NISS

Secure computation with horizontally partitioned data using adaptive regressive splines

Joyee Ghosh, Jerome P. Reiter, and
Alan F. Karr

Technical Report Number 160
June 2006

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

Secure computation with horizontally partitioned data using adaptive regression splines

Joyee Ghosh[a], Jerome P. Reiter[a], Alan F. Karr[b]

[a] *Duke University, Durham, NC, USA.*
[b] *National Institute of Statistical Sciences, Research Triangle Park, NC, USA.*

Abstract

When several data owners possess data on different records but the same variables, known as horizontally partitioned data, the owners can improve statistical inferences by sharing their data with each other. Often, however, the owners are unwilling or unable to share because the data are confidential or proprietary. Secure computation protocols enable the owners to compute parameter estimates for some statistical models, including linear regressions, without sharing individual records' data. A drawback to these techniques is that the model must be specified in advance of initiating the protocol, and the usual exploratory strategies for determining good-fitting models have limited usefulness since the individual records are not shared. In this paper, we present a protocol for secure adaptive regression splines that allows for flexible, semi-automatic regression modeling. This reduces the risk of model misspecification inherent in secure computation settings. We illustrate the protocol with air pollution data.

Key words: Confidentiality, Disclosure, Regression, Secure computation, Spline

1 Introduction

In many contexts, national statistical agencies, survey organizations, businesses, and other data owners (henceforth all called agencies) with related databases can benefit by combining their data. Agencies can use more records or more attributes to fit statistical models when databases are combined than when databases are analyzed separately. Generally, there are two types of data integration settings. Horizontally partitioned databases comprise the same attributes for disjoint sets of data subjects. For example, several local educational agencies might want to combine their students' data to improve the

precision of analyses of the general student population. Vertically partitioned databases comprise the same data subjects, but each database contains different sets of attributes. For example, one agency might have employment information, another health data, and a third information about education, all for the same individuals. A statistical analysis predicting health status from all three sources of attributes is more informative than, or at least complementary to, separate analyses from each data source.

Often the agencies are not able or willing to combine their databases because of concerns about data confidentiality. These concerns can be present even when the agencies cooperate: all may wish to perform integrated analyses, but no one wants to break the confidentiality of others' data. In such cases, the agencies can perform analyses on the combined data without actually sharing the data by utilizing secure computation techniques. Some of these techniques include secure linear regression analyses (Du et al., 2004; Karr et al., 2004b,a), secure data mining with association rules (Kantarcioglu and Clifton, 2002; Vaidya and Clifton, 2002; Evfimievski et al., 2004), and secure model based clustering (Vaidya and Clifton, 2003; Lin et al., 2004). The literature on privacy-preserving data mining (Agrawal and Srikant, 2000; Lindell and Pinkas, 2000) contains related results.

In this paper, we focus on secure regression with horizontally partitioned data, i.e. predicting a continuous attribute from multiple predictors when different subjects are owned by several agencies. The existing protocols for secure regression in this context have a fundamental limitation: they require the agencies to specify the model before starting the protocol. Since the agencies have no opportunities to explore the combined dataset—doing so requires sharing the data, which results in breaches of confidentiality—this limitation increases the risks of model mis-specifications. For example, the agencies may not be able to determine that polynomial terms or other transformations are needed to obtain a good fit. Such transformations may not be evident in each agency's data, particularly when the agencies' data are in different regions of the predictor space.

We propose a protocol for secure regression that uses adaptive regression splines. Adaptive splines provide a flexible, semi-automatic way of fitting regression models, so that agencies employing secure regression spline protocols are less susceptible to model mis-specification than those employing secure linear regression protocols. The paper is organized as follows. Section 2 reviews secure summation and secure linear regression. Section 3 presents the protocol for secure adaptive regression splines. Section 4 illustrates an application of the protocol and compares its predictive performance with secure linear regression. Section 5 concludes with a discussion of this approach.

