ROBUST STATISTICS

Filzmoser P.
Vienna University of Technology, Austria

Rousseeuw P.J.
Universitaire Instelling Antwerpen, Belgium

Keywords: robust estimation, robust regression, breakdown value, multivariate location and scatter, outlier detection.

Contents

Summary

For univariate data it is well known that the sample average can be changed completely by one outlier, whereas the sample median remains useful even when a sizeable fraction of the data is replaced by outliers. The sample average has a breakdown value of zero, whereas the sample median has a positive breakdown value. Also, the least-squares regression method has a breakdown value of zero. In order to attain a positive breakdown value, new regression methods have been developed, such as the least trimmed squares (LTS) method. This approach has had many practical applications.

For multivariate data the estimation of location and scatter can be done by the minimum covariance determinant (MCD) method, which yields high breakdown. This estimator can be used for identifying points with high influence in regression, but also for detecting multivariate outliers. In multivariate analysis one can replace the classical covariance matrix by the MCD estimator, which has successfully been done for example for discriminant analysis, principal components and factor analysis, and canonical correlation analysis.

In robustness, there is currently much activity in generalizing robust methods to other models. Positive-breakdown regression methods such as LTS can be extended to models with several intercepts, to models including dummy regressors, to the zero-intercept regression model, to autoregressive time series, to orthogonal regression, to directional data, and so on. Extensions to nonparametric regression, nonlinear regression, logistic regression, and alternating regression have also been constructed. The latter approach, robust alternating regression, has successfully been used in robustifying factor models and multivariate methods.

1. Motivation and Introduction

The field of robust statistics has gained importance within the last decades. Many researchers are working on robustifying classical statistical methods and on the development of a comprehensive theory of robustness. More and more practitioners are using the advantages offered by robust statistics. Standard statistical software packages include a variety of tools for robust data analysis. Many statisticians have said that statistical data analysis should always consider the aspect of robustness. What is "robustness" and what does "robust statistics" mean?

1.1. The Meaning of Robust Statistics

The classical assumptions of normality, independence, and linearity are often not fulfilled. Statistical estimators and tests which are based on these assumptions will thus give biased results, depending on the "magnitude" of the deviation and on the "sensitivity" of the procedure. To obtain reliable results, a statistical theory is needed that accounts for this kind of deviation from parametric models. *Nonparametric statistics* allows for a whole variety of probability distributions. The restriction to, say, normally distributed data is no longer relevant. However, there are also strong assumptions in nonparametric statistics, like symmetry and absolute continuity. Deviations from these prerequisites again lead to biased and distorted results. Robust statistics works in a "neighborhood" of parametric models. It uses the advantages of parametric models but allows for deviations. Robust statistics can be seen as a theory of approximate parametric models. Hampel et al. gave the definition: "In a broad informal sense, robust statistics is a body of knowledge, partly formalized into 'theories of robustness,' relating to deviations from idealized assumptions in statistics."

1.2. Outliers

The outlier problem is probably as old as statistics. One important task of robust statistics is the identification and proper handling of outliers. Outliers are often thought to be extreme values which are caused by measurement or transmission errors. A definition of the word "outlier" is given in Barnett and Lewis: "We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data." This definition also includes observations which do not follow the majority of the data, such as values that have been measured correctly but are, for one or another reason, far away from the other data values. The cautious formulation "appears to be inconsistent" reflects the subjective

judgment of the observer whether or not an observation is declared to be outlying. One task of robust statistics is to provide methods of detecting outliers.

The detection of outliers can be a very hard problem. Whereas in one dimension observations that are far away from the main data cloud can easily be detected, this is not necessarily the case in higher dimensions, when the outliers are not extreme along the coordinates but in any other direction. With increasing dimensionality, multivariate outliers become harder to detect, yet they can heavily influence the statistical results. Section 5 treats this important problem.

1.3. Aims of Robust Statistics

Classical statistical methods try to fit all data points as well as possible. The usual criterion is least squares, where the sum of the squared residuals has to be minimized to estimate the parameters. If the data set contains outliers, the parameter estimates may deviate strongly from those obtained from the "clean" data. For instance, outliers can attract the regression line. Since all data points obtain the same weight in the least-squares criterion, large deviations are distributed over all the residuals, often making them hard to detect.

