DAVID C. HOAGLIN
FREDERICK MOSTELLER
JOHN W. TUKEY

UNDERSTANDING ROBUST AND EXPLORATORY DATA ANALYSIS

WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

Understanding Robust and Exploratory Data Analysis

Edited by

DAVID C. HOAGLIN
Harvard University and Abt Associates Inc.

FREDERICK MOSTELLER
Harvard University

JOHN W. TUKEY
Princeton University and Bell Laboratories

John Wiley & Sons, Inc.
New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:
Main entry under title:

Understanding robust and exploratory data anlaysis.

(Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6356)
Bibliography: p.
Includes index.
1. Mathematical statistics. I. Hoaglin, David Caster, 1944– . II. Mosteller, Frederick, 1916– . III. Tukey, John Wilder, 1915– . IV. Series.

QA276.U5 1982 519.5 82-8528
ISBN 0-471-09777-2 AACR2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

In current statistical practice, both exploratory data analysis and robust and resistant methods have gained important roles. To apply such methods most effectively, the user needs to understand why they are needed and how they work—and can be helped by some insight into how they were devised. This book provides conceptual, logical, and, sometimes, mathematical support for the simpler of these new techniques.

The techniques of exploratory data analysis, particularly as embodied in the book of that title by Tukey (Addison–Wesley, 1977), may seem to have sprung from nowhere and to be supported only by anecdote. The attitudes underlying exploration, though long used by skilled data analysts, have been little exposed to public view. Many of the purposes parallel those of more conventional techniques. Indeed, we can express some of the justifications for particular techniques by using the concepts of classical statistical theory. This book explains and illustrates such connections.

The robust and resistant techniques that we discuss have considerable support in the statistical research literature, both at a highly abstract mathematical level and in extensive Monte Carlo studies. The book provides the basis for an adequate understanding of these techniques using examples and a much reduced level of mathematical sophistication.

By studying this book the user will become more effective in handling robust and exploratory techniques, the student better able to understand them, and the teacher better able to explain them.

Robust and resistant techniques and those of exploratory data analysis have arisen mainly under the guidance of experience, skilled insight, empirical studies of performance, and even analogy with classical techniques. Today these techniques, in part because of this diversity of guidance, do not seem to follow naturally from any unifying structure. Nevertheless, the connections with classical theory do enable us to explain many of the grounds for choosing techniques.

Classical theory emphasizes large-sample notions of consistency, asymptotic variance, and asymptotic relative efficiency. Knowing the behavior of a statistic as the sample becomes large has some utility when, as so frequently happens, large-sample behavior is simpler than small-sample behavior. Above all, we need to recognize that data sets are usually small and that their behavior often lacks the simplicity of large samples. Useful examination of a technique will, therefore, often require new small-sample studies of its performance. Several of these studies, whose results we include, have formed part of the research leading to this volume.

Large-sample considerations can provide a unifying structure for some robust and resistant techniques. The strongest unifying theme underlying exploratory data analysis is expressed in "Look at the data and think about what you are doing." Proceeding largely from these bases, we provide a broad overview of the simpler aspects of data analysis, emphasizing exploratory and robust techniques.

Our intent is that each chapter be reasonably self-contained except for a few generally applicable techniques from the early chapters. Thus, we explain each method in conjunction with an example or two. We continue to use these examples when we describe variations on the technique, explain connections with other techniques (both classical and exploratory), and present results on performance.

Examples are generally small, and almost all rely on real data. They help introduce the reader to techniques and illustrate why and when one method is preferable to another. Through them we give some empirical evidence for the efficacy of each technique in a concrete application.

A brief collection of exercises at the end of each chapter enables the reader to participate more directly in applying the techniques to other sets of data, establishing their properties, or extending them to new situations.

Although our presentation does not generally involve a great deal of mathematics or theoretical argument, we do use them where they seem appropriate. Sometimes, when more formal approaches are unrewarding, we have recourse to numerical simulation. The mathematical arguments we employ are of three types: (a) a proof that a technique meets some desirable objective, (b) an argument that a certain property of a given technique is valuable, and (c) mathematical analogy to extend a technique.

Books in the mathematical sciences may be pitched at various levels and written using different plans. One plan assumes that readers have a particular level of mathematical background and then consistently maintains this level. The advantages are clear—reader and author have a definite contract. Against it, note that some not so well prepared readers who might benefit from parts of the work can be frozen out because they cannot meet the level chosen. Some almost accessible ideas may then have to be held back for

reasons only of level. Another plan uses different levels in different parts of a work, with the intention of keeping as many readers as possible in touch with each part.