2 Secure summation and secure linear regression

Various assumptions about the participating agencies are possible, for example, whether they use "correct" values in the computations, follow computational protocols, or collude against one another. We assume the agencies wish both to cooperate and to preserve the privacy of their individual databases. We assume that the agencies are "semi-honest:" each follows the agreed-on computational protocols properly, but may retain the results of intermediate computations. The results of analyses of horizontally partitioned data are shared among all participating agencies and possibly disseminated to the broader public.

2.1 Secure summation protocol

Consider $K > 2$ cooperating, semi-honest agencies, such that Agency a has a value v_a. The agencies wish to compute $v = \sum_{a=1}^{K} v_a$ so that each Agency a learns only the minimum possible about the other agencies' values, namely the value of $v_{(-a)} = \sum_{\ell \neq a} v_\ell$. The secure summation protocol (Benaloh, 1987) can be used to perform this computation.

Following the presentation in Karr et al. (2004b), let m be a very large number—which is known to all the agencies—such that $0 \leq v < m$. One agency is designated the master agency and numbered 1. The remaining agencies are numbered $2, \ldots, K$. Agency 1 generates a random number R from $[0, m)$. Agency 1 adds R to its local value v_1 and sends the sum $s_1 = (R + v_1) \bmod m$ to Agency 2. Since the value R is chosen randomly from $[0, m)$, Agency 2 learns essentially nothing about the actual value of v_1.

For the remaining agencies $a = 2, \ldots, K-1$, the algorithm is as follows. Agency a receives $s_{a-1} = (R + \sum_{t=1}^{a-1} v_t) \bmod m$, from which it can learn nothing about the actual values of $v_1, \ldots, v_{a-1}$. Agency a then computes and passes on to Agency $a+1$ the quantity $s_a = (s_{a-1} + v_a) \bmod m = (R + \sum_{t=1}^{a} v_t) \bmod m$. Finally, agency K adds v_K to $s_{K-1} (\bmod\ m)$, and sends the result s_K to agency 1. Agency 1, which knows R, then calculates v by subtraction, $v = (s_K - R) \bmod m$, and shares this value with the other agencies.

For cooperating, semi-honest agencies, the use of arithmetic mod m may be superfluous. It does, however, provide one layer of additional protection: without it, a large value of s_1 would be informative to Agency 2 about the value of R.

This method for secure summation faces an obvious problem if some agencies collude. For example, agencies $j-1$ and $j+1$ can together compare the values

they send and receive to determine the exact value for v_j. Secure summation can be extended to work for an honest majority. Each agency divides v_j into shares. The sum for each share is computed individually. However, the path used is altered for each share so that no agency has the same neighbor twice. To compute v_j, the neighbors of agency j from every iteration would have to collude.

2.2 Secure linear regression via secure summation

Suppose the K agencies wish to combine their data to fit a pre-specified linear regression model,

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \tag{1}$$

where $f(\epsilon_i|x_{i1}, \ldots, x_{ip}) = N(0, \sigma^2)$ for all observations $i = 1, \ldots, n$. The least squares estimate for β is of course $\hat{\beta} = (X'X)^{-1}X'Y$, where

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{2}$$

When the data are horizontally partitioned across K agencies, each agency a has n_a records of X,

$$X^a = \begin{bmatrix} 1 & x^a_{11} & \ldots & x^a_{1p} \\ \vdots & \ddots & \vdots & \\ 1 & x^a_{n_a 1} & \ldots & x^a_{n_a p} \end{bmatrix}, \qquad Y^a = \begin{bmatrix} y^a_1 \\ \vdots \\ y^a_{n_a} \end{bmatrix}. \tag{3}$$

Using (3) and altering indexes as appropriate, we can rewrite (2) in partitioned form as

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix}, \qquad Y = \begin{bmatrix} Y^1 \\ \vdots \\ Y^K \end{bmatrix}. \tag{4}$$

