David J. Olive

Robust Statistics

November 6, 2020

Springer

Preface

Statistics is, or should be, about scientific investigation and how to do it better ....
Box (1990)

Statistics is the science of extracting useful information from data, and a statistical model is used to provide a useful approximation to some of the important characteristics of the population which generated the data.

A *case* or observation consists of the random variables measured for one person or thing. In the location model there is one variable so the ith case is Y_i. For multiple linear regression, the ith case is $(Y_i, \boldsymbol{x}_i^T)^T$ where Y_i is the variable of interest, while for multivariate location and dispersion the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are n cases. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Robust statistics can be tailored to give useful results even when a certain specified model assumption is incorrect. In this text, two assumptions are of great interest: robustness to outliers and robustness to a specified parametric distribution. If a method is robust to outliers, then the method gives useful results even if certain types of outliers are present. If the method is robust to a specified parametric distribution, such as robustness to nonnormality, then there is large sample theory showing that the method is useful on a large class of distributions. For example, central limit type theorems for least squares show that least squares works well for a large class of iid error distributions.

What is in the Book? This online book, a revision of Olive (2008a), finds robust methods that give good results for multiple linear regression or multivariate location and dispersion for a large group of underlying distributions and that are useful for detecting certain types of outliers. Plots for visualizing models and plots for detecting outliers and high leverage cases, and prediction intervals and regions that work for large classes of distributions are also of interest. The emphasis of the text is how to use robust methods in tandem with classical methods for regression, including the special case of the location model. Robust multivariate location and dispersion estimators are derived, and have many applications. A companion volume, Olive (2017b) *Robust Multivariate Analysis*, shows how to use robust methods in tandem with classical methods of multivariate analysis.

Emphasis is on the four following topics. 1) It is shown how to use the response plot to visualize several of the most important regression models including multiple linear regression, binomial regression, Poisson regression, negative binomial regression and their generalized additive model analogs. The response plots are also useful for examining goodness and lack of fit, and for detecting outliers and high leverage groups. 2) The practical robust $\sqrt{n}$ consistent multivariate location and dispersion FCH estimator is developed, along with reweighted versions RFCH and RMVN. These estimators are useful for creating robust multivariate procedures such as robust principal components, for outlier detection and for determining whether the data is from a multivariate normal distribution or some other elliptically contoured distribution. 3) Practical asymptotically optimal prediction intervals and regions are developed. 4) It is shown how to construct the large class of practical $\sqrt{n}$ consistent high breakdown HBREG multiple linear regression estimators.

Chapter 1 is an introduction and Chapter 2 considers the location model with emphasis on the median, the median absolute deviation, the trimmed mean, and the shorth. The dot plot is used to visualize the location model.

Chapter 3 covers the multivariate location and dispersion model, including the multivariate normal and other elliptically contoured distributions. It is also shown that the most used practical “high breakdown” multivariate location and dispersion estimators, such as FMCD (FAST-MCD) and OGK, have not been shown to be consistent or high breakdown. The easily computed outlier resistant $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators are also introduced. These estimators choose between the consistent DGK estimator and the easily computed high breakdown MB estimator. DD plots are used to visualize the model and prediction regions are developed.

Chapters 4-8 consider multiple linear regression. The response plot is used to visualize the model and to detect outliers. The shorth estimator is used to develop prediction intervals that work well for a large class of error distributions. Robust and resistant methods are developed. It is shown that the most used practical “high breakdown” robust regression estimators, such as FLTS (FAST-LTS), have not been shown to be consistent or high breakdown. It is easy to fix the estimators that are not backed by theory, resulting in an easily computed $\sqrt{n}$ consistent high breakdown `hbreg` estimator.

Chapters 9 and 10 show how to visualize many regression models, including generalized linear and generalized additive models, with response plots. These plots are also useful for outlier detection. Chapter 11 provides information on software and suggests some projects for the students.

The text can be used for supplementary reading for courses in regression, multivariate analysis, categorical data analysis, generalized linear models, and exploratory data analysis. The text can also be used to present many statistical methods to students running a statistical consulting lab.

