MODERN DATA ANALYSIS

Edited by
Robert L. Launer
Andrew F. Siegel

MODERN DATA ANALYSIS

Edited by

ROBERT L. LAUNER

Mathematics Division
U.S. Army Research Office
Research Triangle Park, North Carolina

ANDREW F. SIEGEL

Department of Statistics
Princeton University
Princeton, New Jersey

1982

ACADEMIC PRESS
A Subsidiary of Harcourt Brace Jovanovich, Publishers
New York London
Paris San Diego San Francisco São Paulo Sydney Tokyo Toronto

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

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

Modern data analysis.

Papers presented at a Workshop on Modern Data Analysis, organized by the Mathematics Division of the U.S. Army Research Office, and held in Raleigh, N.C., June 2-4, 1980.

Includes indexes.

Contents: Introduction to styles of data analysis techniques / John W. Tukey -- Some multiple Q-Q plotting procedures / Nicholas P. Jewell -- A reader's guide to smoothing scatterplots and graphical methods for regression / William S. Cleveland -- [etc.]

1. Mathematical statistics--Congresses. I. Launer, Robert L. II. Siegel, Andrew F. III. United States. Army Research Office. Mathematics Division. IV. Workshop on Modern Data Analysis (1980: Raleigh, N.C.)
QA276.A1M63 001.64 82-4037
ISBN 0-12-438180-4 AACR2

PRINTED IN THE UNITED STATES OF AMERICA

82 83 84 85 9 8 7 6 5 4 3 2 1

Modern Data Analysis

Academic Press Rapid Manuscript Reproduction

Proceedings of a Workshop on Modern Data Analysis
sponsored by the United States Army Research Office
held in Raleigh, North Carolina, June 2–4, 1980.

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin.

William S. Cleveland (37), *Bell Laboratories, Murray Hill, New Jersey*

Christopher Cox (45), *Department of Statistics, Division of Biostatistics, University of Rochester, Rochester, New York*

Jerome H. Friedman (123), *Stanford Linear Accelerator Center, Stanford, California*

K. Ruben Gabriel (45), *Department of Statistics and Division of Biostatistics, University of Rochester, Rochester, New York*

Nicholas P. Jewell (13), *Department of Statistics, Princeton University, Princeton, New Jersey*

Joseph W. McKean (171), *Department of Mathematics, Western Michigan University, Kalamazoo, Michigan*

Ronald M. Schrader (171), *Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, New Mexico*

Andrew F. Siegel (103), *Department of Statistics, Princeton University, Princeton, New Jersey*

Werner Stuetzle (123), *Department of Statistics, Stanford University; Stanford Linear Accelerator Center, Stanford, California*

John W. Tukey (1, 83), *Research-Communications Principles Division, Bell Laboratories, Murray Hill, New Jersey; Department of Statistics, Princeton University, Princeton, New Jersey*

Roy E. Welsch (149), *Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts*

PREFACE

A variety of techniques which trace their origins back to many different disciplines are loosely grouped together under the heading "data analysis." What they share is the ability to work well in the task of finding useful structure in complicated collections of recorded information. They are designed to help researchers separate important features from randomness and to draw attention to specific aspects of potential interest that might otherwise be lost in a morass of supportive detail. Although some of these methods were originally created in order to solve specific problems, techniques of very general applicability have resulted. It is certainly to the benefit of scientists and statisticians to be aware of and to share theories and methods of data analysis.

Despite its usefulness, the field of data analysis is only now being accorded the acceptance and recognition it deserves as a serious branch of the science of statistics. This may be due to the necessarily fragmented history of a subject whose original contributions came in part from scientists in scattered fields. Another possible explanation for the only recent emergence of this field might be that some techniques are informal and, although they work well in practice, their precise mathematical properties are not yet known. Thus two important aspects in the study of data analysis are the creation of new methods and the derivation of the properties of existing methods.

