

Regression on Distributed Databases via Secure Multi-Party Computation

Alan F. Karr, Xiaodong Lin, Ashish P. Sanil

National Institute of Statistical Sciences

Jerome P. Reiter

Duke University

Abstract

We present a method for performing linear regression on the union of distributed databases that does not entail constructing an integrated database, and therefore preserves confidentiality of the individual databases. The method can be used by statistical agencies to share information from their individual databases, or to make such information available to others.

1 Introduction

In this paper, we show how to perform secure linear regression for “horizontally partitioned data” stored in distributed databases controlled by multiple statistical agencies. The databases that contain the same numerical attributes for disjoint sets of data subjects. Our approach has the additional advantage of being resistant to source identification via attribute values, as discussed in Karr et al. (2004): only data summaries, not data values, are shared.

The approach uses the secure summation protocol, a form of *secure multi-party computation*, to compute the familiar least squares estimators $\hat{\beta} = (X^T X)^{-1} X^T y$ locally. Each agency calculates components of this computation on its own database, and the results are combined in a secure manner.

Assessing the fit of the model is more challenging. Two strategies are outlined in §2.2. One uses global statistics associated with the regression, such as the familiar coefficient of determination R^2 , that can be computed locally. The other uses a secure data integration protocol to build an integrated database of synthetic residuals.

2 Regression using Secure Multi-Party Computation

We assume that the participating agencies wish both to cooperate in order to perform the regression and to preserve the privacy of their individual databases. In particular, each agency wishes to reveal as little as possible in order to effect the regression. Moreover, while an agency can “subtract” its own contribution from integrated computations, it should not be able to distinguish the other agencies’ contributions. We also assume that the agencies are *semi-honest*: they follow agreed-on computational protocols properly, but may retain the results of intermediate computations.

2.1 Secure Linear Regression

We consider the usual linear regression model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon. \quad (1)$$

where y is the response, β_0 is the constant term in the regression, x_1, \dots, x_p are the p predictor variables, and ε is the random error. If the error terms ε_i have the same mean variance and are uncorrelated, then the least squares estimate for β is of course $\hat{\beta} = (X^T X)^{-1} X^T y$.

When the data are horizontally partitioned across K agencies, each agency j has its own share of data, namely, X^j and y^j . To compute $\hat{\beta}$, it is necessary to compute $X^T X$ and $X^T y$. Because of the horizontal partitioning of X and y , this can be done locally and the results combined entry-wise using secure summation. Specifically, since $X^T X = \sum_{j=1}^K (X^j)^T X^j$, each agency j can compute its own $(X^j)^T X^j$, which has dimension $p \times p$, locally, and then the results can be added entry-wise using secure summation to yield $X^T X$, which then can be shared among all the agencies. Similarly, $X^T y$ can be calculated by local computation of the $(X^j)^T y^j$ and secure summation, and shared among all the agencies. Finally, each agency, since it is in possession of $X^T X$ and $X^T y$, can calculate $\hat{\beta}$.

2.2 Model Diagnostics

In the absence of model diagnostics, secure regression would lose much of its appeal to researchers.

Many statistics useful in practice for model diagnosis can be computed using secure summation. These include the coefficient of determination R^2 , the least squares estimate S^2 of the error variance σ^2 , correlations between residuals and predictors and “hat” matrices useful for identifying outliers. It is also possible, using protocols for secure data integration (Karr et al., 2004), to share synthetic residuals that are informative about patterns suggestive of model mis-specification.

Acknowledgements

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

References

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 online at www.niss.org/dgii/technicalreports.html.