JOURNAL OF CHEMOMETRICS, VOL. 5, 1–20 (1991)

TUTORIAL TO ROBUST STATISTICS

PETER J. ROUSSEEUW
University of Antwerp (UIA), Vesaliuslaan 24, B-2650 Edegem, Belgium

SUMMARY

In this tutorial we first illustrate the effect of outliers on classical statistics such as the sample average. This motivates the use of robust techniques. For univariate data the sample median is a robust estimator of location, and the dispersion can also be estimated robustly. The resulting 'z-scores' are well suited to detect outliers. The sample median can be generalized to very large data sets, which is useful for robust 'averaging' of curves or images. For multivariate data a robust regression procedure is described. Its standardized residuals allow us to identify the outliers. Finally, a survey of related approaches is given. (This review overlaps with earlier work by the same author, which appeared elsewhere.)

KEY WORDS Averaging Median Outliers Regression Residuals Robustness

1. INTRODUCTION

The least squares method is currently the most popular approach to estimation because of tradition and ease of computation. However, real data sets frequently contain outliers, which may be mistakes or exceptional observations. In this situation least squares becomes unreliable. Two things often happen: the estimates become totally incorrect and (somewhat surprisingly) the outliers themselves are hidden, which means that one does not notice them at all. To remedy this problem, robust statistical techniques have been developed that (a) still give a trustworthy answer when the data are contaminated and (b) allow us to easily identify the outliers at the same time.

The structure of this paper is as follows. In Section 2 we introduce the notions of outliers and robustness in the special case of estimating a central value of a batch of numbers. The discussion focuses on the comparison of the sample average with the sample median. Also, we consider the situation of very large data sets and its application to robust 'averaging' of curves and images.

In Section 3 we arrive at the regression situation, starting with simple regression and then continuing with multiple regression. We then consider some alternative regression estimators and finally mention other situations in which robust techniques have been made available.

Throughout this review we restrict attention to robust methods that are both intuitively appealing and very powerful (in the sense that they can handle a sizable fraction of outliers if they have to). Our emphasis will be on the application of these methods rather than on their theoretical properties.

0886–9383/91/010001–20$10.00

Received 31 December 1989
Accepted (revised) 22 July 1990

2. ROBUST ESTIMATION IN ONE DIMENSION

2.1. Outliers and robustness

Suppose that we have five measurements of a concentration:

$$5{\cdot}59, \quad 5{\cdot}66, \quad 5{\cdot}63, \quad 5{\cdot}57, \quad 5{\cdot}60 \tag{1}$$

and that we want to estimate its true value. For this one commonly computes the *sample average:*

$$\bar{x} = \frac{5{\cdot}59 + 5{\cdot}66 + 5{\cdot}63 + 5{\cdot}57 + 5{\cdot}60}{5} = 5{\cdot}61$$

A less well-known estimator is the *sample median.* We sort the observations from smallest to largest:

$$5{\cdot}57 \leqslant 5{\cdot}59 \leqslant 5{\cdot}60 \leqslant 5{\cdot}63 \leqslant 5{\cdot}66$$

The sample median is then the middle observation, yielding 5·60. (If the number of observations is even, then we take the average of the *two* observations in the middle.) In this example the median does not coincide with the average, but they are close to each other. The median and the average are both called location estimators because they measure the general position of the data.

Let us now suppose that one of these concentrations has been wrongly recorded, so the data become

$$5{\cdot}59, \quad 5{\cdot}66, \quad 5{\cdot}63, \quad 55{\cdot}7, \quad 5{\cdot}60 \tag{2}$$

Outliers of this kind occur frequently and may be due to copying mistakes (yielding a misplaced decimal point, or the permutation of two digits). It is even possible that the outlying observation is not incorrect but was made under exceptional circumstances (e.g. a seismic quake) or belongs to another population (e.g. it may have been the concentration of a different compound). Anyway, let us look at the effect of such an outlier on the estimate. For the average we find

$$\bar{x} = \frac{5{\cdot}59 + 5{\cdot}66 + 5{\cdot}63 + 55{\cdot}7 + 5{\cdot}60}{5} = 15{\cdot}64$$

which is utterly useless. For the median we sort the data again:

$$5{\cdot}59 \leqslant 5{\cdot}60 \leqslant 5{\cdot}63 \leqslant 5{\cdot}66 \leqslant 55{\cdot}7$$

yielding the value 5·63 which is still quite reasonable. The outlier has changed the median only slightly. We say that the median is a *robust* estimator, unlike the average which is very sensitive to outliers.

In many classical statistics courses the sample median is not even mentioned. The average is typically preferred because of several reasons, such as its ease of computation and the fact that it lends itself to elementary mathematical manipulations. Its usual justification is its optimality at Gaussian distributions, but this is a circular reasoning because Gauss actually introduced the Gaussian distribution as the best framework for the sample average! (The central limit theorem does say that the sum of many small terms tends to a Gaussian distribution, but outliers are often caused by a single large term.) Very few distributions occurring in practice are perfectly Gaussian. In the field of *robust statistics* we try to construct techniques that are not affected much by violations of the Gaussian assumption.

