PETER J. ROUSSEEUW
ANNICK M. LEROY

ROBUST REGRESSION AND OUTLIER DETECTION

WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

Theoretical and
Nonparametric
cond Edition
tics
id Mathematical
and Statistical
tatistics
Applications to
nd Martingales,
ecasting
tistics
and
n Statistics
cations to the

Theoretical and

Nonparametric

cond Edition

tics

d Mathematical

and Statistical

tatistics

Applications to

and Martingales,

ecasting

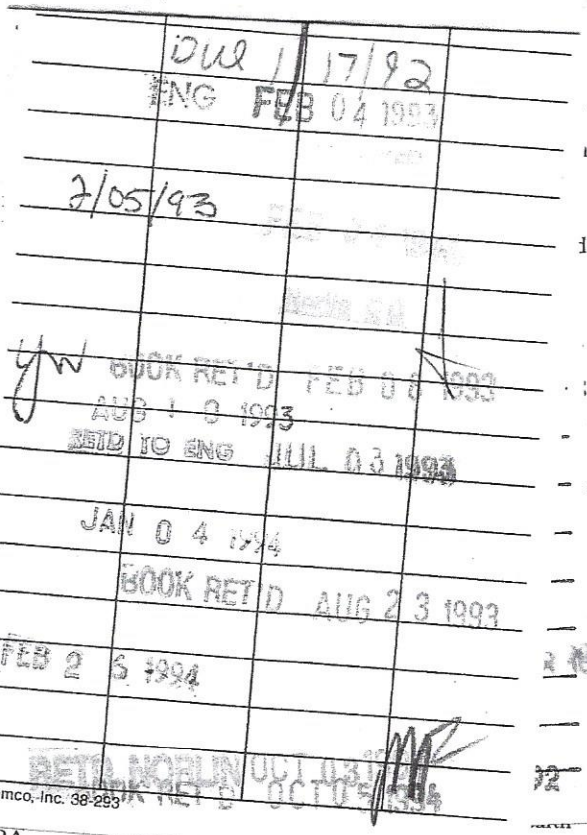
tistics

and

n Statistics

ications to the

BARNETT and LEWIS • Outliers in Statistical Data, *Second Edition*
BARTHOLOMEW • Stochastic Models for Social Processes, *Third Edition*
BARTHOLOMEW and FORBES • Statistical Techniques for Manpower Planning
BECK and ARNOLD • Parameter Estimation in Engineering and Science
BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BHAT • Elements of Applied Stochastic Processes, *Second Edition*
BLOOMFIELD • Fourier Analysis of Time Series: An Introduction
BOX • R. A. Fisher, The Life of a Scientist
BOX and DRAPER • Empirical Model-Building and Response Surfaces
BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER • Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BROWN and HOLLANDER • Statistics: A Biomedical Introduction
BUNKE and BUNKE • Statistical Inference in Linear Models, Volume I
CHAMBERS • Computational Methods for Data Analysis
CHATTERJEE and PRICE • Regression Analysis by E
CHOW • Econometric Analysis by Control Methods
CLARKE and DISNEY • Probability and Random Course with Applications, *Second Edition*
COCHRAN • Sampling Techniques, *Third Edition*
COCHRAN and COX • Experimental Designs, *Secon*
CONOVER • Practical Nonparametric Statistics, *Second Edition*



BARNETT and LEWIS • Outliers in Statistical Data, *Second Edition*
BARTHOLOMEW • Stochastic Models for Social Processes, *Third Edition*
BARTHOLOMEW and FORBES • Statistical Techniques for Manpower Planning
BECK and ARNOLD • Parameter Estimation in Engineering and Science
BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BHAT • Elements of Applied Stochastic Processes, *Second Edition*
BLOOMFIELD • Fourier Analysis of Time Series: An Introduction
BOX • R. A. Fisher, The Life of a Scientist
BOX and DRAPER • Empirical Model-Building and Response Surfaces
BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER • Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BROWN and HOLLANDER • Statistics: A Biomedical Introduction
BUNKE and BUNKE • Statistical Inference in Linear Models, Volume I
CHAMBERS • Computational Methods for Data Analysis
CHATTERJEE and PRICE • Regression Analysis by F
CHOW • Econometric Analysis by Control Methods
CLARKE and DISNEY • Probability and Random Course with Applications, *Second Edition*
COCHRAN • Sampling Techniques, *Third Edition*
COCHRAN and COX • Experimental Designs, *Secon*
CONOVER • Practical Nonparametric Statistics, *Secon*

U18300 8530520

Pr ... in Theoretical and

... of Nonparametric

Second Edition

... istics

... nd Mathematical

... d

... and Statistical

... Statistics

... Applications to

Appl.

... and Martingales,

... recasting

... atistics
... , and
... s

... in Statistics

... lications to the

BARNETT • Interpreting Multivariate Data
BARNETT and LEWIS • Outliers in Statistical Data, *Second Edition*
BARTHOLOMEW • Stochastic Models for Social Processes, *Third Edition*
BARTHOLOMEW and FORBES • Statistical Techniques for Manpower Planning
BECK and ARNOLD • Parameter Estimation in Engineering and Science
BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BHAT • Elements of Applied Stochastic Processes, *Second Edition*
BLOOMFIELD • Fourier Analysis of Time Series: An Introduction
BOX • R. A. Fisher, The Life of a Scientist
BOX and DRAPER • Empirical Model-Building and Response Surfaces
BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER • Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BROWN and HOLLANDER • Statistics: A Biomedical Introduction
BUNKE and BUNKE • Statistical Inference in Linear Models, Volume I
CHAMBERS • Computational Methods for Data Analysis
CHATTERJEE and PRICE • Regression Analysis by F...
CHOW • Econometric Analysis by Control Methods
CLARKE and DISNEY • Probability and Random ... Course with Applications, *Second Edition*
COCHRAN • Sampling Techniques, *Third Edition*
COCHRAN and COX • Experimental Designs, *Second ...*
CONOVER • Practical Nonparametric Statistics, *Second Edition*

Robust Regression and Outlier Detection

PETER J. ROUSSEEUW
University of Fribourg, Switzerland

ANNICK M. LEROY
Vrije Universiteit Brussel, Belgium

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:

Rousseeuw, Peter J.
Robust regression and outlier detection.

(Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6356)
Bibliography: p.
Includes index.
1. Regression analysis. 2. Outliers (Statistics) 3. Least squares. I. Leroy, Annick M. II. Title. III. Series.
QA278.2.R68 1987 519.5'36 87-8234
ISBN 0-471-85233-3

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

Regression analysis is an important statistical tool that is routinely applied in most sciences. Out of many possible regression techniques, the least squares (LS) method has been generally adopted because of tradition and ease of computation. However, there is presently a widespread awareness of the dangers posed by the occurrence of outliers, which may be a result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample. Outliers occur very frequently in real data, and they often go unnoticed because nowdays much data is processed by computers, without careful inspection or screening. Not only the response variable can be outlying, but also the explanatory part, leading to so-called *leverage points*. Both types of outliers may totally spoil an ordinary LS analysis. Often, such influential points remain hidden to the user, because they do not always show up in the usual LS residual plots.

To remedy this problem, new statistical techniques have been developed that are not so easily affected by outliers. These are the robust (or resistant) methods, the results of which remain trustworthy even if a certain amount of data is contaminated. Some people think that robust regression techniques hide the outliers, but the opposite is true because the outliers are far away from the robust fit and hence can be detected by their large residuals from it, whereas the standardized residuals from ordinary LS may not expose outliers at all. The main message of this book is that robust regression is extremely useful in identifying outliers, and many examples are given where all the outliers are detected in a single blow by simply running a robust estimator.

An alternative approach to dealing with outliers in regression analysis is to construct outlier diagnostics. These are quantities computed from

the data with the purpose of pinpointing influential observations, which can then be studied and corrected or deleted, followed by an LS analysis on the remaining cases. Diagnostics and robust regression have the same goal, but they proceed in the opposite order: In a diagnostic setting, one first wants to identify the outliers and then fit the good data in the classical way, whereas the robust approach first fits a regression that does justice to the majority of the data and then discovers the outliers as those points having large residuals from the robust equation. In some applications, both approaches yield exactly the same result, and then the difference is mostly subjective. Indeed, some people feel happy when switching to a more robust criterion, but they cannot accept the deletion of "true" observations (although many robust methods will, in effect, give the outliers zero influence), whereas others feel that it is all right to delete outliers, but they maintain that robust regression is "arbitrary" (although the combination of deleting outliers and then applying LS is itself a robust method). We are not sure whether this philosophical debate serves a useful purpose. Fortunately, some positive interaction between followers of both schools is emerging, and we hope that the gap will close. Personally we do not take an "ideological" stand, but we propose to judge each particular technique on the basis of its reliability by counting how many outliers it can deal with. For instance, we note that certain robust methods can withstand leverage points, whereas others cannot, and that some diagnostics allow us to detect multiple outliers, whereas others are easily masked.