The recent increase in awareness and acceptance of the field of data analysis is due in a large part to the efforts of John W. Tukey. Professor Tukey has practiced data analysis for many years and is responsible for the creation of a variety of new and useful methods. His 1977 book "Exploratory Data Analysis" represents a large collection of both the philosophy and the methods of analyzing data, and could be viewed as formally marking the beginning of the movement. His 1977 book with F. Mosteller, "Data Analysis and Regression," extends many of these ideas and provides a link with the more confirmatory side of data analysis. In addition to these books, Professor Tukey has expended a good deal of energy in helping the field by lecturing at scientific meetings, teaching short courses, and directing research in related areas.

How do the methods of data analysis differ from the more classical statistical techniques? A very significant difference is that traditional methods often require that a specific and often restrictive set of assumptions hold. Should the assumptions fail, the conclusions are not guaranteed to be valid, and serious errors can result without warning. In contrast, exploratory and graphic data analytic methods are designed to help the researcher detect many different types of structures. Thus many good data analytic methods are robust and still work well under a variety of underlying models, especially in the presence of outliers and errors in the data. But perhaps the largest difference between classical statistics and modern data analysis is in the philosophy behind the methods. Many classical methods are designed with one model and one question in mind, and the methods are optimized accordingly. Good data analytic methods are designed with the unexpected in mind, so that potentially crucial facets of the data will not be overlooked.

The flow of material in this volume roughly proceeds from general and exploratory to specific and confirmatory. We begin with an introduction to the styles of data analysis, leading into several papers featuring graphical methods. These are followed by a series of contributions relating to the recognition of mathematical form and physical structure. The final papers are closely concerned with the development of formal theories with application to robustness in regression and the linear model.

These papers were presented at the workshop on Modern Data Analysis in June 1980 in Raleigh, North Carolina, organized by the Mathematics Division of the U.S. Army Research Office.

ABSTRACTS

INTRODUCTION TO STYLES OF DATA ANALYSIS TECHNIQUES

John W. Tukey

We are not used to thinking about data analysis techniques in terms of style. We are not familiar with a good supply of names or acronyms for either the broad purposes of the techniques or the other important "coordinates" in whose terms such techniques can be usefully described. As a result both writer and reader have an unusually difficult task. The development sequence begins with a sketch of three pairs of coordinates. The first pair, "stochastic background" and "stringency," seem to deserve treatment together, many instances falling under one of eight rubrics. Another pair "character" and "flexibility" also go together with six combinations worth emphasis. The combination (interactive) of these two pairs of coordinates is then described. Next we notice that "data structure" is wisely interpreted as covering more than the externals of the data, going on to a brief historical setting for modern robust/resistant techniques.

SOME MULTIPLE Q-Q PLOTTING PROCEDURES

Nicholas P. Jewell

This paper is concerned with some extensions to the idea of a quantile-quantile (Q-Q) plot that is commonly used by statisticians for a variety of purposes. Both single and multiple Q-Q plots are considered. Particular attention is paid to problems involving extreme-value data and to the study of the behavior of sample averages.

A READER'S GUIDE TO SMOOTHING SCATTERPLOTS AND GRAPHICAL METHODS FOR REGRESSION

William S. Cleveland

Comments about smoothing scatterplots and graphical methods for regression are made and pointers to literature relevant to these comments are given.

SOME COMPARISONS OF BIPLOT DISPLAY AND PENCIL-AND-PAPER EXPLORATORY DATA ANALYSIS METHODS

Christopher Cox and K. Ruben Gabriel

This paper uses a number of data sets from Tukey's "Exploratory Data Analysis" to compare their inspection and analysis by biplot display with the exploratory data analysis given by Tukey. The use of biplots for display of two and three way tables is described and the methods of diagnosing models are explained. The illustrations suggest that biplot diagnoses usually result in similar models (additive, multiplicative, degree-of-freedom-for-non-additivity, etc.) to those brought out by pencil-and-paper exploratory data analysis techniques—but biplot diagnostics seem much faster and more immediate.

THE USE OF SMELTING IN GUIDING RE-EXPRESSION

John W. Tukey

Most frameworks, whether or not statistical models, used in data analysis involve some type of functional behavior. Thus it is important to have effective techniques of asking the data what sort of functional behavior we should have in our framework for handling a specific body (or kind) of data. Some sort of smoothing process is essential here; one that is robust/resistant and provides specially smooth input to a recognizer of functional form.