There are several ways to investigate how robust a procedure is. The *influence function*[1] of an estimator describes the effect of one outlier; for a detailed description of this approach see Reference 2. Another possibility is to simulate contaminated data sets and to try out how well the estimator does on them, as in Reference 3. In this paper we will restrict attention to a more simple and yet far-reaching tool, namely the *breakdown point*, which was developed by Hodges,[4] Hampel[5] and Donoho and Huber.[6] The breakdown point of an estimator is the smallest fraction of the observations that have to be replaced to make the estimator unbounded. In this definition one can choose which observations are replaced, as well as the magnitude of the outliers, in the least favourable way.

The breakdown point of the average, applied to a sample $\{x_1, x_2, \ldots, x_n\}$ of n observations, is equal to $1/n$ because it is sufficient to replace a single observation by a large value. On the other hand, the sample median possesses the best possible breakdown point, namely 50%. Indeed, we have to replace at least half of the observations by outlying values in order to be certain that the middle observation is among them. (Here the outliers are chosen in the least favourable way; that is, all on the same side of the original sample.) In fact, when fewer than half of the observations are replaced, the median stays inside the *range* of the original data values.

Of course, when considering breakdown points we do not forget about the usual criteria such as consistency of the estimator (meaning that the result becomes more and more precise when the number of observations increases). Also, any location estimator T should be *equivariant* when the data are multiplied by a constant and when a constant is added to them:

$$T(\{cx_1 + d, \ldots, cx_n + d\}) = cT(\{x_1, \ldots, x_n\}) + d \tag{3}$$

There are quite a few other location estimators satisfying these properties. Many classes exist, such as the types A, D, L, M, P, R, S and W (for a survey see Reference 2, pp. 100–116). However, in this paper we shall focus on the sample median because it is a typical robust estimator.

Next to the sample's location we often want to estimate its *scale* (or 'spread') as well. A scale estimator S should be equivariant in the sense that

$$S(\{cx_1 + d, \ldots, cx_n + d\}) = |c| S(\{x_1, \ldots, x_n\}) \tag{4}$$

where the absolute value is needed because a scale estimate is always positive. The classical scale estimator is the sample standard deviation:

$$S(\{x_1, \ldots, x_n\}) = \sqrt{\left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)} \tag{5}$$

which is, however, notoriously non-robust because it can become very large ('explosion') in the presence of even a single outlier. For our contaminated data set (2) the standard deviation becomes 22·40, whereas it was only 0·0354 for the original data (1). Therefore the breakdown point of the standard deviation is merely $1/n$. An extremely robust estimator of scale is the MAD, given by the median of all absolute distances from the sample median:

$$S = 1{\cdot}483 \underset{j=1,\ldots,n}{\text{median}} \left| x_j - \underset{i=1,\ldots,n}{\text{median}} (x_i) \right| \tag{6}$$

in which 1·483 is a correction factor to make the estimator consistent with the usual scale parameter of Gaussian distributions. Like the sample median, the MAD also has a breakdown point of 50%. The MAD of the contaminated sample (2) is the same as for the original sample (1), namely 0·0445.

When we have estimators of location and scale we can build an outlier identifier. Indeed, what is an outlier? We have to specify with respect to what it is outlying. An outlier is a value which differs from the majority of the points. That is, it lies far from the location T relative to the scale S. We therefore have to compute the standardized observations:

$$z_i = \frac{x_i - T}{S} \tag{7}$$

These z_i (which are sometimes called 'z-scores') then have to be compared to some cut-off value. If $|z_i|$ is larger than 2·5, we will identify the observation x_i. (If there are no outliers and the x_i come from a Gaussian distribution, the probability that $|z_i| > 2{\cdot}5$ is very small. The choice of the cut-off value 2·5 is to a certain extent arbitrary.) Which estimators T and S should we use? If we insert the average for T and the usual standard deviation (5) for S, then we obtain the classical 'studentized deviate', which is quite useless. For our contaminated data set (2) we obtain the following z_i:

$$-0{\cdot}45, \quad -0{\cdot}45, \quad -0{\cdot}45, \quad 1{\cdot}79, \quad -0{\cdot}45$$

none of which come even near the cut-off value. The classical z_i fail because of two reasons: first, because one subtracts a location estimate T which has moved towards the outlier; and secondly, because one divides by a scale estimate S which has exploded. Both deficiencies are easily repaired by inserting robust estimators, such as the sample median for T and the MAD of (6) for S. This robust identifier correctly tells us that there are no outliers in the original sample (1). On the other hand, for the contaminated sample (2) we obtain the z_i-values

$$-0{\cdot}90, \quad 0{\cdot}67, \quad 0{\cdot}00, \quad 1125{\cdot}17, \quad -0{\cdot}67$$

one of which exceeds the cut-off 450 times! This example is but one illustration of a more general principle, which says that *outliers can easily be identified by comparing data with a robust fit.*