In this book we consider methods with high breakdown point, which are able to cope with a large fraction of outliers. The "high breakdown" objective could be considered a kind of third generation in robustness theory, coming after minimax variance (Huber 1964) and the influence function (Hampel 1974). Naturally, the emphasis is on the methods we have worked on ourselves, although many other estimators are also discussed. We advocate the least median of squares method (Rousseeuw 1984) because it appeals to the intuition and is easy to use. No background knowledge or choice of tuning constants are needed: You just enter the data and interpret the results. It is hoped that robust methods of this type will be incorporated into major statistical packages, which would make them easily accessible. As long as this is not yet the case, you may contact the first author (PJR) to obtain an updated version of the program PROGRESS (Program for RObust reGRESSion) used in this book. PROGRESS runs on IBM-PC and compatible machines, but the structured source code is also available to enable you to include it in your own software (Recently, it was integrated in the ROBETH library in Lausanne and in the workstation package S-PLUS of Statistical Sciences,

Inc., P.O. Box 85625, Seattle WA 98145-1625.) The computation time is substantially higher than that of ordinary LS, but this is compensated by a much more important gain of the statistician's time, because he or she receives the outliers on a "silver platter." And anyway, the computation time is no greater than that of other multivariate techniques that are commonly used, such as cluster analysis or multidimensional scaling.

The primary aim of our work is to make robust regression available for everyday statistical practice. The book has been written from an applied perspective, and the technical material is concentrated in a few sections (marked with *), which may be skipped without loss of understanding. No specific prerequisites are assumed. The material has been organized for use as a textbook and has been tried out as such. Chapter 1 introduces outliers and robustness in regression. Chapter 2 is confined to simple regression for didactic reasons and to make it possible to include robustness considerations in an introductory statistics course not going beyond the simple regression model. Chapter 3 deals with robust multiple regression, Chapter 4 covers the special case of one-dimensional location, and Chapter 5 discusses the algorithms used. Outlier diagnostics are described in Chapter 6, and Chapter 7 is about robustness in related fields such as time series analysis and the estimation of multivariate location and covariance matrices. Chapters 1–3 and 6 could be easily incorporated in a modern course on applied regression, together with any other sections one would like to cover. It is also quite feasible to use parts of the book in courses on multivariate data analysis or time series. Every chapter contains exercises, ranging from simple questions to small data sets with clues to their analysis.

PETER J. ROUSSEEUW
ANNICK M. LEROY

October 1986

Contents

CHAPTER 1

Introduction

1. OUTLIERS IN REGRESSION ANALYSIS

The purpose of regression analysis is to fit equations to observed variables. The classical linear model assumes a relation of the type

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + e_i \quad \text{for } i = 1, \ldots, n\,, \tag{1.1}$$

where n is the sample size (number of cases). The variables $x_{i1}, \ldots, x_{ip}$ are called the *explanatory variables* or *carriers*, whereas the variable y_i is called the *response variable*. In classical theory, the *error term* e_i is assumed to be normally distributed with mean zero and unknown standard deviation σ. One then tries to estimate the vector of unknown parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \tag{1.2}$$

from the data:

$$\begin{array}{c} \\ \text{Cases} \end{array} \begin{array}{c} \text{Variables} \\ \begin{bmatrix} x_{11} & \cdots & x_{1p} & y_1 \\ \vdots & & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} & y_i \\ \vdots & & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & y_n \end{bmatrix} \end{array}. \tag{1.3}$$

Applying a regression estimator to such a data set yields

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{bmatrix}, \tag{1.4}$$

where the estimates $\hat{\theta}_j$ are called the *regression coefficients*. (Vectors and matrices will be denoted by boldface throughout.) Although the actual θ_j are unknown, one can multiply the explanatory variables with these $\hat{\theta}_j$ and obtain

$$\hat{y}_i = x_{i1}\hat{\theta}_1 + \cdots + x_{ip}\hat{\theta}_p , \tag{1.5}$$

where $\hat{y}_i$ is called the *predicted* or *estimated* value of y_i. The *residual* r_i of the ith case is the difference between what is actually observed and what is estimated:

$$r_i = y_i - \hat{y}_i . \tag{1.6}$$

The most popular regression estimator dates back to Gauss and Legendre (see Plackett 1972 and Stigler 1981 for some historical discussions) and corresponds to

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2 . \tag{1.7}$$

The basic idea was to optimize the fit by making the residuals very small, which is accomplished by (1.7). This is the well-known *least squares* (LS) method, which has become the cornerstone of classical statistics. The reasons for its popularity are easy to understand: At the time of its invention (around 1800) there were no computers, and the fact that the LS estimator could be computed *explicitly* from the data (by means of some matrix algebra) made it the only feasible approach. Even now, most statistical packages still use the same technique because of tradition and computation speed. Also, in the one-dimensional situation the LS criterion (1.7) yields the arithmetic mean of the observations, which at that time seemed to be the most reasonable location estimator. Afterwards, Gauss introduced the normal (or Gaussian) distribution as the error distribution for which LS is optimal (see the citations in Huber 1972, p. 1042, and Le Cam 1986, p. 79), yielding a beautiful mathematical theory. Since then, the combination of Gaussian assumptions and LS has become a standard mechanism for the generation of statistical techniques

(e.g., multivariate location, analysis of variance, and minimum variance clustering).

More recently, some people began to realize that real data usually do not completely satisfy the classical assumptions, often with dramatic effects on the quality of the statistical analysis (see, e.g., Student 1927, Pearson 1931, Box 1953, and Tukey 1960).

As an illustration, let us look at the effect of outliers in the simple regression model

$$y_i = \theta_1 x_i + \theta_2 + e_i \tag{1.8}$$

in which the slope θ_1 and the intercept θ_2 are to be estimated. This is indeed a special case of (1.1) with $p = 2$ because one can put $x_{i1} := x_i$ and $x_{i2} := 1$ for all $i = 1, \ldots, n$. (In general, taking a carrier identical to 1 is a standard trick used to obtain regression with a constant term.) In the simple regression model, one can make a plot of the (x_i, y_i), which is sometimes called a *scatterplot*, in order to visualize the data structure. In the general multiple regression model (1.1) with large p, this would no longer be possible, so it is better to use simple regression for illustrative purposes.

Figure 1*a* is the scatterplot of five points, $(x_1, y_1), \ldots, (x_5, y_5)$, which almost lie on a straight line. Therefore, the LS solution fits the data very well, as can be seen from the LS line $\hat{y} = \hat{\theta}_1 x + \hat{\theta}_2$ in the plot. However, suppose that someone gets a wrong value of y_4 because of a copying or transmission error, thus affecting, for instance, the place of the decimal point. Then (x_4, y_4) may be rather far away from the "ideal" line. Figure 1*b* displays such a situation, where the fourth point has moved up and away from its original position (indicated by the dashed circle). This point is called an *outlier in the y-direction*, and it has a rather large influence on the LS line, which is quite different from the LS line in Figure 1*a*. This phenomenon has received some attention in the literature because one usually considers the y_i as observations and the $x_{i1}, \ldots, x_{ip}$ as fixed numbers (which is only true when the design has been given in advance) and because such "vertical" outliers often possess large positive or large negative residuals. Indeed, in this example the fourth point lies farthest away from the straight line, so its r_i given by (1.6) is suspiciously large. Even in general multiple regression (1.1) with large p, where one cannot visualize the data, such outliers can often be discovered from the list of residuals or from so-called *residual plots* (to be discussed in Section 4 of Chapter 2 and Section 1 of Chapter 3).

However, usually also the explanatory variables $x_{i1}, \ldots, x_{ip}$ are observed quantities subject to random variability. (Indeed, in many applications, one receives a list of variables from which one then has to choose a

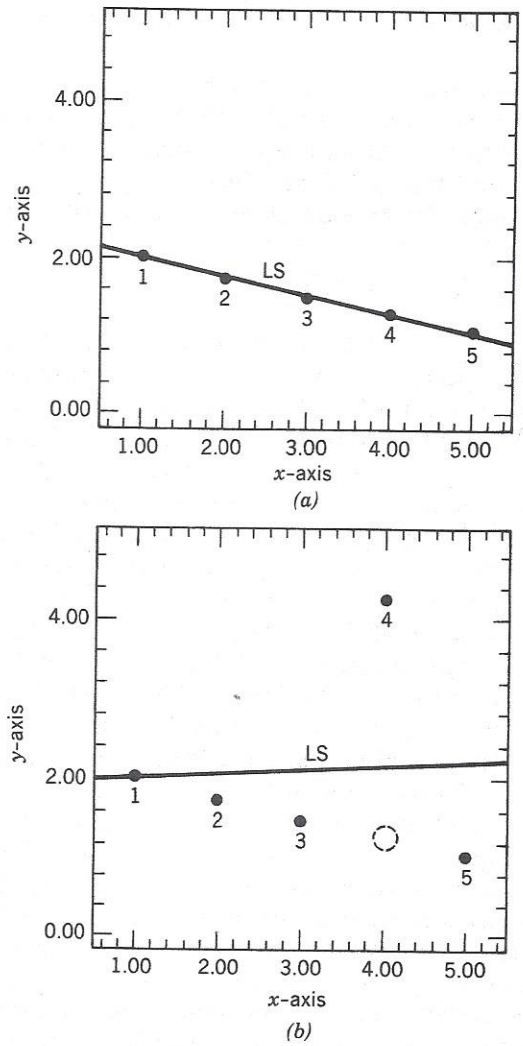


Figure 1. (*a*) Original data with five points and their least squares regression line. (*b*) Same data as in part (*a*), but with one outlier in the *y*-direction.

response variable and some explanatory variables.) Therefore, there is no reason why gross errors would only occur in the response variable y_i. In a certain sense it is even more likely to have an outlier in one of the explanatory variables $x_{i1}, \ldots, x_{ip}$ because usually p is greater than 1, and hence there are more opportunities for something to go wrong. For the effect of such an outlier, let us look at an example of simple regression in Figure 2.

