

# Analysis of Integrated Data without Data Integration

Alan F. Karr, Xiaodong Lin, Ashish P. Sanil  
National Institute of Statistical Sciences  
Research Triangle Park, NC 27709–4006, USA  
karr@niss.org, linxd@samsi.info, ashish@niss.org

Jerome P. Reiter  
Duke University  
Durham, NC 27708 USA  
jerry@stat.duke.edu

May 6, 2004

**Introduction.** Many scientific and policy investigations require statistical analyses that “integrate” data stored in multiple, distributed databases. For example, a regression analysis on integrated state databases about factors influencing student performance would be more insightful than individual analyses, or complementary to them. Other contexts where the same need arises range from homeland security to environmental monitoring.

At the same time, the barriers to actually integrating the databases are numerous. One is confidentiality: the database holders—we term them “agencies”—almost always wish to protect the identities of their data subjects. Another is regulation: the agencies may be forbidden by law to share their data, either with each other or with a trusted third party. A third is scale: despite advances in networking technology, the only way to move a terabyte of data from point A today to point B tomorrow is FedEx.

The good news is that for many analyses it is not necessary to move the data. Instead, using techniques from computer science known generically as secure multi-party computation, the agencies can share summaries of the data anonymously, but in a way that the analysis can be performed in a statistically principled manner.

We illustrate in this paper for linear regression on “horizontally partitioned data.” Only one concept is needed, that of secure summation, which is shown pictorially in Figure 1. There are other approaches to this problem for lower risk situations, as well as similar approaches to related problems, such as vertically partitioned data. For example, NISS has developed techniques for secure data integration, which build the integrated database in such a way that no agency can determine the source of any data elements other than its own, at least under the assumption that the data values themselves do not reveal the source of records.

**The Problem.** We assume that there are  $K > 2$  agencies, each with the same numerical data on its own  $n_j$  data subjects— $p$  predictors  $X^j$  and a response  $y^j$ , and that the agencies wish to fit the usual linear model

$$y = X\beta + \epsilon,$$

to the “global” data

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y^1 \\ \vdots \\ y^K \end{bmatrix}.$$

Figure 2 shows such horizontal partitioning for  $K = 3$  agencies. Each  $X^j$  is  $n_j \times p$ .

We embed the constant term of the regression in the first predictor:  $X_1^j \equiv 1$  for all  $j$ . To illustrate the subtleties of analysis of distributed data, the alternative strategy of centering the predictors and response at their means does not work, at least not directly. The means in this case are the global means, which are not available without another round of secure computation.

Under the condition that  $\text{Cov}(\epsilon) = \sigma^2 I$ , the least squares estimator for  $\beta$  is of course

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

This paper shows how  $\hat{\beta}$  can be computed without integrating the agencies’ databases.

Several assumptions about agency behavior are necessary. First, the agencies agree to cooperate to perform the regression, and none of them is specifically interested in breaking the confidentiality of the others’ data. Second, each reports accurately the results of computations on its own data, and follows the agreed-on computational protocols, such as secure summation, properly. And finally, there is no collusion among agencies.

**Secure Summation.** The simplest secure multi-party computation, and essentially the only one needed for secure regression, is to sum values  $v_j$  held by the agencies. Let  $v$  denote the sum. The method described below, which has appeared recently in the puzzles of the radio shows *Car Talk* and *NPR Weekend Edition Sunday*, lets agency  $j$  learn only the minimum possible about the other agencies’ values, namely, the sum  $v_{(-j)} = \sum_{\ell \neq j} v_\ell = v - v_j$ .

