December 1998

Managing Data Quality: Robust and Resistance Tools for a Data Warehousing Environment

A. Schwarzkopf
University of Oklahoma

Managing Data Quality: Robust and Resistant Tools for a Data Warehouse Environment

A. B. Schwarzkopf
Division of Management
The University of Oklahoma

Poor data quality is one of the fundamental obstacles to effective use of data warehouse approaches. Complete editing of all corporate archived data is impractical so approaches to data applications that minimize the effect of data errors are important. This paper explores the use of robust data analysis techniques to reduce the impact of data quality problems on the business decision results.

The evolution of information technology has taken us from the time when computing power was precious to the current state where hardware is the least expensive component of the technology environment. With this evolution in technology we have now reached the point that we can afford to examine large quantities of corporate data -Data Warehouses -to identify potential costs and new revenue opportunities from archived data

Releasing user analysts to explore current data warehouse architectures has not been a viable approach in most current systems. The size and complexity of the data, the technical nature of the access tools and the problems arising from data quality have all combined to keep data warehouse applications in the hands of specialists. The objective of a data warehouse is to provide information on business problems that were not the original focus of the data system that collected the source records. This means that some of the critical data may not be of good quality. As Sprague and Carlson [Sprague and Carlson] pointed out in their classical work on decision support systems, if a managers suspect that bad data is included in the data warehouse they spend much of their time checking data rather than making decisions

Correcting every attribute collected by an organization over several years is usually not feasible. Editing a sample of very bad data, even with modern data editing packages, is often equivalent to reentering all sample data from original sources. Careful data editing is essential. It is also quite expensive. One company [Horrocks, private communication] estimates that each attribute in a Data Warehouse requires 8 to 12 hours of computer analysis to prepare it. While the quality of data in a warehouse is a fundamental goal, there are other approaches that permit us to get quality information out of the data that is available. We intend to discuss some of these and recommend effective practices in this paper.

In the early 1970's statistical researchers including John Tukey and Peter Huber developed a branch of statistics called *robust* statistics (see for example [Hoaglin, Mosteller and Tukey]). The focus of their approach was to develop statistical techniques that were insensitive to the presence of substantial amounts of bad data. Much of this research involved examination of alternate weighting schemes for individual data items that improved robustness and resistance of estimates of central tendency and spread. A number of references also highlighted the potential of well known data analysis tools as preferable alternatives to traditional least squares estimates used by most statistics packages.

Data Quality

High quality data is expensive. Unless an attribute has been used as part of the operational processing of a company, the values are likely to be incomplete or contaminated. Operational data that is used as a part of the day to day processing of a company will probably have been corrected as orders are assembled and shipped. Tangential data that was collected because it might be important at some time is of much more questionable accuracy and completeness. Erroneous entries do not trigger a correction so that the data quality is primarily a function of the interest and care of the collectors. Until the organization is indoctrinated about the importance of this data to the data warehousing results, there are likely to be a substantial minority of bad (sometimes very bad) entries.

Sample vs Population

Correcting invalid data once it has been entered is time consuming and expensive. While packages exist to help edit and consolidate data as a part of the extraction process from the operational database into the data warehouse, they are expensive and incapable of detecting any but the grossest entry errors. Sometimes the only way to assure quality data is to investigate and reenter every data value in a database. While it is usually not feasible to edit and clean an entire warehouse database, it is often possible to do so with a sample. Sample selection should follow some fundamental statistical principles: identification of subject units, estimation of power needs, and focus on observable phenomena.

Robust Tools: Exploration vs Confirmation

Statistical tools are not alike in the way they treat bad data. Common measures such as means and variances are particularly sensitive to a few bad data points. One technique that can be used is to selectively eliminate apparently outlying data. Modern

exploratory statistical packages facilitate this, but this is both time consuming and delicate. A safer approach is to focus on *robust statistical techniques.* Robustness in statistics refers to the ability of a tool to give accurate answers in spite of violations of the assumptions expected for the tool

One way to illustrate the value of robust statistical approaches is to consider the effect of several simple ones on a typical dataset. Figure 1 below illustrates a dataset in which two linear trends are intermixed. Clearly a correct analysis would identify the bulk of the values as the fundamental trend and the higher sloped line as an outlier or exception. Simple linear regression would produce a line that split the difference While a number of tools are common data mining targets - neural networks, cluster analysis, categorical analysis - we will consider a simple regression problem because of its familiarity to most business analysts. In setting up this analysis it is important to remember that OLAP and data mining are used to *discover* business opportunities rather than to *confirm* hypotheses as is common in the academic environment. Resistance to bad data and the ability to locate families of anomalies is an objective of useful exploratory techniques.