This latter plan, which we have adopted, requires tolerance on the part of the readers: The well prepared need to appreciate that they are not being talked down to, whereas the less well prepared must be willing to skip along when too much background is required. Such a plan can produce seemingly inconsistent writing, such as defining a factorial and yet seeming to assume that the same reader finds the gamma function an old friend. The approach we take also means that rigor, when available, may sometimes be deliberately sacrificed in order to communicate the main idea to more readers.

The mathematical prerequisite for the reader is not high for most chapters. At the same time, mathematical sophistication is matched to the requirement for explaining each technique, and so the level rises especially in Chapters 8 and 11. These we have marked by a star (*) before the chapter number in the table of contents. The reader may omit the starred chapters at first reading with no loss of continuity.

Many of the techniques we discuss also appear in *Exploratory Data Analysis* by John W. Tukey and *Data Analysis and Regression* by Frederick Mosteller and John W. Tukey (Addison–Wesley, 1977). In the present book, the emphasis is more on the rationale and development of the methods, and less on illustrating their use. Our exposition is self-contained, but a reader wishing to see more examples and different applications may find profit in referring to one or both of the books just mentioned or to *Applications, Basics, and Computing of Exploratory Data Analysis* by Paul F. Velleman and David C. Hoaglin (Duxbury Press, 1981).

We are preparing a further volume to provide a similar rationale for additional techniques of exploratory data analysis and other robust and resistant methods.

DAVID C. HOAGLIN
FREDERICK MOSTELLER
JOHN W. TUKEY

Saconesset Hills, Massachusetts
July 1982

Acknowledgments

This book grew out of a working group on exploratory data analysis in the Department of Statistics at Harvard University that began in the spring of 1977 and has involved students, faculty, academic visitors, and others. Those who have participated (at one time or another) are Nancy Romanowicz Cook, John D. Emerson, Miriam Gasko, John P. Gilbert (deceased), Katherine Godfrey, Colin Goodall, David C. Hoaglin, Boris Iglewicz, Lois Kellerman, Guoying Li, Lillian Lin, Frederick Mosteller, Anita Parunak, James L. Rosenberger, Andrew Siegel, Keith A. Soper, Michael A. Stoto, Judith Strenio, John W. Tukey, Paul F. Velleman, George Wong, and Cleo Youtz—ten of whom have spent time at Princeton University. Through their sharing of ideas and friendly criticism, all contributed to the development of the material.

In addition to funding from the National Science Foundation (through grants SES 75-15702 and SES 8023644), these activities received partial support from the Middlebury College Faculty Leave Program, the National Institutes of Health (grant CA-23415), the National Cancer Institute (grant T32 CA09337-01), and for John W. Tukey's activities at Princeton, the U.S. Army Research Office (Durham).

Persi Diaconis, Peter J. Huber, David A. Lax, Lincoln E. Moses, Paul F. Velleman, and especially Erich L. Lehmann generously provided comments on various draft chapters.

Katherine Bell Krystinik and Stephan Morgenthaler assembled numerical results at Princeton University related to Chapter 11. Karen Kafadar made available unpublished results from her work on biweight estimators. Virginia C. Klema and Susan Vinal provided information for checking Table 11-3. Jorge Martinez carried out the simulation work for Chapter 12.

Holly Grano, Marjorie Olson, and Cleo Youtz prepared the manuscript with great care and efficiency.

D. C. H.
F. M.
J. W. T.

Contents

Introduction

The classical statistical techniques are designed to be the best possible when stringent assumptions apply. However, experience and further research have forced us to recognize that classical techniques can behave badly when the practical situation departs from the ideal described by such assumptions. The more recently developed robust and exploratory methods are broadening the effectiveness of statistical analyses.

The techniques of exploratory data analysis help us to cope with a set of data in a fairly informal way, guiding us toward structure relatively quickly and easily. Good statistical practitioners have always looked in detail at the data before producing summary statistics and tests of hypotheses. Exploratory data analysis provides us with an extensive repertoire of methods for the detailed study of a set of data. The emphasis is on flexible probing of the data, often before comparing them to any probabilistic model.

Robust and resistant methods, instead of being the best possible in a narrowly defined situation, are "best" compromises for a broad range of situations and, surprisingly often, are close to "best" for each situation alone. Whereas distribution-free methods treat all distributions equally, robust and resistant methods discriminate between those that are more and less plausible.