To compute $\hat{\beta}$, it is necessary to compute $X'X$ and $X'Y$. Each agency a can compute locally its own $(X^a)'X^a$ and $(X^a)'Y^a$. Since $X'X = \sum_{a=1}^{K}(X^a)'X^a$ and $X'Y = \sum_{a=1}^{K}(X^a)'Y^a$, the results can be added entry-wise using secure summation to yield $X'X$ and $X'Y$, which then can be shared among all the

agencies. This provides all the pieces necessary for each agency to compute $\hat{\beta}$. The estimate of σ^2 also can be computed securely.

It is possible to share via secure summation statistics useful for model diagnostics, including correlations between predictors and the residuals, the coefficient of determination R^2, and the hat matrix $X(X'X)^{-1}X'$. Values of residuals are risky to share, since they reveal information about the dependent variable. Karr et al. (2004b) describe an approach for simulating plots of residuals versus predictors that mimic the real-data plots, based on the techniques of Reiter (2003), which can be used for model diagnostics without releasing genuine residuals. When model diagnostics indicate lack of fit, the secure summation protocol is initiated again with adjusted models. However, running the protocol repeatedly is computationally expensive and generates additional confidentiality risks.

3 Secure adaptive regression splines

In many datasets, the initial form of the model used in the secure summation protocol may not adequately describe the data. It is desirable to use procedures that can fit a variety of data structures with one round of secure computations. To do this we develop a protocol for secure computations using adaptive regression splines (Friedman and Silverman, 1989; Friedman, 1991; Hastie and Tibshirani, 1990; Hastie et al., 2001).

3.1 Adaptive regression splines

For adaptive regression splines, we assume an additive model relating the response to the predictors,

$$y_i = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i, \tag{5}$$

where $E(\epsilon_i|x_{i1}, \ldots, x_{ip}) = 0$ for $i = 1, \ldots, n$. The f_j are piecewise linear functions joined continuously at points called knots. When the number and location of the knots are specified, the model is essentially a linear regression with predictors corresponding to the piecewise linear functions of X. Arbitrarily selecting the knots, however, can degrade performance of the regression splines: too few may not capture the relationship and too many may lead to overfitting. Therefore, it is common to select the knots based on the data values, which is the idea of adaptive regression splines.

We utilize multivariate adaptive regression splines (Friedman, 1991), abbreviated as MARS, to select the knots. To simplify explanation, we do not include

interactions among predictors, although this is straightforward to implement. The resulting version of MARS is equivalent to the TURBO procedure of Friedman and Silverman (1989).

For $j = 1, \ldots, p$, let $X_j = \{x_{1j}, \ldots, x_{nj}\}$ be the vector of data for variable x_j. For $i = 1, \ldots, n$ and $j = 1, \ldots, p$, let λ_{ij} be a vector of length n with each element equal to x_{ij}. We form the collection of all piecewise linear basis functions,

$$BF = \{(X_j - \lambda_{ij})_+, (\lambda_{ij} - X_j)_+ \ : i = 1 \ldots, n; j = 1, \ldots, p\}. \tag{6}$$

The "+" denotes the positive part of the expression, and negative differences are set to zero.

Given some full rank subset of basis functions $B \subset BF$, the linear predictor in (5) can be written as

$$E(y|x_1, x_2, \ldots x_p) = \beta_0 + \sum_{j=1}^{p} B_j'(x_j)\beta_j \tag{7}$$

where B_j and β_j are, respectively, the vectors of basis functions and parameters associated with x_j. The β_j are estimated with $(B'B)^{-1}B'Y$.

The key is to select the knots from the very large set BF, which can be done in a stepwise manner as suggested by Friedman (1991). Beginning with the intercept-only model, use a forward stepwise selection procedure to select basis functions from BF until a pre-specified number of terms, T are added. The value of T is chosen to be large, so that the end model tends to over-fit the data. Perform backwards elimination steps until returning to the intercept-only model, dropping the knot at each stage that produces the smallest increase in residual error. This produces a sequence of T estimated models. The final model is the one that minimizes a generalized cross-validation criterion.