Approached Based on L-estimates

An L-estimate is a statistic calculated as a weighted sum of the observation values: $\sum a_i x_i$ [ ]. The most intuitive L-estimate is a trimmed mean. In this case the weight on the smallest and largest p % of the data observations is 0 and the remaining weights are adjusted to add to 1. Trimmed statistics are useful to eliminate values that are orders of magnitude out of line due to units errors. double keying or entry errors. Trimming gives a consistent process to eliminate the bias due to outliers without having to examine each data point individually and introducing analyst bias. The problem is that trimming reduces the sample size. Early statisticians also used a process called Windsorizing in which the 'trimmed' values were replaced by the boundary of the non trimmed scores leaving a sample of the same size. While this process is intuitively attractive, it generates an influence curve that is discontinuous and is not generally used as a robust estimator (see [Hoaglin], et. al. or [Hampel]. The Windsorizing process does generate a continuous influence curve for stochastic processes, however.).

To trim the data in this example we would group the data by x values and delete the largest and smallest p% of the values. Sometimes repeated trimming (5%, 10%, 15%, etc.) produces results that converge to a limiting value and give a good sense of where the outlying data breaks off. Trimming would be particularly useful for those distributions that have a scattering of (real or invalid) large outliers that are disguising the results. Trimming wastes data, however, so another class of estimates should be considered.

Approaches Based on M-estimates

An M-estimate is a statistic calculated buy minimizing some weighting function. The most common M-estimate is a least squares estimate used in most traditional parametric statistics. The problem with the least squares approach is that it is extremely sensitive to outlying values. Mean values, linear regression coefficients and ANOVA statistics are all least squares estimates. Much of the research literature in robust analysis involved determining weighting functions that behave better for data with outliers or non-normal distributions assumptions. Although sophisticated loss functions such as the biweight are studied in the literature, the simplest M-estimates are statistics such as the median that minimize the mean absolute deviation. They are, however, insensitive to bad data and very useful tools for identifying outlying groups of data.

A recommended M-estimate approach for this simple linear regression problem is [Hoaglin, et. al.]

- divide the sample into thirds based on x values.
- calculate the median x and y values in each third: $(x_L , y_L) , (x_M , y_M) , (x_H , y_H)$
- compute the slope as $(y_H - y_L) / (x_H - x_L)$
- compute the intercept as $(1/3)\{(y_L - b\, x_L) + (y_M - b\, x_M) + (y_H - b\, x_H)\}$

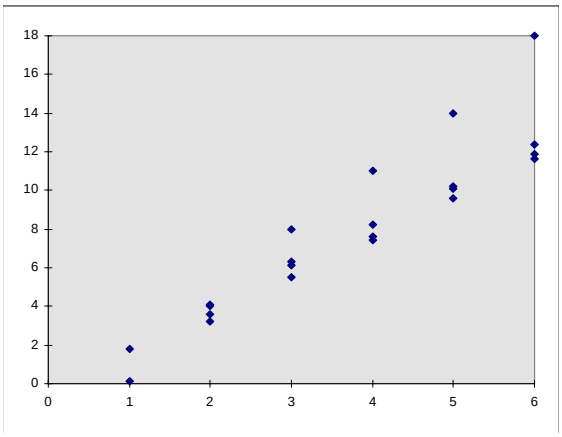
While this seems over simple, it preserves the robustness and breakdown properties of the median.

While medians are robust, they are not particularly powerful. This is not a problem if there is a large quantity of data. In cases where a large number of parameter estimates are necessary or there is not much data, then other techniques have proved useful.

Approaches Based on Reexpressions

Some of the most effective robust analysis tools involve reexpression or transformations of the data. Often the purpose of reexpression is to transform a biased or heavy tailed distribution into one that is more nearly normal. Logarithmic and square root transformations are commonly used to reduce the effect of extreme values on test statistics. But by far the most useful transformation for robust performance is the rank transformation in which the data are replaced by their ranks. Parametric tests such as regression and analysis of variance done on the ranks of the data are easily generated and the statistical parameters (level of significance, power, etc.) are very close to the actual values for the nonparametric tests that these generate [Conover and Iman].

A useful approach to the problem here is to transform both the x values and the y values to ranks and use linear regression to predict the y rank from the x rank [Conover, et. al.]. The predicted rank may be transformed back into the predicted value of y by linear interpolation between the two values of y that have ranks bracketing the predicted rank of y .



References

Conover, W. J. and Ronald L. Iman, (1981) Rank transformations as a bridge between parametric and nonparametric statistics, American Statistician, v35, pp 124-133.

Gendelev, Boris, Closing the OLAP gap, (1998) Database Programming and Design, v.11 no.4, pp 32-39.

Hoaglin, David C., Frederick Mosteller and John W. Tukey, (1983) Understanding Robust and Exploratory Data Analysis, New York, Wiley.

Horrocks, Chris, (1998)Chief Information Officer, Commercial Financial Services, Inc., Tulsa, OK, February 1998.

Huber, Peter J (1981) Robust Statistics, New York, Wiley.

Sprague, R and E. Carlson (1982). Building Effective Decision Support Systems. Englewood Cliffs, NJ: Prentice-Hall, Inc.