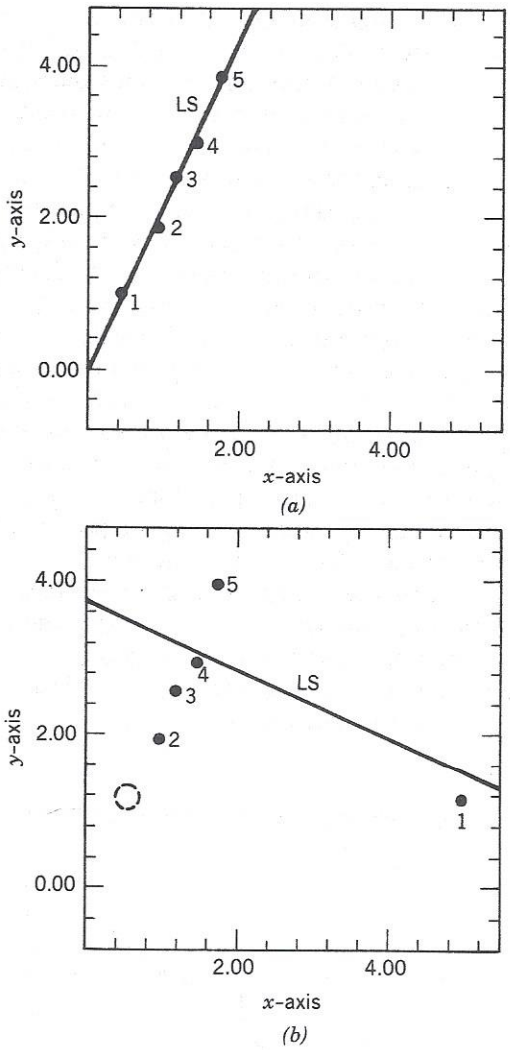


Figure 2. (*a*) Original data with five points and their least squares regression line. (*b*) Same data as in part (*a*), but with one outlier in the *x*-direction ("leverage point").

Figure 2*a* contains five points, $(x_1, y_1), \ldots, (x_5, y_5)$, with a well-fitting LS line. If we now make an error in recording x_1, we obtain Figure 2*b*. The resulting point is called an *outlier in the x-direction*, and its effect on the least squares estimator is very large because it actually tilts the LS line. Therefore the point (x_1, y_1) in Figure 2*b* is called a *leverage point*, in analogy to the notion of leverage in mechanics. This large "pull" on the

LS estimator can be explained as follows. Because x_1 lies far away, the residual r_1 from the original line (as shown in Figure 2*a*) becomes a very large (negative) value, contributing an enormous amount to $\Sigma_{i=1}^{5} r_i^2$ for that line. Therefore the original line cannot be selected from a least squares perspective, and indeed the line of Figure 2*b* possesses the smallest $\Sigma_{i=1}^{5} r_i^2$ because it has tilted to reduce that large r_1^2, even if the other four terms, $r_2^2, \ldots, r_5^2$, have increased somewhat.

In general, we call an observation (x_k, y_k) a leverage point whenever x_k lies far away from the bulk of the observed x_i in the sample. Note that this does not take y_k into account, so the point (x_k, y_k) does not necessarily have to be a regression outlier. When (x_k, y_k) lies close to the regression line determined by the majority of the data, then it can be considered a "good" leverage point, as in Figure 3. Therefore, to say that (x_k, y_k) is a leverage point refers only to its *potential* for strongly affecting the regression coefficients $\hat{\theta}_1$ and $\hat{\theta}_2$ (due to its outlying component x_k), but it does not necessarily mean that (x_k, y_k) will actually have a large influence on $\hat{\theta}_1$ and $\hat{\theta}_2$, because it may be perfectly in line with the trend set by the other data. (In such a situation, a leverage point is even quite beneficial because it will shrink certain confidence regions.)

In multiple regression, the $(x_{i1}, \ldots, x_{ip})$ lie in a space with p dimensions (which is sometimes called the *factor space*). A leverage point is then still defined as a point $(x_{k1}, \ldots, x_{kp}, y_k)$ for which $(x_{k1}, \ldots, x_{kp})$ is outlying with respect to the $(x_{i1}, \ldots, x_{ip})$ in the data set. As before, such

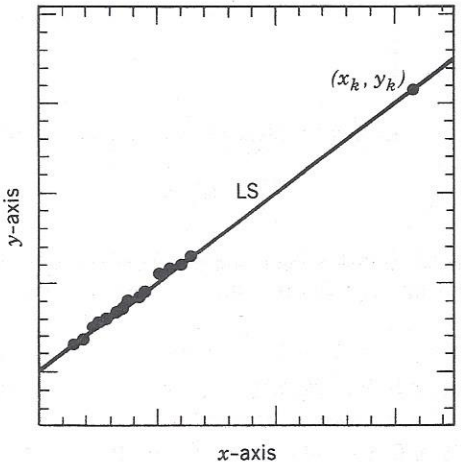


Figure 3. The point (x_k, y_k) is a leverage point because x_k is outlying. However, (x_k, y_k) is not a regression outlier because it matches the linear pattern set by the other data points.

leverage points have a potentially large influence on the LS regression coefficients, depending on the actual value of y_k. However, in this situation it is much more difficult to *identify* leverage points, because of the higher dimensionality. Indeed, it may be very difficult to discover such a point when there are 10 explanatory variables, which we can no longer visualize. A simple illustration of the problem is given in Figure 4, which plots x_{i1} versus x_{i2} for some data set. In this plot we easily see two leverage points, which are, however, invisible when the variables x_{i1} and x_{i2} are considered separately. (Indeed, the one-dimensional sample $\{x_{11}, x_{21}, \ldots, x_{n1}\}$ does not contain outliers, and neither does $\{x_{12}, x_{22}, \ldots, x_{n2}\}$.) In general, it is not sufficient to look at each variable separately or even at all plots of pairs of variables. The identification of outlying $(x_{i1}, \ldots, x_{ip})$ is a difficult problem, which will be treated in Subsection 1*d* of Chapter 7. However, in this book we are mostly concerned with *regression outliers*, that is, cases for which $(x_{i1}, \ldots, x_{ip}, y_i)$ deviates from the linear relation followed by the majority of the data, taking into account both the explanatory variables and the response variable simultaneously.

Many people will argue that regression outliers can be discovered by looking at the least squares residuals. Unfortunately, this is not true when the outliers are leverage points. For example, consider again Figure 2*b*. Case 1, being a leverage point, has tilted the LS line so much that it is now quite close to that line. Consequently, the residual $r_1 = y_1 - \hat{y}_1$ is a

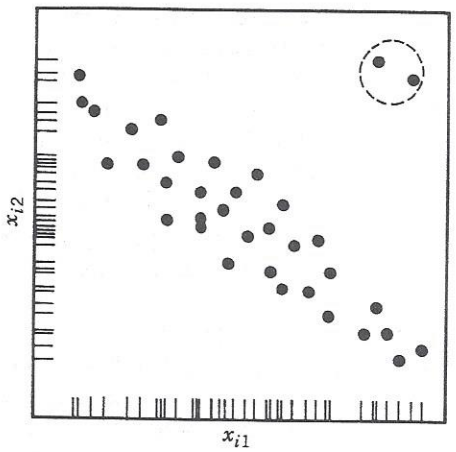


Figure 4. Plot of the explanatory variables (x_{i1}, x_{i2}) of a regression data set. There are two leverage points (indicated by the dashed circle), which are not outlying in either of the coordinates.

small (negative) number. On the other hand, the residuals r_2 and r_5 have much larger absolute values, although they correspond to "good" points. If one would apply a rule like "delete the points with largest LS residuals," then the "good" points would have to be deleted first! Of course, in such a bivariate data set there is really no problem at all because one can actually look at the data, but there are many multivariate data sets (like those of Chapter 3) where the outliers remain invisible even through a careful analysis of the LS residuals.

To conclude, regression outliers (either in x or in y) pose a serious threat to standard least squares analysis. Basically, there are two ways out of this problem. The first, and probably most well-known, approach is to construct so-called *regression diagnostics*. A survey of these techniques is provided in Chapter 6. Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed or corrected, followed by an LS analysis on the remaining cases. When there is only a single outlier, some of these methods work quite well by looking at the effect of deleting one point at a time. Unfortunately, it is much more difficult to diagnose outliers when there are several of them, and diagnostics for such multiple outliers are quite involved and often give rise to extensive computations (e.g., the number of all possible subsets is gigantic). Section 5 of Chapter 6 reports on recent developments in this direction, and in Section 6 of Chapter 6 a new diagnostic is proposed which can even cope with large fractions of outliers.