3.2 Secure MARS

For horizontally partitioned confidential data, it is not possible to construct BF. It requires all observed predictor values, which amounts to the K agencies sharing and revealing their data values. We propose an approximation to the stepwise MARS procedure that circumvents this issue.

The protocol is as follows. First, each agency uses the forward and backwards steps of MARS to determine the optimal set of knots for its own data, $\{\lambda^a\}$, for $a = 1 \ldots, K$. Second, the agencies share their optimal knot values with each other, forming the superset of knots $\Lambda = \{\lambda^1, \ldots, \lambda^K\}$. This can be done

directly or, for an extra layer of confidentiality, via secure data integration (Karr et al., forthcoming), which enables agencies to share knots without revealing the source agency of each knot. Third, each agency constructs its basis matrix B^a, where

$$B^a = \{(X_j^a - \lambda_{ij})_+, (\lambda_{ij} - X_j^a)_+ \ : \lambda_{ij} \in \Lambda\}. \tag{8}$$

Fourth, using the basis matrices $(B^1, \ldots, B^K)$, the agencies use the secure regression protocol of Section 2.2 to compute $(B'B)^{-1}B'Y$, where

$$B = \begin{bmatrix} B^1 \\ \vdots \\ B^K \end{bmatrix}, \qquad Y = \begin{bmatrix} Y^1 \\ \vdots \\ Y^K \end{bmatrix}. \tag{9}$$

Finally, the agencies use a backwards selection to select the final model, for example using an AIC or BIC criteria that penalizes for the number of knots. These criteria values can be computed separately by each agency—after they share $Y'Y$ using secure summation—without additional rounds of secure summation, because the coefficients from any sub-model can be obtained from the appropriate sub-matrices of $B'B$ and $B'Y$.

The protocol requires some additional features to ensure full rank computations. When two or more agencies come up with a repeated knot value, only one of them is kept by the collective. Similarly, when any x_j is split at $(x_j - \lambda_{ij})_+$ and at $(\lambda_{ij} - x_j)_+$ by different agencies, only one of these pairwise basis functions can be included in B when both pairs of another knot λ_{hj}, $h \neq i$, are selected for x_j. These features can be coded in the backward elimination algorithm.

Because knot values are released without ties to identifiers, the risks to data confidentiality from releasing knot values should be low. If any agency fears that a record could be identified by the release of one of its data values as a knot, the agency could choose not to release that knot. Releasing nearby values, or possibly adding fictitious knots to make it hard to determine whether all released knots are genuine values, may preserve the utility of the regression spline and protect that record's confidentiality.

4 Illustrative simulation

We illustrate secure MARS using the ozone data used by Friedman and Silverman (1989). The data comprise measurements on ground level ozone concentrations and meteorological variables taken on 330 days in the Los Angeles

Table 1
Description of variables used in illustrative simulation

Name	Description
day	the day of the year
dpg	the pressure gradient (mm Hg)
hum:	the humidity (in percent)
ibh	the inversion base height (feet)
ibt	the inversion base temperature (degrees F)
temp	the temperature (degrees F)
vh	altitude at which the pressure is 500 millibars
vis	the visibility (miles)
wind	the wind speed (mph)

basin in 1976. The inferential goal is to predict ground ozone level from nine explanatory variables. The predictors are described in Table 1. Their relationships with the response variable are displayed in Figure 1. Most of the relationships between the response and the predictors are nonlinear, so that multiple linear regression using the original explanatory variables is likely to yield poor predictions.