2.2. Large data sets and averaging curves and images

Up to now we have assumed that we keep the data at our disposition and that we can go back to them several times. For instance, in order to compute the sample median we need to store all the values in the computer's memory so that we can sort them. However, one sometimes encounters situations where so many data are arriving in real time that they cannot be stored. In such situations one has nearly always restricted attention to the sample average, which needs very little memory space. The average can be computed with an updating mechanism, so only a single pass through the data is necessary. For instance, the following Fortran lines may be used:

```
    SUM = 0
    DO 10 I = 1,N
10  SUM = SUM + ENTER(I)
    AVERA = SUM/N
```

where ENTER is a function that reads, records, generates or otherwise accesses the ith observation. It is therefore not necessary to store the data in central memory. It is commonly thought that all robust estimators need to store at least the data themselves, thereby consuming n storage spaces, which in certain applications is unfeasible.

To remedy this problem, Rousseeuw and Bassett[7] introduced a new robust estimator which

can also be computed by means of a single-pass updating mechanism, without having to store all the observations. Let us assume that $n = b^k$ where b and k are integers. The *remedian with base b* proceeds by computing medians of groups of b observations, yielding b^{k-1} estimates on which this procedure is iterated, and so on, until only a single estimate remains. When implemented properly, this method merely needs k arrays of size b which are continuously reused. Figure 1 illustrates the remedian with base 11 and exponent 4. The data enter at the top and array 1 is filled with the first eleven observations. Then the median of these eleven observations is stored in the first element of array 2, and array 1 is used again for the second group of eleven observations, the median of which will be put in the second position of array 2. After some time array 2 is full too and its median is stored in the first position of array 3, and so on. When $11^4 = 14\,641$ data values have passed by, array 4 is complete and its median becomes the final estimate. This method used only 44 storage positions and its speed is of the same order of magnitude as that of the average or the median. A physical analogy is the mileage recorder in a car: compare the first array with the rightmost wheel that counts the individual miles, the second array with the wheel indicating tens of miles, etc. This also explains why b is called the *base*, as in the terminology of positional number systems. (We could take $b = 10$, but we prefer odd b because then the medians are easy to compute.)

The remedian with base b merely needs bk storage spaces for data sets with $n = b^k$ values. The total storage only increases as the logarithm of n because $bk = b \log_b(n)$. (When n is not a power of b, Rousseeuw and Bassett compute a weighted median at the last step, which does not need more storage.) The following lines implement the remedian of Figure 1:

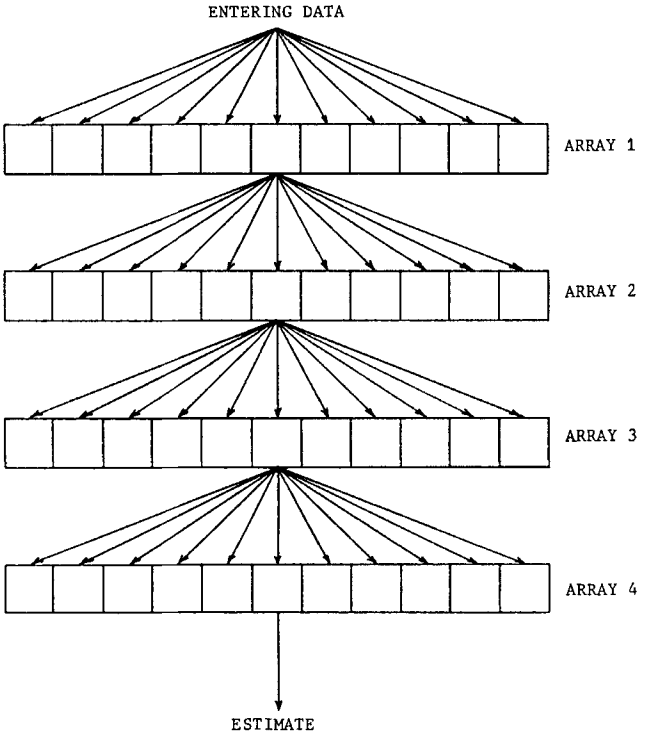

Figure 1. Mechanism of the remedian with base 11 and exponent 4, using 44 storage spaces for a data set of size $n = 11^4 = 14\,641$

```
CC    A PROGRAM FOR THE REMEDIAN
CC    ----------------------
      DIMENSION A1(11),A2(11),A3(11),A4(11)
      DO 40 M=1,11
      DO 30 L=1,11
      DO 20 K=1,11
      DO 10 J=1,11
      I=I+1
   10 A1(J)=ENTER(I)
   20 A2(K)=FMED(A1)
   30 A3(L)=FMED(A2)
   40 A4(M)=FMED(A3)
      REMED=FMED(A4)
      WRITE(*,*)REMED
      STOP
      END
```

where FMED is a function which returns the median of an array of eleven numbers.