Smoothing is usually thought of as value change, but it can also be done by eliminating less typical points and keeping more typical ones. *Smelting* is a specific class of techniques for smoothing by excision, in which the qualitative nature of the series is used to tell us which (x, y) pairs to keep and which to set aside.

Combined with a good selection of diagnostic plots, smelting offers the best route we have today toward functional form recognition. Done reasonably, we can go far toward the use of functional forms invertible in closed form, avoiding, for example, dangerous polynomials.

GEOMETRIC DATA ANALYSIS: AN INTERACTIVE GRAPHICS PROGRAM FOR SHAPE COMPARISON

Andrew F. Siegel

Two shapes, each consisting of n homologous points, can be rotated, scaled, and translated to obtain a close fit to each other by several methods. An interactive graphical computer program is presented here that implements two methods: least squares and repeated medians, a robust method. Examples are given and the use of the system is discussed.

PROJECTION PURSUIT METHODS FOR DATA ANALYSIS

Jerome H. Friedman and Werner Stuetzle

Projection pursuit methods iteratively construct a model for structure in multivariate data, based on suitably chosen lower dimensional projections. At each step of the iteration, the model is updated to agree with the data in the corresponding projection. Projections can be chosen either by numerical optimization (automatic projection pursuit) or interactively by a user at a computer graphics terminal (manual projection pursuit). The projection pursuit paradigm has been applied to clustering, regression, classification, and density estimation.

INFLUENCE FUNCTIONS AND REGRESSION DIAGNOSTICS

Roy E. Welsch

Influential-data diagnostics are becoming an accepted part of data analysis. In this paper we show how these diagnostic techniques are connected with the ideas of qualitative robustness described by Hampel and the concept of bounded-influence regression as developed by Krasker and Welsch. Asymptotic influence functions are discussed, and the identification of influential subsets of data points is considered.

THE USE AND INTERPRETATION OF ROBUST ANALYSIS OF VARIANCE

Joseph W. McKean and Ronald M. Schrader

Robust analysis of variance procedures are discussed for the general non-full-rank linear model. These procedures are described in terms of their mathematical structure, demonstrating that the analysis has the same uses and interpretation as classical analysis of variance. This structure also leads to efficient computational algorithms. Necessary standardizing constants for the test statistics are motivated by consideration of likelihood ratio tests. An example of an experimental design illustrates the similarity between the robust and classical analyses, emphasizing the advantages of the robust method. Some Monte Carlo results attest to the validity of the robust methods for the example.

CONTENTS

INTRODUCTION TO STYLES OF DATA ANALYSIS TECHNIQUES

John W. Tukey

Research-Communications Principles Division
Bell Laboratories
Murray Hill, New Jersey
and
Department of Statistics
Princeton University*
Princeton, New Jersey

We are not used to thinking about data analysis techniques in terms of style. We are not familiar with a good supply of names or acronyms for either the broad purposes of the techniques or the other important "coordinates" in whose terms such techniques can be usefully described. As a result both writer and reader have an unusually difficult task.

One way to, I hope, ease that task is to offer readers two different ways to read what follows: *either* Sections I to VI followed by Sections VII and VIII, thus developing concepts before their exemplification, *or* first Sections VII and VIII, followed by Sections I to VI, thus illustrating the concepts before defining them.

The development sequence begins [Section I] with a sketch of three pairs of coordinates. The first pair, "stochastic background" and "stringency", seem to deserve treatment together [Section II], many instances falling under one of 8 rubrics. Another pair "character" and "flexibility" also go together [Section III] with 6 combinations worth emphasis. Section IV then describes the combination (interactive) of these two pairs of coordinates. Next we notice [Section V] that "data structure" is wisely interpreted as covering more than the externals of the data, going on [Section VI] to a brief historical setting for modern robust/resistant techniques.

The illustration sequence begins [Section VII] with a brief account of the more important classes of data-handling components -- ADEs, OCONs, DDAPs, OUTs, CDAPs and SDAPs. Section VIII discusses how more or less familiar techniques, when used to analyze 10 observations on each of 7 quantities, fit into the classification set up in Section VI and, in part, into the coordinates described earlier.

*Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham).