Although these data were originally collected as one dataset, we use them to simulate a horizontally partitioned data setting. We create $K = 3$ agencies, each of which has 100 randomly partitioned observations. The remaining 30 observations are not used to fit the regressions; rather, we use them as an evaluation dataset for comparisons of the methods. We examine the performance of the secure linear regression of Karr et al. (2004b), the secure adaptive regression spline using both AIC and BIC for backwards elimination, and the adaptive regression spline obtained from the combined data assuming no confidentiality restrictions. All computations involving MARS are performed using the *mda* package in the statistical software R.

Table 2 displays the sum of squared errors based on the training sample and the evaluation sample for two replications of this simulation, with two different partitions and evaluation samples. As expected, the secure linear regression results in the highest prediction errors. The secure regression splines outperform the secure linear regression. They even predict reasonably well relative to using MARS on the combined data. Further replications verified the superior performance of the secure splines over secure linear regression.

Fig. 1. Relationships between response and predictors in the air pollution data

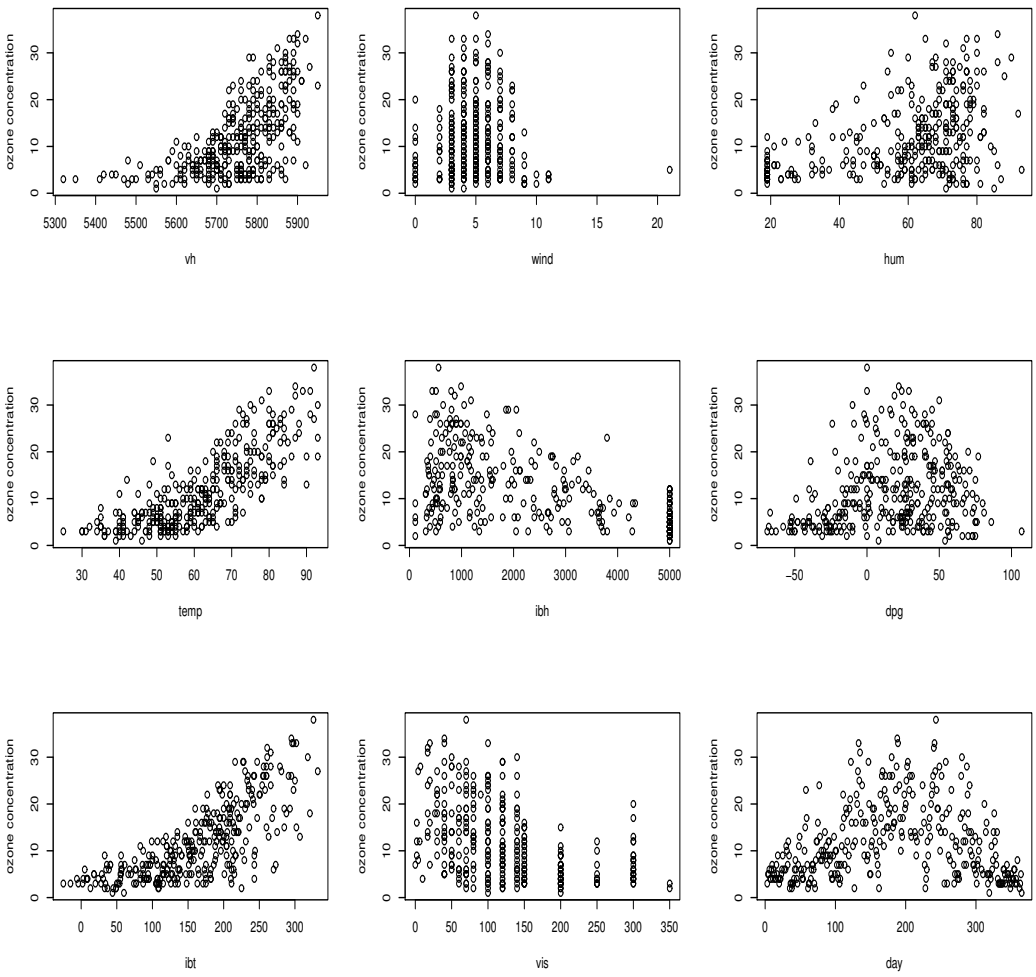

Table 2
Sums of squared errors after using secure regression and secure splines in ozone data