The remedian transforms properly when all observations x_i are replaced by $cx_i + d$. Like the sample median, it is even equivariant with respect to any *monotone* transformation of the x_i, such as a power function or an exponential. Rousseeuw and Bassett showed that the remedian is a consistent estimator of the underlying population median, investigated its sampling distribution and computed its breakdown point. Alternative approaches are due to Martin and Masreliez,[8] Pearl,[9] Tierney[10] and Tukey.[11]

An important application of the remedian is to curve averaging. Suppose we want to obtain a certain curve corresponding to a physical phenomenon. A curve can be registered by means of a list of its function values $x(t)$ at equally spaced arguments t (usually t stands for time). Unfortunately, the observed values of $x(t)$ are subject to noise of various sources. Therefore one repeats the experiment several times, yielding n curves in all, so the data are of the form

$$\{x_i(t);\ t = 1, \ldots, M\} \quad \text{for} \quad i = 1, \ldots n \tag{8}$$

One wants to combine the n curves to estimate the true underlying shape. The classical approach is *averaging*, by which one computes the curve

$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t), \quad t = 1, \ldots, M \tag{9}$$

In the case of Gaussian noise and no outliers, averaging makes good sense because then the noise goes down for large n. The averaging technique is built into many special-purpose recorders (e.g. in hospitals).

Usually M and n are quite large, making it impractical to store all the observed curves in central memory. This precludes calculation of the 'median curve'

$$\underset{i=1,\ldots,n}{\text{median}}\ x_i(t), \quad t = 1, \ldots, M \tag{10}$$

as well as many other robust summaries. We propose to compute the remedian curve

$$\underset{i=1,\ldots,n}{\text{remedian}}\ x_i(t), \quad t = 1, \ldots, M \tag{11}$$

instead, because it is a robust single-pass method. The above Fortran program can be adapted quite easily to produce the remedian curve, by replacing the arrays A1, A2, A3, and A4 of

length 11 by matrices with eleven rows and M columns. For $b = 11$ and $k = 4$ the total storage becomes $44M$, whereas the plain median would have needed $14\,641M$ positions.

Let us consider an example. The *electroretinagram* (ERG) is used in ophthalmology to examine disorders of the visual system. When the patient's eye is exposed to a white flash of light, it develops an electric potential that may be recorded by a contact lens electrode. The ERG curve shows the evolution of the evoked potential (expressed in microvolts) as a function of the time (in milliseconds) elapsed after the flash. The bottom curve in Figure 2(b) is a typical ERG of a healthy patient (from Reference 12). The important features are the four peaks (denoted by a, b, OP_1, and OP_2), in particular their t-co-ordinates, which are used for medical diagnosis.