The website (http://parker.ad.siu.edu/Olive/robbook.html) for this book provides more than 30 data sets, and over 115 *R* programs in the file *rpack.txt*. Section 11.2 discusses how to get the data sets and programs into the software, but the following commands will work.

Downloading the book's R functions *rpack.txt* and *R* data sets *robdata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/rpack.txt")
source("http://parker.ad.siu.edu/Olive/robdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 110 *R* functions from *rpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but the functions and data are easily obtained with the source commands).

Background: This course assumes that the student has had considerable exposure to Statistics, but is at a much lower level than most texts on robust statistics. Calculus and a course in linear algebra are essential. Familiarity with least squares regression is also assumed, and the matrix representation of the multiple linear regression model should be familiar. See Olive (2010, 2017a) and Weisberg (2005). An advanced course in statistical inference, especially one that covered convergence in probability and distribution, is needed for several sections of the text. See Casella and Berger (2002), White (1984), and Olive (2008b, 2014).

If most of the large sample theory in the text is covered, then the course should be limited to Ph.D. students who want to do research in high breakdown multivariate robust statistics.

I suggest skipping the theory so that graduate students from many fields can benefit from the course, and I have taught the course three times to undergraduates and graduate students where the prerequisite was a calculus based course in Statistics (e.g. Wackerly, Mendenhall and Scheaffer 2008). For such a course, cover Ch. 1, 2.1–2.5, 3.1, 3.2, 3.3, 3.6, 3.7, 3.10, 3.12, Ch. 4, Ch. 5, 6.2, 7.6, part of 8.2, Ch. 10 and selected topics from Ch. 9. (This will cover the most important material in the text. Many of the remaining sections are for Ph.D. students and experts in robust statistics.) The text problems can be done by graduate and undergraduate students.

The Rousseeuw and Yohai Paradigm: This book is an alternative to the Rousseeuw Yohai paradigm for high breakdown multivariate Robust Statistics which is to approximate an impractical brand name estimator by computing a fixed number of easily computed trial fits and then use the brand name estimator criterion to select the trial fit to be used in the final robust estimator. The resulting estimator will be called an F-brand name estimator or F-estimator where the F indicates that a fixed number of trial fits was used. For example, generate 500 easily computed estimators of multivariate location and dispersion as trial fits. Then choose the trial fit with the dispersion estimator that has the smallest determinant. Since the minimum covariance determinant (MCD) criterion is used, call the resulting estimator the FMCD estimator. These practical estimators are typically not yet backed by large sample or breakdown theory. Most of the literature follows the Rousseeuw Yohai paradigm, using estimators like FMCD, FLTS, FMVE, F-S, FLMS, F-τ, F-Stahel-Donoho, F-Projection, F-MM, FLTA, F-Constrained M, ltsreg, lmsreg, cov.mcd, cov.mve or OGK that are not backed by theory. Maronna, Martin, and Yohai (2006, ch. 2, 6) and Hubert, Rousseeuw, and Van Aelst (2008) provide references for the above estimators.

Problems with these estimators have been pointed out many times. See, for example, Olive (2017b), Huber and Ronchetti (2009, p. xiii, 8-9, 152-154, 196-197) and Hawkins and Olive (2002) with discussion by Hubert, Rousseeuw, and Van Aelst (2002), and Maronna and Yohai (2002). As a rule of thumb, if $p > 2$ then the brand name estimators take too long to compute, so researchers who claim to be using a practical brand name estimator are actually using an F-brand name estimator.

Need for the book: Most of the literature on high breakdown multivariate robust statistics follows the Rousseeuw and Yohai paradigm. See Maronna et al. (2019). The Olive and Hawkins paradigm, as illustrated by this book, is to give theory for the estimator actually used. Practical robust methods backed by theory are needed since so many data sets contain outliers that can ruin a classical analysis. Wilcox (2017) covers material from both paradigms.

This text also simplifies bootstrap theory and theory for variable selection estimators.

Acknowledgments

This work has been partially supported by NSF grants DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware, including R Core Team (2011). A 1997 preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis and visualizing data.

Material from the text has also been used for courses in Regression Graphics, Multiple Linear Regression, Categorical Data, Robust Multivariate Analysis, Robust Statistics, and Statistical Learning.

Contents