The other approach is *robust regression*, which tries to devise estimators that are not so strongly affected by outliers. Many statisticians who have vaguely heard of robustness believe that its purpose is to simply ignore the outliers, but this is not true. On the contrary, it is by looking at the residuals from a robust (or "resistant") regression that outliers may be identified, which usually cannot be done by means of the LS residuals. Therefore, diagnostics and robust regression really have the same goals, only in the opposite order: When using diagnostic tools, one first tries to delete the outliers and then to fit the "good" data by least squares, whereas a robust analysis first wants to fit a regression to the majority of the data and then to discover the outliers as those points which possess large residuals from that robust solution.

The following step is to think about the structure that has been uncovered. For instance, one may go back to the original data set and use subject-matter knowledge to study the outliers and explain their origin. Also, one should investigate if the deviations are not a symptom for model failure, which could, for instance, be repaired by adding a quadratic term or performing some transformation.

There are almost as many robust estimators as there are diagnostics, and it is necessary to measure their effectiveness in order to differentiate between them. In Section 2, some robust methods will be compared, essentially by counting the number of outliers that they can deal with. In subsequent chapters, the emphasis will be on the application of very robust methods, which can be used to analyze extremely messy data sets as well as clean ones.

2. THE BREAKDOWN POINT AND ROBUST ESTIMATORS

In Section 1 we saw that even a single regression outlier can totally offset the least squares estimator (provided it is far away). On the other hand, we will see that there exist estimators that can deal with data containing a certain percentage of outliers. In order to formalize this aspect, the *breakdown point* was introduced. Its oldest definition (Hodges 1967) was restricted to one-dimensional estimation of location, whereas Hampel (1971) gave a much more general formulation. Unfortunately, the latter definition was asymptotic and rather mathematical in nature, which may have restricted its dissemination. We prefer to work with a simple finite-sample version of the breakdown point, introduced by Donoho and Huber (1983). Take any sample of n data points,

$$Z = \{(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}\,, \tag{2.1}$$

and let T be a regression estimator. This means that applying T to such a sample Z yields a vector of regression coefficients as in (1.4):

$$T(Z) = \hat{\boldsymbol{\theta}}\,. \tag{2.2}$$

Now consider all possible corrupted samples Z' that are obtained by replacing any m of the original data points by arbitrary values (this allows for very bad outliers). Let us denote by bias$(m; T, Z)$ the maximum bias that can be caused by such a contamination:

$$\text{bias}\,(m; T, Z) = \sup_{Z'} \|T(Z') - T(Z)\|\,, \tag{2.3}$$

where the supremum is over all possible Z'. If bias$(m; T, Z)$ is infinite, this means that m outliers can have an arbitrarily large effect on T, which may be expressed by saying that the estimator "breaks down." Therefore, the (finite-sample) breakdown point of the estimator T at the sample Z is defined as

$$\varepsilon_n^*(T, Z) = \min\left\{\frac{m}{n};\ \text{bias}\,(m; T, Z)\ \text{is infinite}\right\}. \tag{2.4}$$

In other words, it is the smallest fraction of contamination that can cause the estimator T to take on values arbitrarily far from $T(Z)$. Note that this definition contains no probability distributions!

For least squares, we have seen that one outlier is sufficient to carry T over all bounds. Therefore, its breakdown point equals

$$\varepsilon_n^*(T, Z) = 1/n \tag{2.5}$$

which tends to zero for increasing sample size n, so it can be said that LS has a breakdown point of 0%. This again reflects the extreme sensitivity of the LS method to outliers.

A first step toward a more robust regression estimator came from Edgeworth (1887), improving a proposal of Boscovich. He argued that outliers have a very large influence on LS because the residuals r_i are squared in (1.7). Therefore, he proposed the *least absolute values* regression estimator, which is determined by

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} |r_i|. \tag{2.6}$$

(This technique is often referred to as L_1 regression, whereas least squares is L_2.) Before that time, Laplace had already used the same criterion (2.6) in the context of one-dimensional observations, obtaining the sample median (and the corresponding error law, which is now called the *double exponential* or *Laplace distribution*). The L_1 regression estimator, like the median, is not completely unique (see, e.g., Harter 1977 and Gentle et al. 1977). But whereas the breakdown point of the univariate median is as high as 50%, unfortunately the breakdown point of L_1 regression is still no better than 0%. To see why, let us look at Figure 5.

Figure 5 gives a schematic summary of the effect of outliers on L_1 regression. Figure 5*a* shows the effect of an outlier in the y-direction, in the same situation as Figure 1. Unlike least squares, the L_1 regression line is robust with respect to such an outlier, in the sense that it (approximately) remains where it was when observation 4 was still correct, and still fits the remaining points nicely. Therefore, L_1 protects us against outlying y_i and is quite preferable over LS in this respect. In recent years, the L_1 approach to statistics appears to have gained some ground (Bloomfield

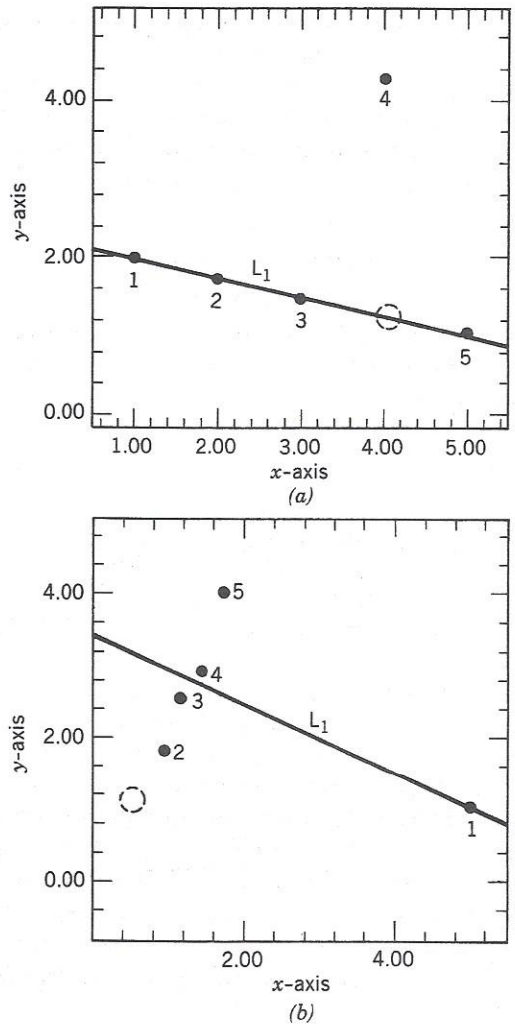


Figure 5. (*a*) Robustness of L_1 regression with respect to an outlier in the y-direction. (*b*) Sensitivity of L_1 regression to an outlier in the x-direction ("leverage point").

and Steiger 1980, 1983; Narula and Wellington 1982; Devroye and Gyorfi 1984). However, L_1 regression does not protect against outlying x, as we can see from Figure 5*b*, where the effect of the leverage point is even stronger than on the LS line in Figure 2. It turns out that when the leverage point lies far enough away, the L_1 line passes right through it (see exercise 10 below). Therefore, a single erroneous observation can

totally offset the L_1 estimator, so its finite-sample breakdown point is also equal to $1/n$.

The next step in this direction was the use of *M-estimators* (Huber 1973, p. 800; for a recent survey see Huber 1981). They are based on the idea of replacing the squared residuals r_i^2 in (1.7) by another function of the residuals, yielding

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} \rho(r_i) , \tag{2.7}$$

where ρ is a symmetric function [i.e., $\rho(-t) = \rho(t)$ for all t] with a unique minimum at zero. Differentiating this expression with respect to the regression coefficients $\hat{\theta}_j$ yields

$$\sum_{i=1}^{n} \psi(r_i)\mathbf{x}_i = \mathbf{0} , \tag{2.8}$$

where ψ is the derivative of ρ, and $\mathbf{x}_i$ is the row vector of explanatory variables of the ith case:

$$\begin{aligned} \mathbf{x}_i &= (x_{i1}, \ldots, x_{ip}) \\ \mathbf{0} &= (0, \ldots, 0) . \end{aligned} \tag{2.9}$$

Therefore (2.8) is really a system of p equations, the solution of which is not always easy to find: In practice, one uses iteration schemes based on reweighted LS (Holland and Welsch 1977) or the so-called *H-algorithm* (Huber and Dutter 1974, Dutter 1977, Marazzi 1980). Unlike (1.7) or (2.6), however, the solution of (2.8) is not equivariant with respect to a magnification of the y-axis. (We use the word "equivariant" for statistics that transform properly, and we reserve "invariant" for quantities that remain unchanged.) Therefore, one has to standardize the residuals by means of some estimate of σ, yielding

$$\sum_{i=1}^{n} \psi(r_i/\hat{\sigma})\mathbf{x}_i = \mathbf{0} , \tag{2.10}$$

where $\hat{\sigma}$ must be estimated simultaneously. Motivated by minimax asymptotic variance arguments, Huber proposed to use the function

$$\psi(t) = \min(c, \max(t, -c)) . \tag{2.11}$$

M-estimators with (2.11) are statistically more efficient (at a model with

Gaussian errors) than L_1 regression, while at the same time they are still robust with respect to outlying y_i. However, their breakdown point is again $1/n$ because of the effect of outlying $\mathbf{x}_i$.