The secure summation protocol, which is depicted graphically in Figure 1, is almost more complicated to describe than to implement. Number the agencies  $1, \dots, K$ . Agency 1 generates a very large random integer  $R$ , adds  $R$  to its value  $v_1$ , and sends the sum to agency 2. Since  $R$  is random, Agency 2 learns effectively nothing about  $v_1$ . Agency 2 adds its value  $v_2$  to  $R + v_1$ , sends the result to agency 3, and so on. Finally, agency 1 receives  $R + v_1 + \dots + v_K = R + v$  from agency  $K$ , subtracts  $R$ , and shares the result  $v$  with the other agencies. Here is one place where cooperation matters: agency 1 is obliged to share  $v$  with the other agencies.

Figure 1 contains an extra layer of protection. Suppose that  $v$  is known to lie in the range  $[0, m)$ , where  $m$  is a very large number, say  $2^{100}$ , known to all the agencies. Then  $R$  can be chosen randomly from  $\{0, \dots, m - 1\}$  and all computations performed modulo  $m$ .

Here is a simple application: the agencies have income data and wish to compute the global average income. Let  $n_j$  be the number of records in agency  $j$ ’s database and  $I_j$  be the sum of

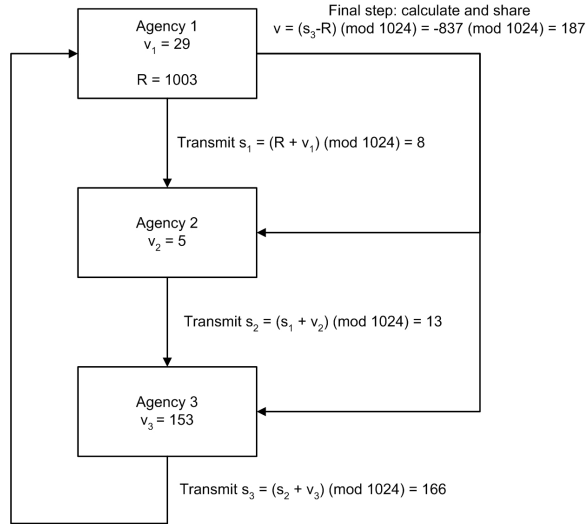


Figure 1: Values computed at each agency during secure computation of a sum initiated by Agency 1. Here  $v_1 = 29$ ,  $v_2 = 5$ ,  $v_3 = 152$  and  $v = 187$ . All arithmetic is modulo  $m = 1024$ .

their incomes. The quantity to be computed is  $\bar{I} = \sum_j I_j / \sum_j n_j$ , whose numerator can be computed using secure summation on the  $I_j$ 's, and whose denominator can be computed using secure summation on the  $n_j$ 's.

**Secure Regression.** To compute  $\hat{\beta}$ , it is necessary to compute  $X^T X$  and  $X^T y$ . Because of the horizontal partitioning of the data,

$$X^T X = \sum_{j=1}^K (X^j)^T X^j.$$

Therefore, agency  $j$  simply computes its own  $(X^j)^T X^j$ , which has dimensions  $p \times p$ , where  $p$  is the number of predictors, and these are combined entrywise using secure summation. This computation is illustrated with  $K = 3$  in Figure 2. Of course, because of symmetry, only  $\binom{p}{2} + p$  secure summations are needed. Similarly,  $X^T y$  can be computed by secure, entry-wise summation of the  $(X^j)^T y^j$ .

Finally, each agency can calculate  $\hat{\beta}$  from the shared values of  $X^T X$  and  $X^T y$ . Note that no agency learns any other agency's  $(X^j)^T X^j$  or  $(X^j)^T y^j$ , but only the sum of these over all the other agencies.

**Example.** We illustrate the secure regression protocol using the “Boston housing data” (Harrison and Rubinfeld, 1978). There are 506 data cases, representing towns around Boston, which we partitioned among  $K = 3$  agencies. The agencies might, for example, represent regional governmental authorities.