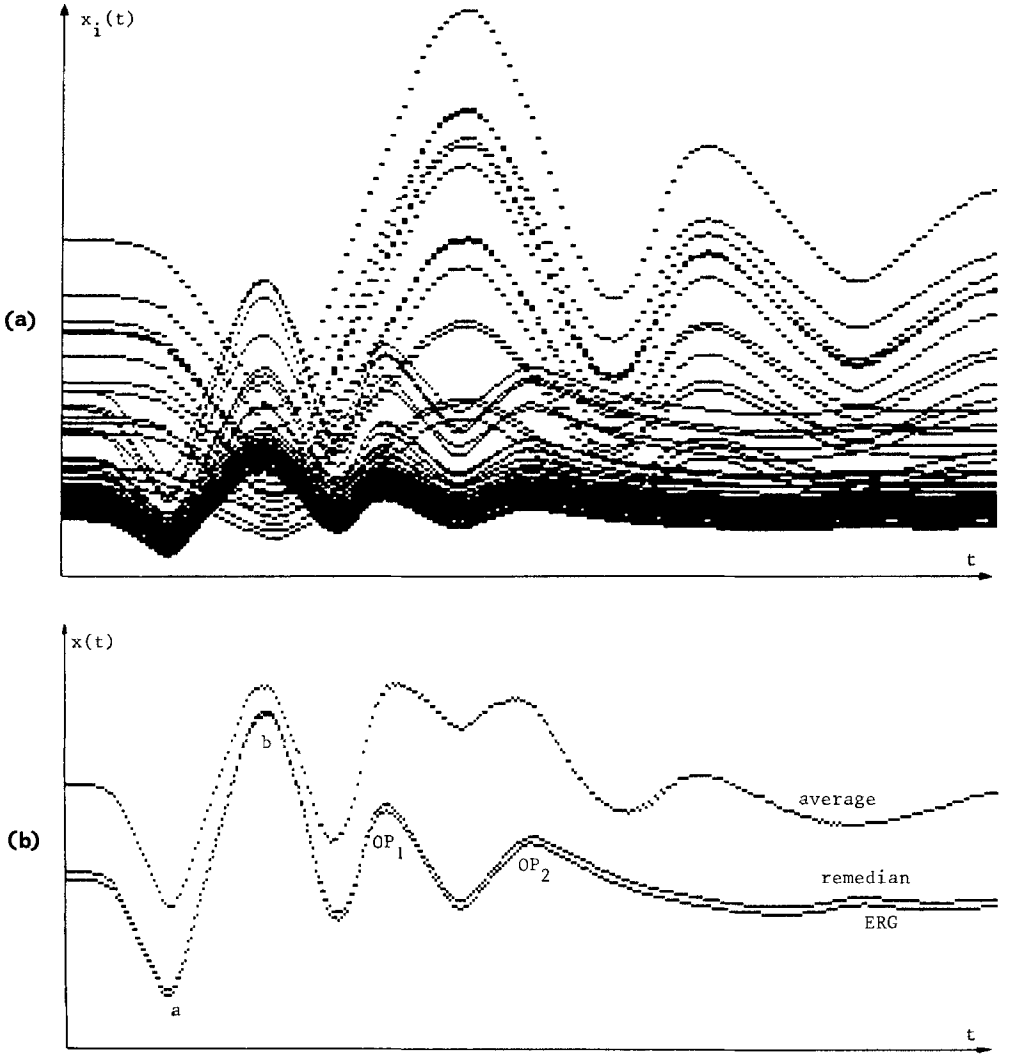

Figure 2. (a) Bundle of simulated electroretinagrams (ERG), some with Gaussian noise and others with various kinds of contamination. (b) Plot with typical ERG (bottom curve), the average of the simulated ERGs (upper curve) and their remedian (middle curve)

When the ERG curve is recorded only once, the noise typically dominates the signal so that no peak can be found. The current solution is to record many curves by repeating the stimulus flash light and then to average them. However, the average curve is often deformed and difficult to interpret owing to a high amount of contamination caused by electrical interference, involuntary eye movements and other artefacts.

It is quite feasible to replace the averaging routine in the recording instrument by the remedian, because the latter is equally fast and needs little storage. To verify if this replacement is worthwhile, computer simulations were performed in which both the average and the remedian were calculated for a bundle of curves, some of which were contaminated. The basic curve was the ERG of Figure 2(b) measured at $M = 320$ time units. Figure 2(a) contains $n = 81$ curves (in ophthalmology more curves are used, but this would make the display overcrowded). The curves were generated as follows: with probability 0·7, curve i is the basic ERG plus some Gaussian noise with modest scale. With probability 0·1 the $x_i(t)$-values are multiplied by a random factor greater than unity. With probability 0·2 the curve models a response at half the usual speed, again with magnified $x_i(t)$-values.

The upper curve in Figure 2(b) is the average of the ERG curves in Figure 2(a). It has been greatly affected by the contamination, which caused a substantial upward shift. What is worse, the average has one peak too many, rendering medical diagnosis difficult. Averaging often produces results like this in actual clinical practice. On the other hand, the 3^4 remedian lies very near to the original ERG and is virtually undamaged by the contamination.

Many other applications of robust curve averaging could be envisaged, for instance in spectroscopy. Median-type procedures can also be used to estimate horizontal shifts between spectrograms.[13,14]

Averaging also occurs in image analysis. An image may be described as a rectangular grid of pixels, each with a corresponding number $x(r, c)$ indicating its grey intensity. When n images are observed one after another, the data are

$$\{x_i(r, c);\ r = 1, \ldots, R \quad \text{and} \quad c = 1, \ldots, C\} \quad \text{for} \quad i = 1, \ldots, n \tag{12}$$

where R is the number of rows and C the number of columns. In one application, images of a crystallographic lattice were recorded by means of an electron microscope, with $R = 512$, $C = 512$ and $n \approx 10\,000$. Usually such images are averaged to obtain a sharp result, but in this case averaging did not work well because in many images a part of the lattice was contaminated or even destroyed by the radiation of the microscope itself. Computing plain medians was not feasible because there were $nRC \approx 2\,621\,440\,000$ data values in all, which could not be stored in central memory. One can, however, compute the remedian image given by

$$\underset{i=1,\ldots,n}{\text{remedian}}\ x_i(r, c) \quad \text{for} \quad r = 1, \ldots, R \quad \text{and} \quad c = 1, \ldots, C \tag{13}$$

The computation of remedian curves (or images) may be speeded up if one has access to parallel computing facilities, because one can let each processor work on a different portion of the curve.

3. ROBUST REGRESSION

3.1. Simple regression

In simple regression one assumes a relation of the type

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{14}$$

in which x_i is called the *explanatory variable* or *regressor* and y_i is the *response variable.* The intercept β_0 and the slope β_1 have to be estimated from the data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Classical theory assumes the ε_i to be Gaussian.