Because of this vulnerability to leverage points, *generalized M-estimators* (GM-estimators) were introduced, with the basic purpose of bounding the influence of outlying $\mathbf{x}_i$ by means of some weight function w. Mallows (1975) proposed to replace (2.10) by

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi(r_i/\hat{\sigma})\mathbf{x}_i = \mathbf{0}\,, \tag{2.12}$$

whereas Schweppe (see Hill 1977) suggested using

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi(r_i/(w(\mathbf{x}_i)\hat{\sigma}))\mathbf{x}_i = \mathbf{0}\,. \tag{2.13}$$

These estimators were constructed in the hope of bounding the influence of a single outlying observation, the effect of which can be measured by means of the so-called *influence function* (Hampel 1974). Based on such criteria, optimal choices of ψ and w were made (Hampel 1978, Krasker 1980, Krasker and Welsch 1982, Ronchetti and Rousseeuw 1985, and Samarov 1985; for a recent survey see Chapter 6 of Hampel et al. 1986). Therefore, the corresponding GM-estimators are now generally called *bounded-influence estimators*. It turns out, however, that the breakdown point of all GM-estimators can be no better than a certain value that decreases as a function of p, where p is again the number of regression coefficients (Maronna, Bustos, and Yohai 1979). This is very unsatisfactory, because it means that the breakdown point diminishes with increasing dimension, where there are more opportunities for outliers to occur. Furthermore, it is not clear whether the Maronna–Bustos–Yohai upper bound can actually be attained, and if it can, it is not clear as to which GM-estimator can be used to achieve this goal. In Section 7 of Chapter 2, a small comparative study will be performed in the case of simple regression ($p = 2$), indicating that not all GM-estimators achieve the same breakdown point. But, of course, the real problem is with higher dimensions.

Various other estimators have been proposed, such as the methods of Wald (1940), Nair and Shrivastava (1942), Bartlett (1949), and Brown and Mood (1951); the median of pairwise slopes (Theil 1950, Adichie 1967, Sen 1968); the resistant line (Tukey 1970/1971, Velleman and Hoaglin 1981, Johnstone and Velleman 1985b); R-estimators (Jureckovà 1971, Jaeckel 1972); L-estimators (Bickel 1973, Koenker and Bassett

1978); and the method of Andrews (1974). Unfortunately, in simple regression, none of these methods achieves a breakdown point of 30%. Moreover, some of them are not even defined for $p > 2$.

All this raises the question as to whether robust regression with a high breakdown point is at all possible. The affirmative answer was given by Siegel (1982), who proposed the *repeated median* estimator with a 50% breakdown point. Indeed, 50% is the best that can be expected (for larger amounts of contamination, it becomes impossible to distinguish between the "good" and the "bad" parts of the sample, as will be proven in Theorem 4 of Chapter 3). Siegel's estimator is defined as follows: For any p observations

$$(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_p}, y_{i_p})$$

one computes the parameter vector which fits these points exactly. The jth coordinate of this vector is denoted by $\theta_j(i_1, \ldots, i_p)$. The repeated median regression estimator is then defined coordinatewise as

$$\hat{\theta}_j = \underset{i_1}{\text{med}} (\ldots (\underset{i_{p-1}}{\text{med}} (\underset{i_p}{\text{med}} \theta_j(i_1, \ldots, i_p))) \ldots) . \tag{2.14}$$

This estimator can be computed explicitly, but requires consideration of all subsets of p observations, which may cost a lot of time. It has been successfully applied to problems with small p. But unlike other regression estimators, the repeated median is not equivariant for linear transformations of the $\mathbf{x}_i$, which is due to its coordinatewise construction.

Let us now consider the equivariant and high-breakdown regression methods that form the core of this book. To introduce them, let us return to (1.7). A more complete name for the LS method would be *least sum of squares*, but apparently few people have objected to the deletion of the word "sum"—as if the only sensible thing to do with n positive numbers would be to add them. Perhaps as a consequence of its historical name, several people have tried to make this estimator robust by replacing the square by something else, not touching the summation sign. Why not, however, replace the sum by a median, which is very robust? This yields the *least median of squares* (LMS) estimator, given by

$$\underset{\hat{\theta}}{\text{Minimize}} \underset{i}{\text{med}} r_i^2 \tag{2.15}$$

(Rousseeuw 1984). This proposal was essentially based on an idea of Hampel (1975, p. 380). It turns out that this estimator is very robust with respect to outliers in y as well as outliers in $\mathbf{x}$. It will be shown in Section

4 of Chapter 3 that its breakdown point is 50%, the highest possible value. The LMS is clearly equivariant with respect to linear transformations on the explanatory variables, because (2.15) only makes use of the residuals. In Section 5 of Chapter 3, we will show that the LMS is related to projection pursuit ideas, whereas the most useful algorithm for its computation (Section 1 of Chapter 5) is reminiscent of the bootstrap (Diaconis and Efron 1983). Unfortunately, the LMS performs poorly from the point of view of asymptotic efficiency (in Section 4 of Chapter 4, we will prove it has an abnormally slow convergence rate).

To repair this, Rousseeuw (1983, 1984) introduced the *least trimmed squares* (LTS) estimator, given by

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{h} (r^2)_{i:n} , \tag{2.16}$$

where $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ are the ordered squared residuals (note that the residuals are first squared and then ordered). Formula (2.16) is very similar to LS, the only difference being that the largest squared residuals are not used in the summation, thereby allowing the fit to stay away from the outliers. In Section 4 of Chapter 4, we will show that the LTS converges at the usual rate and compute its asymptotic efficiency. Like the LMS, this estimator is also equivariant for linear transformations on the $\mathbf{x}_i$ and is related to projection pursuit. The best robustness properties are achieved when h is approximately $n/2$, in which case the breakdown point attains 50%. (The exact optimal value of h will be discussed in Section 4 of Chapter 3.)

Both the LMS and the LTS are defined by minimizing a robust measure of the scatter of the residuals. Generalizing this, Rousseeuw and Yohai (1984) introduced so-called *S-estimators*, corresponding to

$$\underset{\hat{\theta}}{\text{Minimize}}\ S(\boldsymbol{\theta}) , \tag{2.17}$$

where $S(\boldsymbol{\theta})$ is a certain type of robust M-estimate of the scale of the residuals $r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})$. The technical definition of S-estimators will be given in Section 4 of Chapter 3, where it is shown that their breakdown point can also attain 50% by a suitable choice of the constants involved. Moreover, it turns out that S-estimators have essentially the same asymptotic performance as regression M-estimators (see Section 4 of Chapter 3). However, for reasons of simplicity we will concentrate primarily on the LMS and the LTS.

Figure 6 illustrates the robustness of these new regression estimators with respect to an outlier in y or in $\mathbf{x}$. Because of their high breakdown

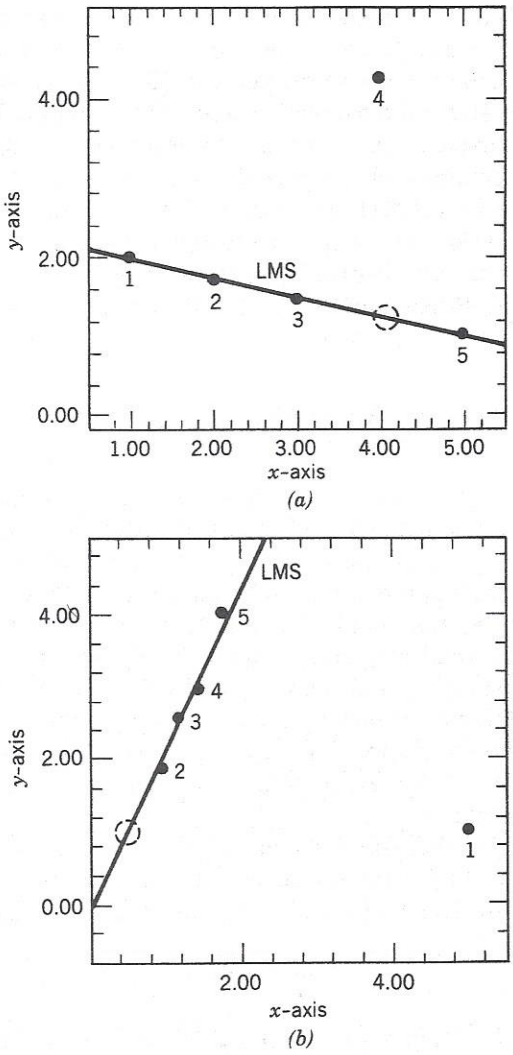


Figure 6. Robustness of LMS regression with respect to (*a*) an outlier in the *y*-direction, and (*b*) an outlier in the *x*-direction.

point, these estimators can even cope with several outliers at the same time (up to about $n/2$ of them, although, of course, this will rarely be needed in practice). This resistance is also independent of p, the number of explanatory variables, so LMS and LTS are reliable data analytic tools that may be used to discover regression outliers in such multivariate situations. The basic principle of LMS and LTS is to fit the *majority* of the

data, after which outliers may be identified as those points that lie far away from the robust fit, that is, the cases with large positive or large negative residuals. In Figure 6*a*, the 4th case possesses a considerable residual, and that of case 1 in Figure 6*b* is even more apparent.