The database sizes are  $n_1 = 172$ ,  $n_2 = 182$  and  $n_3 = 152$ . The response  $y$  is median housing value, and three predictors were selected:  $X_1 = \text{CRIME}$  per capita,  $X_2 = \text{IND[USTRIALIZATION]}$ ,

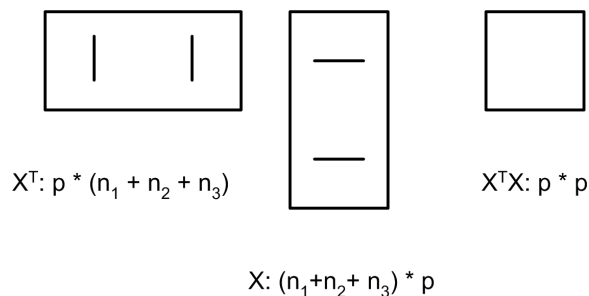


Figure 2: Pictorial representation of the secure regression protocol. The dimensions of various matrices are shown.

the proportion of non-retail business acres, and  $X_3 = \text{DIST}[\text{ANCE}]$ , a weighted sum of distances to five Boston employment centers.

Figure 3 shows the results of the computations performed by the three agencies, of their respective  $(X^j)^T X^j$  and  $(X^j)^T y^j$ . The agencies then use the secure regression protocol to produce the global values

$$X^T X = (X^1)^T X^1 + (X^2)^T X^2 + (X^3)^T X^3 = \begin{bmatrix} 506.00 & 1828.44 & 5635.21 & 1920.29 \\ 1828.44 & 43970.34 & 32479.10 & 3466.28 \\ 5635.21 & 32479.10 & 86525.63 & 16220.67 \\ 1920.29 & 3466.28 & 16220.67 & 9526.77 \end{bmatrix}$$

and

$$X^T y = (X^1)^T y^1 + (X^2)^T y^2 + (X^3)^T y^3 = \begin{bmatrix} 11401.60 \\ 25687.10 \\ 111564.08 \\ 45713.87 \end{bmatrix}.$$

These global objects are shared among the three agencies, each of which can then calculate the estimated values of the regression coefficients.

Figure 4 contains these estimators, as well as, for comparison purposes, the estimators for the three agency-specific local regressions. The intercept is  $\hat{\beta}_{\text{CONST}}$ , the coefficient corresponding to the constant predictor  $X_1$ . Each agency  $j$  ends up knowing both—but only—the global coefficients and its own local coefficients. To the extent that these differ, it can infer some information about the other agencies’ regressions collectively, but not individually. In this example, agency 2 can detect that its regression differs from the global one, but is not able to determine that agency 1 is the primary cause for the difference.

**Model Diagnostics.** In the absence of model diagnostics, secure regression loses much of its appeal, especially to statisticians. We describe briefly two strategies for producing informative diagnostics. The first is to use diagnostics that can be computed using secure summation from corresponding local statistics. The second uses “secure data integration” (Karr et al., 2004) to share synthetic residuals (Reiter, 2003).

Agency $j$	$n_j$	$(X^j)^T X^j$	$(X^j)^T y^j$
1	172	$\begin{bmatrix} 172.00 & 49.03 & 1581.19 & 781.52 \\ 49.03 & 40.42 & 556.29 & 180.95 \\ 1581.19 & 556.29 & 23448.60 & 5631.35 \\ 781.52 & 180.95 & 5631.35 & 4186.07 \end{bmatrix}$	$\begin{bmatrix} 4057.90 \\ 909.24 \\ 32227.19 \\ 18996.12 \end{bmatrix}$
2	182	$\begin{bmatrix} 182.00 & 94.47 & 1563.50 & 746.12 \\ 94.47 & 160.90 & 1433.20 & 231.87 \\ 1563.50 & 1433.20 & 18970.98 & 5224.19 \\ 746.12 & 231.87 & 5224.19 & 3882.02 \end{bmatrix}$	$\begin{bmatrix} 4691.10 \\ 2299.13 \\ 37949.83 \\ 19193.18 \end{bmatrix}$
3	152	$\begin{bmatrix} 152.00 & 1684.95 & 2490.52 & 392.64 \\ 1684.95 & 43769.02 & 30489.61 & 3053.46 \\ 2490.52 & 30489.61 & 44106.05 & 5365.14 \\ 392.64 & 3053.46 & 5365.14 & 1458.68 \end{bmatrix}$	$\begin{bmatrix} 2652.60 \\ 22478.73 \\ 41387.06 \\ 7524.57 \end{bmatrix}$