	First replication		Second replication	
	Training	Evaluation	Training	Evaluation
Secure linear regression	5638	709	5864	475
Secure splines, AIC	3611	506	3762	369
Secure splines, BIC	3773	510	3899	393
Spline on full data	3870	609	3849	358

5 Discussion

The secure regression spline protocol provides a flexible, semi-automatic way to implement secure regression in horizontally partitioned data. It reduces

the reliance on model pre-specification, which can be problematic in secure contexts because agencies cannot explore the combined dataset to search for good-fitting models. Hence, when relationships are suspected to be non-linear, agencies are likely to be better off using secure adaptive regression splines over secure linear regression.

The approach presented here focuses on point estimates of predictions. Interval estimates are also desirable. One approach is to generate many bootstrapped datasets, run the protocol on each dataset, then compute prediction intervals from the bootstrapped predictions. This requires repeated applications of the protocol, which are computationally expensive and risky from a confidentiality standpoint.

Secure regression in the vertically partitioned data setting—when data owners possess different variables on the same subjects—faces similar model pre-specification dilemmas. Implementing a secure regression spline for vertically partitioned data is more complicated than for horizontally partitioned data, as there are no obvious starting points for the knots. Given the existence of other algorithms for vertically partitioned data, this is a topic worthy of further research.

6 Acknowledgments

This research was supported by NSF grants EIA–0131884 to the National Institute of Statistical Sciences (NISS) and DMS–0112069 to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Agrawal, R., Srikant, R., 2000. Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD on Management of Data. pp. 439–450.

Benaloh, J., 1987. Secure sharing homomorphisms: Keeping shares of a secret secret. In: Odlyzko, A. M. (Ed.), Advances in Cryptography: CRYPTO86. Vol. 263. New York: Springer-Verlag, pp. 251–260.

Du, W., Han, Y., Chen, S., 2004. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: Proceedings of the Fourth SIAM Conference on Data Mining. pp. 222–233.

Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J., 2004. Privacy-preserving mining of association rules. Information Systems 29, 343–364.

Friedman, J. H., 1991. Multivariate adaptive regression splines (with discussion). The Annals of Statistics 19, 1–141.

Friedman, J. H., Silverman, B. W., 1989. Flexible parsimonious smoothing and additive modeling. Technometrics 31, 3–39.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. New York: Chapman & Hall.

Hastie, T. J., Tibshirani, R. J., Friedman, J. H., 2001. The Elements of Statistical Learning. Springer.

Kantarcioglu, M., Clifton, C., 2002. Privacy-preserving distributed mining of association rules on horizontally-partitioned data. In: Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 24–31.

Karr, A. F., Fulp, W. J., Lin, X., Reiter, J. P., Vera, F., Young, S. S., forthcoming. Secure, privacy-preserving analysis of distributed databases. Technometrics.

Karr, A. F., Lin, X., Reiter, J. P., 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., Reiter, J. P., 2004b. Secure regressions on distributed databases. Journal of Computational and Graphical Statistics 14, 263–279.

Lin, X., Clifton, C., Zhu, Y., 2004. Privacy-preserving clustering with distributed EM models. Knowledge and Information Systems 8, 68–81.

Lindell, Y., Pinkas, B., 2000. Privacy-preserving data mining. In: Advances in Cryptology: CRYPTO2000. New York: Springer-Verlag, pp. 36–54.

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

Vaidya, J., Clifton, C., 2002. Privacy-preserving association mining over vertically-partitioned data. In: Proceedings of Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 639–644.

Vaidya, J., Clifton, C., 2003. Privacy-preserving k-means clustering over vertically-partitioned data. In: Proceedings of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 206–215.