Broad Phases of Data Analysis

One description of the general steps and operations that make up practical data analysis identifies two broad phases: exploratory and confirmatory. Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst. It often provides the first contact with the data, preceding any firm choice of models for either structural or stochastic components, and it also serves to uncover unexpected departures from familiar models. An important element of the exploratory approach is flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive steps of analysis uncover.

Confirmatory data analysis assesses the reproducibility of the observed patterns or effects. Its role is closer to that of traditional statistical inference in providing statements of significance and confidence; but the confirmatory phase often includes steps such as (a) incorporating information from an analysis of another, closely related body of data and (b) validating a result by collecting and analyzing new data.

In brief, exploratory data analysis emphasizes flexible searching for clues and evidence, whereas confirmatory data analysis stresses evaluating the available evidence.

A cycle of alternating uses of exploratory and confirmatory techniques, either on successive smaller bodies of data or on a single substantial one, is not uncommon and is often very desirable.

Four Themes

Throughout exploratory data analysis, four main themes appear and often combine. These are *resistance*, *residuals*, *re-expression*, and *revelation*.

Resistance provides insensitivity to localized misbehavior in data. A resistant method produces results that change only slightly when a small part of the data is replaced by new numbers, possibly very different from the original ones. Resistant methods pay much attention to the main body of the data and little to outliers. The median is a resistant statistic, whereas the sample mean is not. Attention to resistance reflects the understanding that "good" data seldom contain less than a few percent of gross errors or blunders, so that protection against the adverse effects of such errors should always be available.

In theoretical discussions, one seeks to limit the effect of any "small" change in the sample. In this sense, small changes include minor perturbations in all the data, drastic shifts in a small fraction of the data, and numerous possibilities between these two extremes. In a particular instance, we may need to be concerned with only some of the possible small changes. Thus we might speak of "resistance to wild values" or "resistance to rounding and grouping." Most commonly, because their presence can so easily produce serious distortion, we have wild values in mind when we discuss resistance. Also, it is generally easier to overcome the lesser difficulties that arise when an estimator, such as the sample median, is not resistant to rounding and grouping.

We distinguish between resistance and the related notion of robustness. Robustness generally implies insensitivity to departures from assumptions surrounding an underlying probabilistic model. (Some discussions regard resistance as one aspect of "qualitative robustness.")

In summarizing the location of a sample, the median is highly resistant. A number of exploratory techniques for more structured forms of data

provide resistance because they are based on the median. In terms of efficiency, the median, for all its resistance, is not highly robust because other estimators achieve appreciably greater efficiency across a broader range of distributions. By contrast, the mean is both badly nonresistant and badly nonrobust.

Residuals are what remain after a summary or fitted model has been subtracted out of the data according to the schematic equation

$$\text{residual} = \text{data} - \text{fit}.$$

For example, if the data are the pairs (x_i, y_i) and the fit is the line $\hat{y}_i = a + bx_i$, then the residuals are $r_i = y_i - \hat{y}_i$.

A key attitude of exploratory data analysis asserts that an analysis of a set of data is not complete without a careful examination of the residuals. This analysis can and should take advantage of the tendency of resistant analyses to provide a clear separation between dominant behavior and unusual behavior in the data. When the bulk of the data follows a consistent pattern, that pattern determines a resistant fit. The resistant residuals then contain any drastic departures from the pattern, as well as chance fluctuations. Unusual residuals call for a check on the details of how the corresponding observations were made and handled. As in more traditional practice, the residuals—properly analyzed and displayed—can warn of important systematic aspects of data behavior that may need attention, such as curvature, nonadditivity, and nonconstancy of variability.

Re-expression involves finding what scale (e.g., logarithmic or square root) would simplify the analysis of the data. Exploratory data analysis emphasizes the benefits of considering, at an early stage, whether the original scale of measurement for the data is satisfactory. If not, a re-expression into another scale may help to promote symmetry, constancy of variability, straightness of relationship, or additivity of effect, depending on the structure of the data. A view that the original scale of measurement has a preferred status may cause reluctance to consider re-expression. That view will often not stand examination. True, the physicist sometimes has a cogent theoretical basis for deciding whether to work with volts or $(\text{volts})^2$. However, in circumstances where cogent theory does not guide the choice, the original scale of measurement does not have a similar claim to preferred status. Thus the response of an animal's liver to some treatment may be no more naturally reflected in w, the weight, than in $\log w$ or $\sqrt{w}$, at least until quantitative understanding has advanced.

