

David J. Olive

Robust Statistics

November 6, 2020



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 i th case is Y_i . For multiple linear regression, the i th case is $(Y_i, \mathbf{x}_i^T)^T$ where Y_i is the variable of interest, while for multivariate location and dispersion the i th case is $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^T$. 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

1	Introduction	1
	1.1 Outlier....s	4
	1.2 Applications	7
	1.3 Complements	17
	1.4 Problems	17
2	The Location Model	19
	2.1 Four Essential Statistics	20
	2.2 A Note on Notation	22
	2.3 The Population Median and MAD	23
	2.4 Prediction Intervals and the Shorth	29
	2.5 Bootstrap Confidence Intervals and Tests	34
	2.6 Robust Confidence Intervals	39
	2.7 Large Sample CIs and Tests	41
	2.8 Some Two Stage Trimmed Means	44
	2.9 Asymptotics for Two Stage Trimmed Means	48
	2.10 L, R, and M Estimators	51
	2.11 Asymptotic Theory for the MAD	53
	2.12 Some Other Estimators	57
	2.12.1 The Median of Estimators Estimator	57
	2.12.2 LMS, LTA, LTS	57
	2.13 Asymptotic Variances for Trimmed Means	61
	2.14 Simulation	64
	2.15 Sequential Analysis	71
	2.16 Summary	71
	2.17 Complements	74
	2.18 Problems	77
3	The Multivariate Location and Dispersion Model	85
	3.1 The Multivariate Normal Distribution	87
	3.2 Elliptically Contoured Distributions	90

3.3	The Sample Mean and Sample Covariance Matrix . . .	94
3.4	Mahalanobis Distances	97
3.5	Equivariance and Breakdown	102
3.6	The Concentration Algorithm	105
3.7	Theory for Practical Estimators	110
3.8	DD Plots	121
3.9	Outlier Resistance and Simulations	130
3.10	Outlier Detection if $p > n$	140
3.11	The RMVN Set, RFCH Set, and covmb2 Set	141
3.12	Summary	146
3.13	Complements	150
3.14	Problems	152
4	Prediction Regions and Bootstrap Confidence Regions . . .	165
4.1	Prediction Regions	165
4.2	Bootstrap Confidence Regions	176
4.3	Theory for Bootstrap Confidence Regions	180
4.4	Data Splitting	184
4.5	Summary	185
4.6	Complements	187
4.7	Problems	188
5	Multiple Linear Regression	191
5.1	Predictor Transformations	193
5.2	A Graphical Method for Response Transformations . .	198
5.3	A Review of Multiple Linear Regression	203
5.3.1	The ANOVA F Test	207
5.3.2	The Partial F Test	211
5.3.3	The Wald t Test	214
5.3.4	The OLS Criterion	215
5.4	Asymptotically Optimal Prediction Intervals	217
5.5	Numerical Diagnostics	226
5.6	Graphical Diagnostics	228
5.7	MLR Outlier Detection	230
5.8	MLR Breakdown and Equivariance	236
5.9	MLR Concentration Algorithms	242
5.10	Complements	250
5.11	Problems	252
6	Robust and Resistant Regression	257
6.1	Resistant Multiple Linear Regression	257
6.1.1	The rmreg2 Estimator	261
6.2	A Practical High Breakdown Consistent Estimator . .	263
6.3	High Breakdown Estimators	272
6.3.1	Theoretical Properties	274