However, in general the y_i (and hence the residuals) may be in any unit of measurement, so in order to decide if a residual r_i is "large" we need to compare it to an estimate $\hat{\sigma}$ of the error scale. Of course, this scale estimate $\hat{\sigma}$ has to be robust itself, so it depends only on the "good" data and does not get blown up by the outlier(s). For LMS regression, one could use an estimator such as

$$\hat{\sigma} = C_1 \sqrt{\underset{i}{\text{med}}\, r_i^2}\,, \tag{2.18}$$

where r_i is the residual of case i with respect to the LMS fit. The constant C_1 is merely a factor used to achieve consistency at Gaussian error distributions. (In Section 1 of Chapter 5, a more refined version of $\hat{\sigma}$ will be discussed, which makes a correction for small samples.) For the LTS, one could use a rule such as

$$\hat{\sigma} = C_2 \sqrt{\frac{1}{n} \sum_{i=1}^{h} (r^2)_{i:n}}\,, \tag{2.19}$$

where C_2 is another correction factor. In either case, we shall identify case i as an outlier if and only if $|r_i/\hat{\sigma}|$ is large. (Note that this ratio does not depend on the original measurement units!)

This brings us to another idea. In order to improve on the crude LMS and LTS solutions, and in order to obtain standard quantities like t-values, confidence intervals, and the like, we can apply a *weighted least squares* analysis based on the identification of the outliers. For instance, we could make use of the following weights:

$$w_i = \begin{cases} 1 & \text{if } |r_i/\hat{\sigma}| \le 2.5 \\ 0 & \text{if } |r_i/\hat{\sigma}| > 2.5\,. \end{cases} \tag{2.20}$$

This means simply that case i will be retained in the weighted LS if its LMS residual is small to moderate, but disregarded if it is an outlier. The bound 2.5 is, of course, arbitrary, but quite reasonable because in a Gaussian situation there will be very few residuals larger than $2.5\hat{\sigma}$. Instead of 'hard" rejection of outliers as in (2.20), one could also apply "smooth" rejection, for instance by using continuous functions of $|r_i/\hat{\sigma}|$, thereby allowing for a region of doubt (e.g., points with $2 \le |r_i/\hat{\sigma}| \le 3$ could be given weights between 1 and 0). Anyway, we then apply

3. Find (or construct) an example of a good leverage point and a bad leverage point. Are these points easy to identify by means of their LS residuals?

Section 2

4. Show that least squares and least absolute deviations are M-estimators.
5. When all y_i are multiplied by a nonzero constant, show that the least squares (1.7) and least absolute deviations (2.6) estimates, as well as the LMS (2.15) and LTS (2.16) estimates, are multiplied by the same factor.
6. Obtain formula (2.8) by differentiating (2.7) with respect to the coefficients $\hat{\theta}_j$, keeping in mind that r_i is given by (1.6).
7. In the special case of simple regression, write down Siegel's repeated median estimates of slope and intercept, making use of (2.14). What does this estimator reduce to in a one-dimensional location setting?
8. The following real data come from a large Belgian insurance company. Table 1 shows the monthly payments in 1979, made as a result of the end of period of life-insurance contracts. (Because of company rules, the payments are given as a percentage of the total amount in that year.) In December a very large sum was paid, mainly because of one extremely high supplementary pension.

Table 1. Monthly Payments in 1979

Month (x)	Payment (y)
1	3.22
2	9.62
3	4.50
4	4.94
5	4.02
6	4.20
7	11.24
8	4.53
9	3.05
10	3.76
11	4.23
12	42.69

Source: Rousseeuw et al. (1984a)

(a) Make a scatterplot of the data. Is the December value an outlier in the x-direction or in the y-direction? Are there any other outliers?

(b) What is the trend of the good points (of the majority of the data)? Fit a robust line by eye. For this line, compute $\Sigma_{i=1}^{n} r_i^2$, $\Sigma_{i=1}^{n} |r_i|$, and $\text{med}_i\, r_i^2$.

(c) Compute the LS line (e.g., by means of a standard statistical program) and plot it in the same figure. Also compute $\Sigma_{i=1}^{n} r_i^2$, $\Sigma_{i=1}^{n} |r_i|$, and $\text{med}_i\, r_i^2$ for this line, and explain why the lines are so different.

(d) Compute both the Pearson product-moment correlation coefficient and the Spearman rank correlation coefficient, and relate them to (b) and (c).

9. Show that the weighted least squares estimate defined by (2.21) can be computed by replacing all $(\mathbf{x}_i, y_i)$ by $(w_i^{1/2}\mathbf{x}_i, w_i^{1/2}y_i)$ and then applying ordinary least squares.

10. (E. Ronchetti) Let $\bar{x}$ be the average of all x_i in the data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Suppose that x_1 is an outlier, which is so far away that all remaining x_i lie on the other side of $\bar{x}$ (as in Figure 5*b*). Then show that the L_1 regression line goes right through the leverage point (x_1, y_1). (Hint: assume that the L_1 line does not go through (x_1, y_1) and show that $\Sigma_{i=1}^{n} |r_i|$ will decrease when the line is tilted about $\bar{x}$ to reduce $|r_1|$.)

CHAPTER 2

Simple Regression

1. MOTIVATION

The simple regression model

$$y_i = \theta_1 x_i + \theta_2 + e_i \qquad (i = 1, \ldots, n) \tag{1.1}$$

has been used in Chapter 1 for illustrating some problems that occur when fitting a straight line to a two-dimensional data set. With the aid of some scatterplots, we showed the effect of outliers in the y-direction and of outliers in the x-direction on the ordinary least squares (LS) estimates (see Figures 1 and 2 of Chapter 1). In this chapter we would like to apply high-breakdown regression techniques that can cope with these problems. We treat simple regression separately for didactic reasons, because in this situation it is easy to see the outliers. In Chapter 3, the methods will be generalized to the multiple regression model.

The phrase "simple regression" is also sometimes used for a linear model of the type

$$y_i = \theta_1 x_i + e_i \qquad (i = 1, \ldots, n)\,, \tag{1.2}$$

which does not have a constant term. This model can be used in applications where it is natural to assume that the response should become zero when the explanatory variable takes on the value zero. Graphically, it corresponds to a straight line passing through the origin. Some examples will be given in Section 6.

The following example illustrates the need for a robust regression technique. We have resorted to the so-called *Pilot-Plant data* (Table 1) from Daniel and Wood (1971). The response variable corresponds to the

Table 1. Pilot-Plant Data Set

Observation (i)	Extraction (x_i)	Titration (y_i)
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88

Source: Daniel and Wood (1971).

acid content determined by titration, and the explanatory variable is the organic acid content determined by extraction and weighing. Yale and Forsythe (1976) also analyzed this data set.

The scatterplot (Figure 1) suggests a strong statistical relationship between the response and the explanatory variable. The tentative assumption of a linear model such as (1.1) appears to be reasonable.

The LS fit is

$$\hat{y} = 0.322x + 35.458 \qquad \text{(dashed line)}.$$

The least median of squares (LMS) line, defined by formula (2.15) of Chapter 1, corresponds to

$$\hat{y} = 0.311x + 36.519 \qquad \text{(solid line)}.$$

In examining the plot, we see no outliers. As could be expected in such

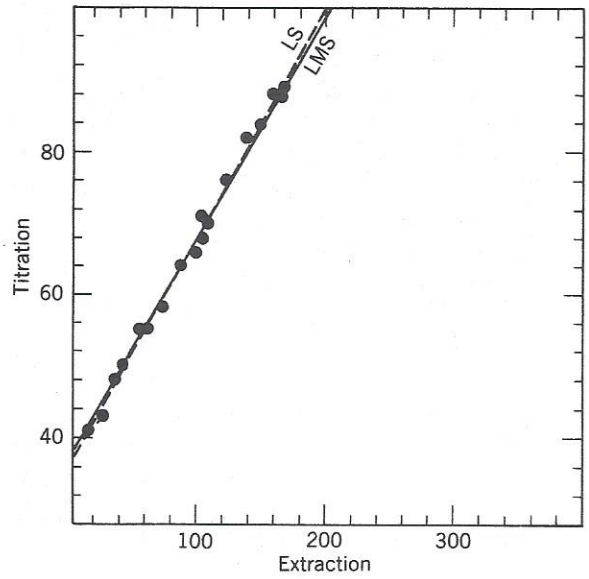


Figure 1. Pilot-Plant data with LS fit (dashed line) and LMS fit (solid line).

a case, only marginal differences exist between the robust estimates and those based on least squares.

Suppose now that one of the observations has been wrongly recorded. For example, the x-value of the 6th observation might have been registered as 370 instead of 37. This error produces an outlier in the x-direction, which is surrounded by a dashed circle in the scatterplot in Figure 2.