Figure 3: Illustration of the secure regression protocol for the “Boston housing data” Harrison and Rubinfeld (1978). As discussed in the text, there are three agencies, each of which computes its local  $(X^j)^T X^j$  and  $(X^j)^T y^j$ . These are combined entrywise using secure summation to produce shared global values  $X^T X$  and  $X^T y$ , from which each agency calculates the global regression coefficients.

Among diagnostics computable by secure summation are the coefficient of determination  $R^2$ , the least squares estimate  $S^2 = (y - X\hat{\beta})^T (y - X\hat{\beta}) / (n - p)$  of the error variance  $\sigma^2$ , correlations between predictors and residuals, and the hat matrix  $H = X(X^T X)^{-1} X^T$ , which can be used to identify  $X$ -outliers.

For diagnosing some types of assumption violations, only patterns in relationships among the residuals and predictors suggestive of model mis-specification are needed, rather than exact values of the residuals and predictors. Such diagnostics can be produced for the global database using secure data integration protocols (Karr et al., 2004) to share synthetic diagnostics proposed for remote access computer servers (Gomatam et al., 2003).

The synthetic diagnostics are generated in three steps. First, each agency simulates values of its predictors. Second, using the global regression coefficients, each agency simulates residuals associated with these synthetic predictors in a way—and this is the hard part—that mimics the relationships between the predictors and residuals in its own data. Finally, the agencies share their synthetic predictors and residuals using secure data integration.

**Discussion.** In this paper we have presented a framework for secure linear regression in a cooperative environment. A huge number of variations is possible. For example, in order to give the agencies flexibility, it may be important to give them the option of withdrawing from the computation when their perceived risk becomes too great. To illustrate, agency  $j$  may wish to withdraw if its sample size  $n_j$  is too large relative to the global sample size  $n$ . This is the classical

Regression	$\hat{\beta}_{\text{CONST}}$	$\hat{\beta}_{\text{CRIME}}$	$\hat{\beta}_{\text{IND}}$	$\hat{\beta}_{\text{DIST}}$
Global	35.505	-0.273	-0.730	-1.016
Agency 1	39.362	-8.792	-0.720	-1.462
Agency 2	35.611	2.587	-0.896	-0.849
Agency 3	34.028	-0.241	-0.708	-0.893

Figure 4: Estimated global and agency-specific regression coefficients for the partitioned Boston housing data. The intercept is  $\hat{\beta}_{\text{CONST}}$ .

$p$ -rule in the statistical disclosure limitation literature (Willenborg and de Waal, 2001). But,  $n$  can be computed using secure summation, and so agencies may then “opt out” according to whatever criteria they wish to employ. It is even possible, at least under a scenario that the process does not proceed if any of the agencies opts out, to allow the opting out itself to be anonymous.

**Acknowledgements.** This research was supported by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS).

## References

- Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2003). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.* Submitted for publication. Available on-line at [www.niss.org/dgii/technicalreports.html](http://www.niss.org/dgii/technicalreports.html).
- Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *J. Environ. Econ. Mgt.*, 5:81–102.
- Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004). Secure regression on distributed databases. *J. Computational and Graphical Statist.* Submitted for publication. Available on-line at [www.niss.org/dgii/technicalreports.html](http://www.niss.org/dgii/technicalreports.html).
- Reiter, J. P. (2003). Model diagnostics for remote access regression servers. *Statistics and Computing*, 13:371–380.
- Willenborg, L. C. R. J. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer–Verlag, New York.