In simple regression the observations (x_i, y_i) are two-dimensional, so they can be plotted. Regression users should always draw this plot first because it shows whether the data are roughly linear, and any unusual structures will be clearly visible. In this section we begin with the simple regression situation in which the phenomena are most easily visualized, after which we continue with multiple regression for which robust methods are much more necessary.

Applying a regression estimator to such a bivariate data set with n observations yields the *regression coefficients* $\hat{\beta}_0$ and $\hat{\beta}_1$. Although the true parameters β_0 and β_1 are unknown, one can insert these $\hat{\beta}_0$ and $\hat{\beta}_1$ in (14) to yield

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{15}$$

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

$$r_i = y_i - \hat{y}_i \tag{16}$$

The most popular regression estimator (dating back to Gauss and Legendre, around 1800) is the *least squares method* (LS) given by

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{minimize}} \sum_{i=1}^{n} r_i^2 \tag{17}$$

The basic idea was to make all the residuals very small. Gauss preferred the LS criterion to other objective functions because in this way the regression coefficients could be computed *explicitly* from the data. Afterwards, Gauss introduced the Gaussian distribution as the distribution for which LS is optimal. Since then, the LS method has been theoretically justified in many other ways (e.g. the Gauss–Markov theorem) that are equally circular.

More recently, people began to realize that actual data often do not satisfy the Gaussian assumption, with dramatic effects on the LS results. Let us look at some plots illustrating the effect of outliers. Figure 3(a) is the scatterplot of five points, $(x_1, y_1), \ldots, (x_5, y_5)$, which almost

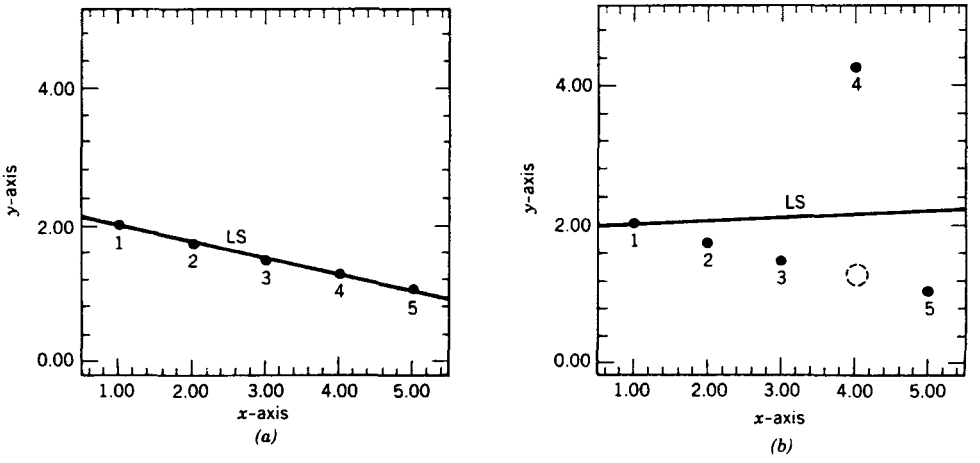

Figure 3. (a) Five points and their least squares regression line, (b) Same data with one outlier in the y-direction

lie on a straight line. In such a situation the LS solution fits the data very well, as can be seen from the LS line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ in the plot. However, suppose that someone gets a wrong value of y_4 because of a recording or copying mistake. Then (x_4, y_4) may be rather far away from the 'ideal' line. Figure 3(b) displays such a situation, where the fourth point has moved up from its original position (indicated by the dashed circle). This point is called an *outlier in the y-direction* and has a rather large effect on the LS line in Figure 3(b). Such vertical outliers have received most attention in the literature, because in designed experiments the values of x_i are preselected and then one only expects outliers in the y_i.

In observational studies ('happenstance data') the x_i are not fixed but are themselves observed quantities subject to random variability. In that case, outliers can also occur in the x_i. (And note that even in designed experiments one may have data entry mistakes in the x_i!) For the effect of such an outlier we turn to Figure 4. In Figure 4(a) we again see five points with a well-fitting LS line. If we now record x_1 wrongly, we obtain Figure 4(b). The resulting point is called an *outlier in the x-direction* and its effect on LS is very large because it actually tilts the LS line. Therefore the point (x_1, y_1) 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 4(a)) becomes a very large (negative) value, contributing very much to Σr_i^2 for that line. Therefore the original line cannot be selected from an LS perspective, and indeed the line of Figure 4(b) possesses the smallest Σr_i^2 because it has tilted to reduce that large r_1^2, even if the other 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 from the bulk of the observed x_i in the data. Note that this does not take y_k into account. In Figure 5 the point (x_k, y_k) lies close to the linear pattern set by the majority of the data, so it can be considered a 'good' leverage point. On the other hand, the point (x_4, y_4) in Figure 4(b) is a bad leverage point. Therefore, to say that an observation (x_k, y_k) is a leverage point refers only to its *potential* for influencing the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ due to its outlying component x_k.