What will happen with the regression coefficients for this contaminated sample? The least squares result

$$\hat{y} = 0.081x + 58.939$$

corresponds to the dashed line in Figure 2. It has been attracted very strongly by this single outlier, and therefore fits the other points very badly. On the other hand, the solid line was obtained by applying least median of squares, yielding

$$\hat{y} = 0.314x + 36.343\,.$$

This robust method has succeeded in staying away from the outlier, and

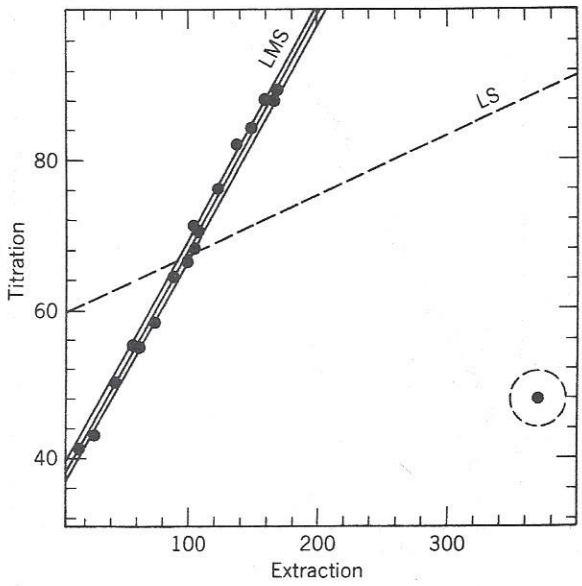


Figure 2. Same data set as in Figure 1, but with one outlier. The dashed line corresponds to the LS fit. The solid LMS line is surrounded by the narrowest strip containing half of the points.

yields a good fit to the majority of the data. Moreover, it lies close to the LS estimate applied to the original uncontaminated data. It would be wrong to say that the robust technique ignores the outlier. On the contrary, the LMS fit exposes the presence of such points.

The LMS solution for simple regression with intercept is given by

$$\underset{\hat{\theta}_1, \hat{\theta}_2}{\text{Minimize}} \underset{i}{\text{med}} (y_i - \hat{\theta}_1 x_i - \hat{\theta}_2)^2 . \tag{1.3}$$

Geometrically, it corresponds to *finding the narrowest strip covering half of the observations*. (To be precise, by "half" we mean $[n/2] + 1$, where $[n/2]$ denotes the integer part of $n/2$. Moreover, the thickness of this strip is measured in the vertical direction.) The LMS line lies exactly at the middle of this band. (We will prove this fact in Theorem 1 of Chapter 4, Section 2.) Note that this notion is actually much easier to explain to most people than the classical LS definition. For the contaminated Pilot-Plant data, this strip is drawn in Figure 2.

The outlier in this example was artificial. However, it is important to realize that this kind of mistake appears frequently in real data. Outlying

data points can be present in a sample because of errors in recording observations, errors in transcription or transmission, or an exceptional occurrence in the investigated phenomenon. In the two-dimensional case (such as the example above), it is rather easy to detect atypical points just by plotting the observations. This visual tracing is no longer possible for higher dimensions. So in practice, one needs a procedure that is able to lessen the impact of outliers, thereby exposing them in the residual plots (examples of this are given in Section 3). In addition, when no outliers occur, the result of the alternative procedure should hardly differ from the LS solution. It turns out that LMS regression does meet these requirements.

Let us now look at some real data examples with outliers. In the Belgian Statistical Survey (published by the Ministry of Economy), we found a data set containing the total number (in tens of millions) of international phone calls made. These data are listed in Table 2 and plotted in Figure 3.

The plot seems to show an upward trend over the years. However, this time series contains heavy contamination from 1964 to 1969. Upon inquiring, it turned out that from 1964 to 1969, another recording system

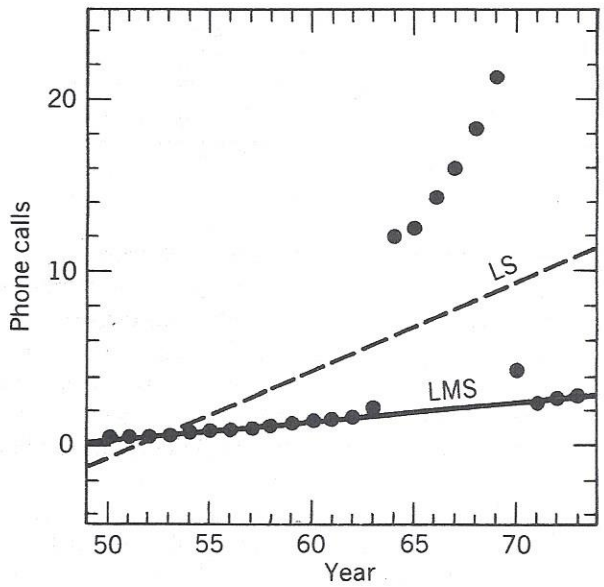


Figure 3. Number of international phone calls from Belgium in the years 1950–1973 with the LS (dashed line) and LMS fit (solid line).

Table 2. Number of International Calls from Belgium

Year (x_i)	Number of Calls[a] (y_i)
50	0.44
51	0.47
52	0.47
53	0.59
54	0.66
55	0.73
56	0.81
57	0.88
58	1.06
59	1.20
60	1.35
61	1.49
62	1.61
63	2.12
64	11.90
65	12.40
66	14.20
67	15.90
68	18.20
69	21.20
70	4.30
71	2.40
72	2.70
73	2.90

[a]In tens of millions.

was used, giving the total number of *minutes* of these calls. (The years 1963 and 1970 are also partially affected because the transitions did not happen exactly on New Year's Day, so the number of calls of some months were added to the number of minutes registered in the remaining months!) This caused a large fraction of outliers in the y-direction.

The ordinary LS solution for these data is given by $\hat{y} = 0.504x - 26.01$ and corresponds to the dashed line in Figure 3. This dashed line has been affected very much by the y values associated with the years 1964–1969. As a consequence, the LS line has a large slope and does not fit the good or the bad data points. This is what one would obtain by not looking critically at these data and by applying the LS method in a routine way. In fact, some of the good observations (such as the 1972 one) yield even larger LS residuals than some of the bad values! Now let us apply the

LMS method. This yields $\hat{y} = 0.115x - 5.610$ (plotted as a solid line in Figure 3), which avoids the outliers. It corresponds to the pattern one sees emerging when simply looking at the plotted data points. Clearly, this line fits the majority of the data. (This is not meant to imply that a linear fit is necessarily the best model, because collecting more data might reveal a more complicated kind of relationship.)

Another example comes from astronomy. The data in Table 3 form the Hertzsprung–Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. Here x is the logarithm of the effective temperature at the surface of the star (T_e), and y is the logarithm of its light intensity (L/L_0). These numbers were given to us by C. Doom (personal communication), who extracted the raw data from Humphreys (1978) and performed the calibration according to Vansina and De Greve (1982).

Table 3. Data for the Hertzsprung–Russell Diagram of the Star Cluster CYG OB1

Index of Star (i)	$\log T_e$ (x_i)	$\log [L/L_0]$ (y_i)	Index of Star (i)	$\log T_e$ (x_i)	$\log [L/L_0]$ (y_i)
1	4.37	5.23	25	4.38	5.02
2	4.56	5.74	26	4.42	4.66
3	4.26	4.93	27	4.29	4.66
4	4.56	5.74	28	4.38	4.90
5	4.30	5.19	29	4.22	4.39
6	4.46	5.46	30	3.48	6.05
7	3.84	4.65	31	4.38	4.42
8	4.57	5.27	32	4.56	5.10
9	4.26	5.57	33	4.45	5.22
10	4.37	5.12	34	3.49	6.29
11	3.49	5.73	35	4.23	4.34
12	4.43	5.45	36	4.62	5.62
13	4.48	5.42	37	4.53	5.10
14	4.01	4.05	38	4.45	5.22
15	4.29	4.26	39	4.53	5.18
16	4.42	4.58	40	4.43	5.57
17	4.23	3.94	41	4.38	4.62
18	4.42	4.18	42	4.45	5.06
19	4.23	4.18	43	4.50	5.34
20	3.49	5.89	44	4.45	5.34
21	4.29	4.38	45	4.55	5.54
22	4.29	4.22	46	4.45	4.98
23	4.42	4.42	47	4.42	4.50
24	4.49	4.85			

The Hertzsprung–Russell diagram itself is shown in Figure 4. It is the scatterplot of these points, where the log temperature x is plotted from right to left. In the plot, one sees two groups of points: the majority, which seems to follow a steep band, and the four stars in the upper right corner. These parts of the diagram are well known in astronomy: The 43 stars are said to lie on the main sequence, whereas the four remaining stars are called giants. (The giants are the points with indices 11, 20, 30, and 34.)

Application of our LMS estimator to these data yields the solid line $\hat{y} = 3.898x - 12.298$, which fits the main sequence nicely. On the other hand, the LS solution $\hat{y} = -0.409x + 6.78$ corresponds to the dashed line in Figure 4, which has been pulled away by the four giant stars (which it does not fit well either). These outliers are leverage points, but they are not errors: It would be more appropriate to say that the data come from two different populations. These two groups can easily be distinguished on the basis of the LMS residuals (the large residuals correspond to the giant stars), whereas the LS residuals are rather homogeneous and do not allow us to separate the giants from the main-sequence stars.

