Using Regression to Compromise Statistical Databases: A Modification of the Attribute Correlation Modeling Approach

Myun J. Cheon
University of South Carolina

Patrick R. Philipoom
University of South Carolina

It has been shown that statistical databases can be compromised by regression using the Attribute Correlation Modeling method. This method, however, does not work well when only a small portion of the database can be sampled. Some of the shortcomings of the Attribute Correlation Modeling approach are discussed. A modification of the Attribute Correlation Modeling method which addresses these shortcomings is presented and compared to the original approach using Census Data. The modified Attribute Correlation Modeling method significantly outperforms the previous approach.

Introduction

Many organizations collect information for on-line statistical databases. The U.S. Census Bureau, for example, collects information on such items as household income, education level, number of cars per household, etc. To obtain this information, the Census Bureau promises that the information will be released in aggregate form only and that information about any one household will not be disclosed. It has been shown by several researchers that these promises are difficult to keep (Hoffman and Miller, 1970; Schlorer, 1975). Since statistical databases are widely used for epidemiology studies, health care planning, demographics, marketing research and so forth, it is crucial that the compromise risks of these systems be determined.

The focus of this research is the disclosure of information about a model or a population. These are called by Duncan and Lambert (1987) "model disclosure" and "population disclosure," respectively. Duncan and Lambert (1987) in a report for the Internal Revenue Service (IRS) warn that by accessing the IRS database, one could disclose the model used to determine which taxpayers are audited. An example of a population disclosure problem is the availability of government health care records to allow insurance companies to determine the health care risks of various groups so that they can "redline" or refuse coverage to these groups. This is a serious problem especially with government databases that will only get worse in the future as groups use the Freedom of Information Act to force government agencies to allow on-line access to their databases (Grodsky, 1990).

The present research reexamines the vulnerability of statistical databases to compromise using regression. This paper will show that statistical databases can be compromised with less information than previous research would lead one to believe. The emphasis of this research is not to show how to expose confidential information, but rather to illustrate how easily statistical databases can be compromised. The paper proceeds as follows. The literature on compromising statistical databases is reviewed and the Attribute Correlation Modeling (ACM) method is described. Then, the shortcomings of the ACM method are discussed.
and a modified version of this method is presented. The two methods are compared using data from the 1980 U.S. Census on South Carolina. The paper concludes with a summary of the results and their implications.

**Literature Review**

A statistical database is "a collection of records about which queries concerning certain subsets of the records may be answered" (Kam and Ulman, 1977, p. 1). Ulman (1980) defines a statistical database as "a database from which aggregate information about large subsets of entities of an entity set is to be obtained, such as a database of census data, or for certain applications, a file of employees, tax returns or hospital patients" (p. 314). Statistical databases typically contain both parameter data such as social security number or name and measured variables such as income or number of dependents.

The purpose of a statistical database is to allow users access to statistical summaries of information to support activities such as marketing research and economic planning (Fernandez et al., 1981). Statistical summaries may include SUM, COUNT, MEAN and STANDARD DEVIATION of a subset of the database. One of the major problems with statistical databases is that they often contain sensitive and confidential information about individuals, groups or corporations which must not be disclosed. A major security problem for statistical databases is to limit the amount and type of information that may be obtained from a database. The database manager must control user inferences so that it is not possible to infer from responses to legitimate queries information that is explicitly hidden from the user (Leiss, 1982).

The term inference, according to Morganstern (1987), refers to the process of obtaining some information which reduces the degree of freedom of the other data values. For example, knowing a person’s profession gives information that narrows the range of possible salaries that individual. Denning (1980) defines the inference problem as "the deduction of confidential data by correlating the declassified statistical summaries and prior information" (p. 291). Denning (1978) further states that a compromise of confidential information occurs "whenever it is possible to deduce from the responses to one or more queries, information not previously known about an individual" (p. 16). For example, using queries of COUNT and MEAN, it is possible to compare the average salary of two groups differing by only a single person and determine the salary of the person whose record is in one group and not the other (Hoffman and Miller, 1970).

Several authors suggest that user inference or compromise must be defined in relative terms. For example, Traub et. al (1984) state that even though an error of $4,000 may not be considered a compromise if the person’s salary is $20,000, an error of $4,000 would be considered a compromise if the person made $175,000. Beck (1980) makes a similar argument. Thus, what constitutes a compromise is specific to a particular statistical database.

Palley (1986; 1987) and Palley and Simonoff (1987) demonstrate how statistical databases can be compromised using an approach which they call the Attribute Correlation Modeling method (ACM) even if the database manager forbids users from performing regression on the statistical database. By making a series of legitimate queries, a synthetic database can be constructed as an approximation of the statistical database. Once the synthetic database is constructed, the data can be analyzed outside of the control of the database manager. Regressions can be performed on the synthetic database and confidence intervals can be constructed about any point of interest. If the resulting prediction intervals are sufficiently tight, the statistical database is compromised.

Using ACM, a regression model is constructed by selecting r-1 independent variables that are regarded as non-

---

Figure 1: Interim Tuple Table (ITT) Completed