When a point (x_i, y_i) violates the linear relation of the majority, we will call it a *regression*

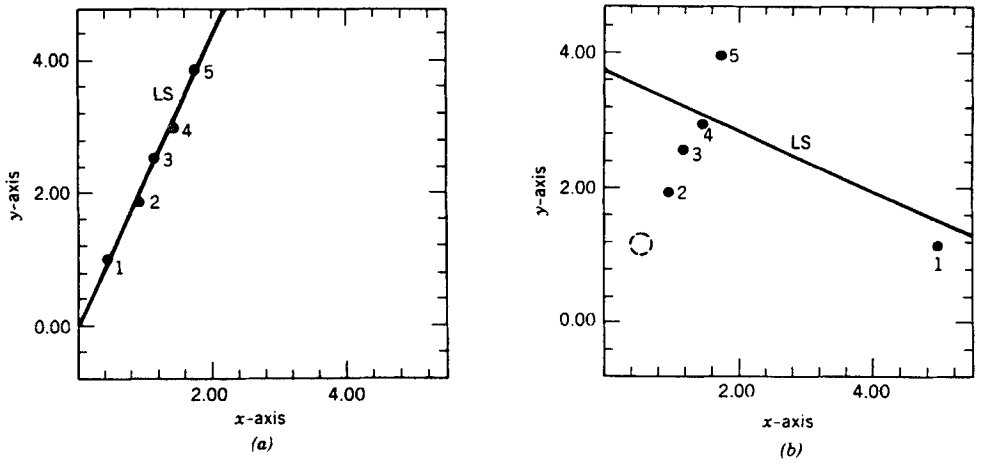

Figure 4. (a) Five points with their least squares regression line. (b) Same data with one outlier in the x-direction

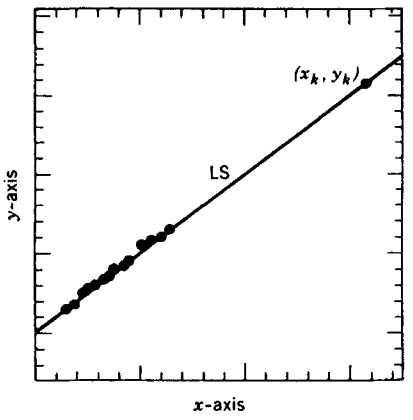

Figure 5. The point (x_k, y_k) is a leverage point because x_k is outlying. However, (x_k, y_k) is not a regression outlier because it matches the linear pattern set by the other data points.

outlier, taking into account both x_i and y_i simultaneously. In other words, a regression outlier is either a vertical outlier or a bad leverage point.

It is often thought that regression outliers can be identified by looking at the LS residuals. Unfortunately, things are not that simple. For example, consider again Figure 4(b). Case 1, being a bad 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 small (negative) number. The residuals r_2 and r_5 are much larger, although they correspond to good points. If one were to apply a rule such as 'delete the points with largest LS residuals', then the good points would have to be deleted first! Of course, in simple regression there is really no problem at all because one can actually look at the data, but we shall see that in multiple regression the outliers often remain invisible in spite of a careful inspection of LS residuals.

From the examples in Figures 3 and 4 we know that even a single regression outlier can totally offset the LS estimator (provided it is far away). This implies that the breakdown point of the LS method is merely $1/n$ (which is not so surprising because the LS estimator generalizes the sample average we studied above). A first step toward a more robust regression estimator came from Edgeworth,[15] who argued that outliers have a very large effect on LS because the residuals r_i are being squared in (17). Therefore he proposed the *least absolute values* method given by

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{minimize}} \sum_{i=1}^{n} |r_i| \tag{18}$$

This technique is often referred to as L_1 regression, whereas least squares is called L_2. (For recent surveys on L_1 see Reference 16.) Unfortunately, L_1 is only robust with respect to vertical outliers, but it does not protect against bad leverage points. Indeed, in the example of Figure 4(b) the leverage point attracts the L_1 line as much as the LS line. Therefore the breakdown point of the L_1 method is still no better than $1/n$.