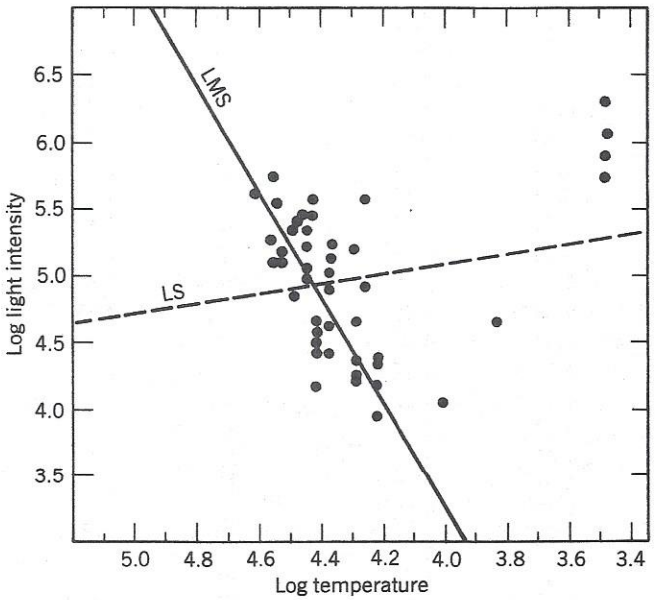


Figure 4. Hertzsprung–Russell diagram of the star cluster CYG OB1 with the LS (dashed line) and LMS fit (solid line).

2. COMPUTATION OF THE LEAST MEDIAN OF SQUARES LINE

The present section describes the use of PROGRESS, a program implementing LMS regression. (Its name comes from Program for RObust reGRESSion.) The algorithm itself is explained in detail in Chapter 5. Without the aid of a computer, it would never have been possible to calculate high-breakdown regression estimates. Indeed, one does not have an explicit formula, such as the one used for LS. It appears there are deep reasons why high-breakdown regression cannot be computed cheaply. [We are led to this assertion by means of partial results from our own research and because of some arguments provided by Donoho (1984) and Steele and Steiger (1986).] Fortunately, the present evolution of computers has made robust regression quite feasible.

PROGRESS is designed to run on an IBM-PC or a compatible microcomputer. At least 256K RAM must be available. The boundaries of the arrays in the program allow regression analysis with at most 300 cases and 10 coefficients. PROGRESS starts by asking the data specifications and the options for treatment and output. This happens in a fully interactive way, which makes it very easy to use the program. The user only has to answer the questions appearing on the screen. No knowledge of informatics or computer techniques is required. Nevertheless, we will devote this section to the input. [The mainframe version described in Leroy and Rousseeuw (1984) was written in very portable FORTRAN, so it was not yet interactive.] We will treat the Pilot-Plant example (with outlier) of the preceding section. The words typed by the user are printed in italics to distinguish them from the words or lines coming from PROGRESS.

The first thing to do, of course, is to insert the diskette containing the program. In order to run PROGRESS, the user only has to type A:PROGRESS in case the program is on drive A. (Other possibilities would be drive B or hard disk C.) Then the user has to press the ENTER key. Having done this, the program generates the following screen:

```
****************
*  PROGRESS    *
****************

ENTER THE NUMBER OF CASES PLEASE: 20
```

The user now has to enter the number of cases he or she wants to handle in the analysis. Note that there are limits to the size of the data

sets that can be treated. (This restriction is because of central memory limitations of the computer.) Therefore, PROGRESS gives a warning when the number of cases entered by the user is greater than this limit. When PROGRESS has accepted the number of cases, the following question appears:

```
DO YOU WANT A CONSTANT TERM IN THE REGRESSION?
PLEASE ANSWER YES OR NO: YES
```

When the user answers YES to this question, PROGRESS performs a regression with a constant term. Otherwise, the program yields a regression through the origin. The general models for regression with and without a constant are, respectively,

$$y_i = x_{i1}\theta_1 + \cdots + x_{i,p-1}\theta_{p-1} + \theta_p + e_i \qquad (i = 1, \ldots, n) \qquad (2.1)$$

and

$$y_i = x_{i1}\theta_1 + \cdots + x_{i,p-1}\theta_{p-1} + x_{ip}\theta_p + e_i \qquad (i = 1, \ldots, n). \qquad (2.2)$$

In (2.2) the estimate of y_i is equal to zero when all x_{ij} $(j = 1, \ldots, p)$ are zero. [Note that (2.1) is a special case of (2.2), obtained by putting the last explanatory variable x_{ip} equal to 1 for all cases.]

It may happen that the user has a large data set, consisting of many more variables than those he or she wishes to insert in a regression model. PROGRESS allows the user to select some variables out of the entire set. Furthermore, for each variable in the regression, PROGRESS asks for a label in order to facilitate the interpretation of the output. Therefore the user has to answer the following questions:

```
WHAT IS THE TOTAL NUMBER OF VARIABLES IN YOUR DATA SET?
 ----------------------------------------
PLEASE GIVE A NUMBER BETWEEN 1 AND 50: 5

WHICH VARIABLE DO YOU CHOOSE AS RESPONSE VARIABLE?
 ------------------------------------
OUT OF THESE 5 GIVE ITS POSITION: 4

GIVE A LABEL FOR THIS VARIABLE (AT MOST 10 CHARACTERS): TITRATION

HOW MANY EXPLANATORY VARIABLES DO YOU WANT TO USE IN THE ANALYSIS?
 ------------------------------------------------
(AT MOST 4): 1
```

The answer to each question is verified by PROGRESS. This means that a message is given when an answer is not allowed. For example,

when the user answers 12 to the question

```
WHICH VARIABLE DO YOU CHOOSE AS RESPONSE VARIABLE?
 ------------------------------------
OUT OF THESE 5 GIVE ITS POSITION: 12
```

the following prompt will appear:

```
NOT ALLOWED! ENTER YOUR CHOICE AGAIN: 4
```

Also, the program checks whether the number of cases is more than twice the number of regression coefficients (including the constant term if there is one). If there are fewer cases, the program stops.

The question

```
HOW MANY EXPLANATORY VARIABLES DO YOU WANT TO USE IN THE ANALYSIS?
 ------------------------------------------------
(AT MOST 4):
```

may be answered with 0. In that situation the response variable is analyzed in a one-dimensional way, yielding robust estimates of its location and scale. (More details on this can be found in Chapter 4.)

When the number of explanatory variables is equal to the number of remaining variables (this means, all but the response variable) in the data set, the user has to fill up a table containing one line for each explanatory variable. Each of these variables is identified by means of a label of at most 10 characters. These characters have to be typed below the arrows. On the other hand, when the number of explanatory variables is less, the user also has to give the position of the selected variable in the data set together with the corresponding label. For our example, this table would be

```
EXPLANATORY VARIABLES     :  POSITION  LABEL (AT MOST 10 CHARACTERS)
-----------------------↓↓↓↓-- ↓↓↓↓↓↓↓↓↓↓----------
NUMBER 1                  :    2       EXTRACTION
```

An option concerning the amount of output can be chosen in the following question:

```
HOW MUCH OUTPUT DO YOU WANT?
 --------------------
0=SMALL OUTPUT       : LIMITED TO BASIC RESULTS
1=MEDIUM-SIZED OUTPUT: ALSO INCLUDES A TABLE WITH THE OBSERVED VALUES OF Y,
                       THE ESTIMATES OF Y, THE RESIDUALS AND THE WEIGHTS
2=LARGE OUTPUT       : ALSO INCLUDES THE DATA ITSELF
ENTER YOUR CHOICE: 2
```

If the user types 0, the output is limited to the basic results, namely the LS, the LMS, and the reweighted least squares (RLS) estimates, with their standard deviations (in order to construct confidence intervals around the estimated regression coefficients) and t-values. The scale estimates are also given. In the case of regression with one explanatory variable, a plot of y versus x is produced. This permits us to detect a pattern in the data.

Setting the print option at 1 yields more information: a table with the observed values of y, the estimated values of y, the residuals, and the residuals divided by the scale estimate (which are called *standardized residuals*); and for reweighted least squares, an additional column with the weight (resulting from LMS) of each observation. Apart from the output produced with print option 1, print option 2 also lists the data itself.

A careful analysis of residuals is an important part of applied regression. Therefore we have added a plot option that permits us to obtain a plot of the standardized residuals versus the estimated value of y (this is performed when the plot option is set at 1) or a plot of the standardized residuals versus the index i of the observation (which is executed when the plot option is set at 2). If the plot option is set at 3, both types of plots are given. If the plot option is set at 0, the output contains no residual plots. The plot option is selected by means of the following question:

```
DO YOU WANT TO LOOK AT THE RESIDUALS?
 ----------------------------
0=NO RESIDUAL PLOTS
1=PLOT OF THE STANDARDIZED RESIDUALS VERSUS THE ESTIMATED VALUE OF Y
2=PLOT OF THE STANDARDIZED RESIDUALS VERSUS THE INDEX OF THE OBSERVATION
3=PERFORMS BOTH TYPES OF RESIDUAL PLOTS
ENTER YOUR CHOICE: 0
```

When the following question is answered with YES, the program yields some outlier diagnostics, which will be described in Chapter 6.

```
DO YOU WANT TO COMPUTE OUTLIER DIAGNOSTICS?
PLEASE ANSWER YES OR NO: NO
```

When the data set has already been stored in a file, the user only has to give the name of that file in response to the following question. If such a file does not already exist, the user still has the option of entering his or her data by keyboard in an interactive way during a PROGRESS session. In that case the user has to answer KEY. The entered data set has to contain as many variables as mentioned in the third question of the