ISBN 0-12-438180-4

I. COORDINATES FOR USES OF DATs

If we are to think about uses of data analysis techniques (DATs), we need to have several kinds of coordinates in mind. These are conveniently grouped as follows:

styles
- *stochastic background*
- *indication, conclusion, etc.*

data structure
- *formal arrangement*
- *type of phenomena*

specifiers
- *output wanted*
- *algorithms used*

Any or all may be important. Of the six, the 3rd, 5th and 6th are, relatively at least, well understood. What we might need to discuss then, will be the 1st, 2nd, and, in less detail, the 4th.

Some readers may want to begin with Section VII, reading to the end before returning to this point. Others will prefer to read the sections in order of their numbers.

II. STOCHASTIC BACKGROUND AND STRINGENCY COMBINED: THE FIRST STYLE COORDINATE

It would have been possible for techniques to be common with any combination of stochastic background and stringency.

Here "stringency" is a deliberately vague term (cp. Mosteller and Tukey 1977, pp. 17ff) covering "efficiency", "power", "minimum variance" and the like. In our current world, however, only a few combinations are at all common, namely the eight in exhibit 1.

Two comments need to be made about this coordinate:

- where a DAT (data analysis technique) belongs may depend on the circumstances where it was invented -- and it may depend on how it is thought about. Thus moment-matching requires little if any formal background for its invention (and development), but *is* sometimes of high narrow stringency against an overutopian background.
- one reason for the absence of "nonparametric, high" is that we have not found any good way to seek out such a behavior.

exhibit 1

The 8 common combinations of stochastic background and stringency

Stochastic background	Narrow stringency*	Broad stringency*	Examples** (say for $n \geq 8$)
No formal	(inapplicable)	(inapplicable)	midhinge
Overutopian	Not much	Dubious	s^2 and mean of random subsample of 3
Overutopian	High or nearly so	Dubious	mean and s^2
"Nonparametric"	Unknown	Unknown	median and sign test for a random subsample
"Nonparametric"	Some	Dubious	midhinge and hingespread
"Nonparametric"	Some	Some	—
Robust/resistant	Medium	Medium	midmean and midspread
Robust/resistant	High	High	biweight

*"Stringency" here means "degree of success in wringing out all the information that is there". It is narrow if assessed for a narrowly specified situation, as for instance, for samples from a Gaussian distribution. It is broad if assessed against each and any of a broad set of situations.

**Most unfamiliar terms are defined in either Tukey 1977, Mosteller and Tukey 1977, or both.

Illustrations

A few brief illustrations may help us. Let us consider:

- Using Student's t is a matter of *critical* data analysis. This is usually thought of as deriving from a tight Gaussian specification via sufficient statistics (which would make its stochastic background "overutopian", its narrow stringency "high", and its broad stringency "dubious"), but, since the paper of Pitman (1937), using Student's t can also be thought of -- in the two-sample case, at least -- as almost nonparametric in stochastic background but still, of course, with broad stringency, dubious.
- Looking hard at a sample median (unaccompanied); this has to be a matter of *exploratory* data analysis, probably with no formal stochastic background for small batches, however, medians also come from a robust/resistant stochastic background; the broad stringency would vary with batch size (#=3 or 4, evermore relatively high stringency; #=5 or 6, today relatively high stringency. $n = 7$ up, moderate stringency.)
- Looking hard at a modern (say 6-biweight) estimate of center unaccompanied by assessment of width; this still has also to be a matter of *exploratory* data analysis, but the stochastic background is almost surely robust/resistant, and the stringency is high (at least for $n \geq 7$ or 8).

III. CHARACTER AND FLEXIBILITY COMBINED: THE SECOND STYLE COORDINATE

Here we find character and flexibility even more closely related to one another (than stochastic background and stringency were). The six alternatives of exhibit 2 cover all that is today common.

exhibit 2

6 common combinations of character and flexibility

Tag	Character	Flexibility
RDDA	rigidly descriptive (with no exploration)	negligible
EDA	truly exploratory	large
OCDA→E	"Overlapping-critical" used for exploratory purposes	yes
OCDA→C	"Overlapping-critical" used for confirmatory purposes	yes
SCDA	"simple confirmatory" "separate critical"	not here
CCDA	"careful confirmatory"	eliminated