Revelation through displays meets the analyst's need to see behavior—of data, of fits, of diagnostic measures, and of residuals—and thus to grasp the unexpected features as well as the familiar regularities. Emphasis on visual

displays, including many new graphical techniques, has been a major contribution of exploratory data analysis.

Terminology

Readers who have become acquainted with exploratory data analysis (EDA) after studying traditional statistical methods may wonder why relatively few of the traditional technical terms carry over. Because this book often makes connections between EDA and existing background in statistics, the new words and the familiar ones frequently appear almost side by side. Although such passages may help to clarify how EDA terms relate to traditional ones, they do not always attempt to explain the need for new words. We now offer a few of the more general reasons for the EDA terminology.

First, EDA is more concerned with earlier stages in the overall process of working on data and with different operations and emphases. New techniques often require new technical terms. Stem-and-leaf displays, letter values, and boxplots all illustrate this fact.

Second, some EDA terms are related to, but not equivalent to, more traditional notions. Here the preferred approach is to avoid disturbing the definitions already embedded in the literature. For example, the "hinge" or "fourth" is not exactly a "quartile," and "batch" does not include the assumptions of independence and identical distribution usually associated with "sample."

Third, gains may come from avoiding misleading words. A primary example is "normal," as used in "normal distribution" and in "normal equations" (in least-squares regression), which conflict with this word's usage in normal parlance. Thus we frequently refer to the "Gaussian distribution," rather than the "normal distribution," to avoid any suggestion that this shape customarily underlies actual data.

Finally, some terms, although perhaps unfamiliar, are standard in a particular area of statistics whose results are related to EDA. Some examples in Chapters 9 through 12 come from the field of robustness, where a spurt of research in recent years has established new concepts and many valuable theoretical results.

Estimation

When we engage in a careful discussion of estimation, as in Chapters 9 through 12, we often need to distinguish between the procedure (which we would apply to any sample) and the numerical value (which we obtain by applying the procedure to a particular sample). We use "estimator" for the procedure and "estimate" for the value, and we have tried to maintain this

usage throughout the book. On occasion we use "estimand" for the value that an estimator would produce if it were applied, conceptually, to an entire theoretical distribution. Often the estimand is a familiar parameter of the distribution, but we would also use the term for other quantitative descriptions of what an estimator seems to be estimating.

Sampling Situations

Some of the data sources assembled to challenge proposed robust estimators are technically not distributions. For example, sets of 20 observations in which exactly 19 always come from the standard Gaussian distribution and exactly one always comes from a Gaussian distribution with a larger variance are not samples from either of these two distributions or from a mixture distribution. When we must be careful about this distinction (as in Chapters 10 and 11), we refer to such a data source as a "situation."

Iteration

Resistant and robust techniques more often involve iteration than do classical ones. Thus instead of finding a solution in a single step, we often take an initial value and successively refine it, bringing it closer and closer to the final answer.

In this book the main examples are the three-group resistant line (Chapter 5), median polish (Chapter 6), and the M-estimators of location (Chapter 11). The resistant line procedure defines the slope of the fitted line in terms of the residuals in a way that requires calculating the residuals from a preliminary fit and then adjusting the preliminary slope if necessary. Similarly, in two-way tables, the process of adjusting a preliminary row effect may leave some column medians nonzero. Here, median polish makes alternate adjustments of row and column effects. In doing this, it seeks a two-way table of residuals whose rows and columns all have their medians equal to zero, although we conventionally stop after taking only a few steps toward this goal. In general, the calculations for an M-estimator involve solving, again usually approximately, a nonlinear equation in which the previous estimate enters through the definition of the corresponding residuals.

Often, a certain amount of iteration is to be expected as part of the price of resistance or robustness; the procedures that yield a fit or an estimate without iteration may not be adequately resistant or robust. Happily, the iterative adjustment procedures in this book are simple, and they seldom require many steps.

ADDITIONAL LITERATURE

Besag, J. (1981). "On resistant techniques and statistical analysis," *Biometrika*, **68**, 463–469.

Hampel, F. R. (1971). "A general qualitative definition of robustness," *Annals of Mathematical Statistics*, **42**, 1887–1896.

Tukey, J. W. (1972). "Data analysis, computation, and mathematics," *Quarterly of Applied Mathematics*, **30**, 51–65.

——— (1979). "Robust techniques for the user." In R. L. Launer and G. N. Wilkinson (Eds.), *Robustness in Statistics*. New York: Academic, pp. 103–106.

——— (1980). "We need both exploratory and confirmatory," *The American Statistician*, **34**, 23–25.