Data,* University of Chicago Press, Chicago, Illinois.
13. Kantarcioglu, M. Clifton, C. 2004. "Privacy preserving data mining of association rules on horizontally partitioned data," *Transactions on Knowledge and Data Engineering*, 16, 1026–1037.
14. Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004). "Privacy preserving analysis of vertically partitioned data using secure matrix products," *J. Official Statist*, Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.
15. Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2005a). "Secure regressions on distributed databases," *Journal of Computational and Graphical Statistics*, 14, 263– 279.
16. Karr, A.F., Fulp, W.J., Vera, F., Young, S.S. (2005b). "Secure, Privacy-Preserving Analysis of Distributed Databases." Available on-line at www.niss.org/dgii/techreports.html.
17. Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2006). "Secure statistical analysis of distributed databases," In *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication.* Edited by A. Wilson, G. Wilson, and D. Olwell. Springer, New York.

18. Koch, G., Amara, J., Atkinson, S. and Stanish, W. 1983. Overview of categorical analysis methods. SAS-SUGI, 8:785-795.
19. Kohnen, C.N., Reiter, J.P., Karr, A.F., Lin, X., Sanil, A.P. (2005). "Secure regression for vertically partitioned, partially overlapping data," Available on-line at www.niss.org/dgii/techreports.html.
20. Reiter, J.P. (2003). Model diagnostics for remote access regression servers. *Statistics and Computing*, 13, 371–380.
21. Reiter, J.P. (2004). Secure regression on distributed databases. Unpublished manuscript.
22. Reiter, J.P. and Kohnen, C. (2005). "Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75, 889–903.
23. Rinaldo, A. (2005). *Maximum Likelihood Estimation for Log-linear Models.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
24. Slavkovic, A.B. (2004) "Statistical disclosure limitation with released marginal and conditionals for contingency tables," *Proceedings of Workshop on Privacy and Security Aspects of Data Mining ICDM 04.* IEEE Computer Society Press, 13-20.
25. Vaidya, J. and Clifton, C. (2002). "Privacy preserving association rule mining in vertically partitioned data," *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.
26. Vaidya, J.; Clifton, C. (2004). "Privacy-preserving data mining: Why, how, and when," *IEEE Security and Privacy*, 2 No. 6, 19–27.
27. Vaidya, J., Clifton, C., Zhu, M. (2006). *Privacy Preserving Data Mining.* Springer Verlag, New York.