6.3.2	Computation and Simulations	278
6.4	Complements	280
6.5	Problems	282
7	MLR Variable Selection and Lasso	285
7.1	Introduction	285
7.2	OLS Variable Selection	287
7.3	Large Sample Theory for Some Variable Selection Estimators	295
7.4	Bootstrapping Variable Selection	300
7.4.1	The Parametric Bootstrap	302
7.4.2	The Residual Bootstrap	303
7.4.3	The Nonparametric Bootstrap	305
7.4.4	Bootstrapping OLS Variable Selection	306
7.4.5	Simulations	310
7.5	Data Splitting	314
7.6	Some Alternative MLR Estimators	314
7.7	Forward Selection	321
7.8	Ridge Regression	324
7.9	Lasso	331
7.10	Lasso Variable Selection	335
7.11	The Elastic Net	337
7.12	Prediction Intervals	342
7.13	Outlier Resistant MLR Methods	346
7.14	Summary	346
7.15	Complements	351
7.16	Problems	353
8	AER and Time Series	363
8.1	Additive Error Regression	363
8.1.1	Response Transformations	365
8.2	Time Series	367
8.3	Summary	367
8.4	Complements	367
8.5	Problems	367
9	1D Regression	371
9.1	Estimating the Sufficient Predictor	374
9.2	Visualizing 1D Regression	379
9.3	Predictor Transformations	387
9.4	Variable Selection	389
9.5	Inference	400
9.6	Complements	409
9.7	Problems	410

10	GLMs and GAMs	417
	10.1 Introduction	417
	10.2 Multiple Linear Regression	419
	10.3 Logistic Regression	423
	10.4 Poisson Regression	434
	10.5 Inference	441
	10.6 Variable Selection	448
	10.7 Generalized Additive Models	455
	10.7.1 Response Plots	459
	10.7.2 The EE Plot for Variable Selection	460
	10.7.3 An EE Plot for Checking the GLM	461
	10.7.4 Examples	461
	10.8 Overdispersion	465
	10.9 Complements	468
	10.10 Problems	471
11	Appendix	477
	11.1 Tips for Doing Research	477
	11.2 R	478
	11.3 Projects	482
	11.4 Some Useful Distributions	483
	11.4.1 The Binomial Distribution	484
	11.4.2 The Burr Type XII Distribution	484
	11.4.3 The Cauchy Distribution	485
	11.4.4 The Chi Distribution	485
	11.4.5 The Chi-square Distribution	485
	11.4.6 The Double Exponential Distribution	486
	11.4.7 The Exponential Distribution	487
	11.4.8 The Two Parameter Exponential Distribution ..	487
	11.4.9 The Gamma Distribution	488
	11.4.10 The Half Cauchy Distribution	489
	11.4.11 The Half Logistic Distribution	490
	11.4.12 The Half Normal Distribution	490
	11.4.13 The Inverse Exponential Distribution	491
	11.4.14 The Largest Extreme Value Distribution	491
	11.4.15 The Logistic Distribution	491
	11.4.16 The Log-Cauchy Distribution	492
	11.4.17 The Log-Logistic Distribution	492
	11.4.18 The Lognormal Distribution	493
	11.4.19 The Maxwell-Boltzmann Distribution	493
	11.4.20 The Normal Distribution	494
	11.4.21 The One Sided Stable Distribution	494
	11.4.22 The Pareto Distribution	495
	11.4.23 The Poisson Distribution	495
	11.4.24 The Power Distribution	496

11.4.25	The Rayleigh Distribution	496
11.4.26	The Smallest Extreme Value Distribution	497
11.4.27	The Student's t Distribution	497
11.4.28	The Topp-Leone Distribution	498
11.4.29	The Truncated Extreme Value Distribution	498
11.4.30	The Uniform Distribution	498
11.4.31	The Weibull Distribution	499
11.5	Truncated Distributions	499
11.5.1	The Truncated Exponential Distribution	502
11.5.2	The Truncated Double Exponential Distribution	504
11.5.3	The Truncated Normal Distribution	504
11.5.4	The Truncated Cauchy Distribution	506
11.6	Large Sample Theory	507
11.6.1	The CLT and the Delta Method	507
11.6.2	Modes of Convergence and Consistency	510
11.6.3	Slutsky's Theorem and Related Results	518
11.6.4	Multivariate Limit Theorems	521
11.7	Mixture Distributions	525
11.8	Complements	527
11.9	Problems	527
11.10	Hints for Selected Problems	533
11.11	Tables	543
	Index	573