One aim of robust statistics is to reduce the impact of outliers. Robust methods try to fit the bulk of the data, which assumes that the good observations outnumber the outliers. Outliers can then be identified by looking at the residuals, which are large in the robust analysis. An important task afterwards is to ask what has caused these outliers. They should not be ignored, but they have to be analyzed and interpreted.

As already mentioned in Section 1.2, robust statistics should ensure reliable results in the case of deviations from idealized assumptions. Apart from outliers, other deviations include unsuspected serial correlations that are due to deviations from the independence assumption. Hence, robust statistics entails much more than just removing some extreme data points. Good robust statistical methods should also prevent efficiency loss, which means loss in precision of the statistical estimation.

1.4. History

Aside from visual inspection of the data, which had already been done in the prehistory of statistics, the beginning of robust statistics dates back to the eighteenth century, when the first rules for the rejection of outliers were developed. These rules were formalized in the nineteenth century, and techniques for robustly estimating "means" were used. Later on, estimators that downweight outliers were developed. Robustness of statistical testing was considered in the first half of the twentieth century. Box (1953) and Tukey (1960) demonstrated the need for robust methods. Their work can be seen as a breakthrough in robust statistics. A few years later, Huber (1964) and Hampel (1974) laid the foundations of a comprehensive theory of robust statistics. Since then the number of papers on robustness has exploded, and the field of robust statistics achieved vast importance. In recent years, previous approaches have been combined, the computational complexity of algorithms has been intensified, and new fields of applications have been opened up.

-
-
-

Bibliography

Barnett V. and Lewis T. (1994). *Outliers in Statistical Data, 3rd edition*, 584 pp. Chichester, UK: Wiley. [Describes many kinds of outliers.]

Box G.E.P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318–335. [The first paper to use the term "robust statistics."]

Hampel F.R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics* **42**, 1887–1896. [Introduces the breakdown value.]

Hampel F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393. [Introduces the influence curve, an important tool in robust statistics.]

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, 502 pp. New York: Wiley. [Gives a survey of robust statistical techniques.]

Huber P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101. [Introduces M-estimators. Started much work on robustness.]

Huber P.J. (1981). *Robust Statistics*, 308 pp. New York: Wiley. [Presents a summary of mathematical concepts of robust statistics.]

Rousseeuw P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880. [Introduces high-breakdown regression estimators.]

Rousseeuw P.J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, Vol. B (ed. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz), pp. 283–297. Dordrecht: Reidel. [Introduces high-breakdown estimators for location and scatter.]

Rousseeuw P.J. and Leroy A.M. (1987). *Robust Regression and Outlier Detection*, 329 pp. New York: Wiley-Interscience. [Provides methods, algorithms, and programs for robust regression, with many examples.]

Tukey J.W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: Essays in Honor of Harald Hotelling* (ed. I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann), pp. 448–485. Stanford: Stanford University Press. [Showed how inefficient classical estimators can be in the presence of outliers, and asked for a systematic study of robustness.]

Biographical Sketches

Peter Filzmoser was born in Wels, Austria in 1968. After his graduation in Mathematics in 1993, he became Assistent at the Department of Statistics, Vienna University of Technology. In 1996 he obtained his Ph.D. degree, and received an award for his thesis on "Principal Planes" from the Vienna University of Technology. In 2001 he became Associate Professor at the Vienna University of Technology. He visited the University of Antwerp (Prof. Rousseeuw) and the Université Libre de Bruxelles (Prof. Croux) several times, resulting in joint papers on robust multivariate methods. In 2001 he organized an international conference on robust statistics in Vorau, Austria.

Peter J. Rousseeuw was born in 1956 in Antwerp (Belgium). After studying mathematics at the University of Brussels, he did his Ph.D. research in statistics at the ETH Zurich. His Ph.D. in 1981 was about robust estimators and tests based on influence functions. Afterwards he carried out research on robust regression analysis and on cluster analysis. From 1984 onwards he was Professor of Statistics, first at the Technical University of Delft (the Netherlands) and then at the University of Fribourg (Switzerland). Since 1989 he has been at the University of Antwerp. He has authored three books (Wiley, New York) and over 140 papers. He is a Fellow of IMS, ISI and ASA. He has acted as adviser for 14 Ph.D. dissertations. Several methods and algorithms that he developed are included in the software packages SAS and S-Plus. His recent research focuses on depth functions and the analysis of economic and financial data.