There exist many other techniques, with varying degree of robustness. To avoid a digression at this stage, we will concentrate on one approach first and leave descriptions of other methods to Section 3.3 below. We want an estimator with a high breakdown point, which can be used to analyse messy data sets as well as clean ones. Let us look again at (17). A more natural name for the LS method (in fact, the name you would expect when seeing its formula for the first time) is *least sum of squares*. Apparently, few people have objected to the deletion of the word

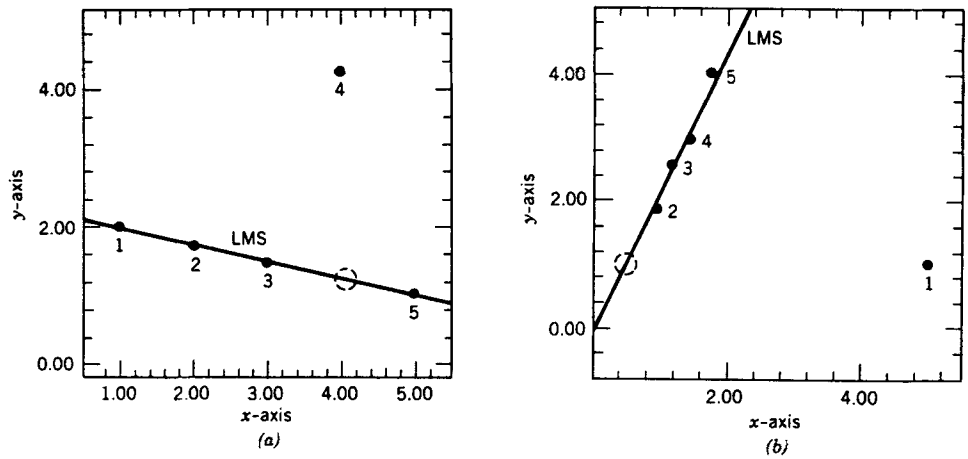

Figure 6. Robustness of LMS regression with respect to (a) an outlier in the y-direction and (b) an outlier in the x-direction

'sum' – as if the only sensible thing to do with n numbers would be add them. However, adding the r_i^2 is equivalent to using their average (dividing by n does not change the minimization) and we have seen that the average is not robust. Why not replace the sum by a median, which is very robust? This yields the *least median of squares* (LMS) method proposed by Rousseuw:[17]

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\text{minimize}} \ \underset{i=1,\ldots,n}{\text{median}} \ r_i^2 \tag{19}$$

It turns out that this estimator is very robust with respect to outliers in y as well as outliers in x. In Figure 6 we see that the LMS yields the desired fit for the examples of Figures 3(b) and 4(b). Its breakdown point is 50%, the highest possible value. This means that the LMS can cope with several outliers at the same time, in the sense that the result will still be trustworthy. Its basic principle 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 fourth case possesses a considerable LMS residual, and that of case 1 in Figure 6(b) is even more apparent.

The LMS line has an intuitive geometric interpretation because it lies at the centre of the narrowest strip covering half of the points. Note that this interpretation is simpler than that of least squares! It also provides some insight as to why the LMS is not much attracted by outliers. Using this geometric interpretation, it is possible to construct an algorithm for the LMS line, which is described in Chap. 5 of Reference 18. That book is devoted to LMS and other high-breakdown methods. It is also a user manual for the Fortran program PROGRESS (Program for RObust reGRESSion) which computes LMS regression on IBM-PCs and other machines.

Let us look at an example from astronomy. The data in Table 1 form the Hertzsprung–Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. Here x_i is the logarithmic temperature at the star surface as determined by spectroscopy and y_i is the logarithm of its light intensity. The data originally appeared in Reference 19.

The Hertzsprung–Russell diagram itself is shown in Figure 7. It is the scatterplot of these

points, where the log temperature is plotted from right to left. In the plot we see two groups of points: the majority, which seem to follow a steep band; and the four stars in the upper right corner. These parts in the diagram are well known in astronomy. The 43 stars are said to belong to the main sequence, whereas the four remaining stars are called giants. (The giants are the points with indices 11, 20, 30 and 34.)

Applying the LMS estimator to these data yields the solid line $\hat{y} = 3{\cdot}898x - 12{\cdot}298$, which fits the main sequence nicely. On the other hand, the LS solution $\hat{y} = -0{\cdot}409x + 6{\cdot}78$ corresponds to the dashed line in Figure 7, which has been pulled away by the four giant stars

Table 1. Simple regression data: the Hertzsprung–Russell diagram of the star cluster CYG OB1

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	4·37	5·23	17	4·23	3·94	33	4·45	5·22
2	4·56	5·74	18	4·42	4·18	34	3·49	6·29
3	4·26	4·93	19	4·23	4·18	35	4·23	4·34
4	4·56	5·74	20	3·49	5·89	36	4·62	5·62
5	4·30	5·19	21	4·29	4·38	37	4·53	5·10
6	4·46	5·46	22	4·29	4·22	38	4·45	5·22
7	3·84	4·65	23	4·42	4·42	39	4·53	5·18
8	4·57	5·27	24	4·49	4·85	40	4·43	5·57
9	4·26	5·57	25	4·38	5·02	41	4·38	4·62
10	4·37	5·12	26	4·42	4·66	42	4·45	5·06
11	3·49	5·73	27	4·29	4·66	43	4·50	5·34
12	4·43	5·45	28	4·38	4·90	44	4·45	5·34
13	4·48	5·42	29	4·22	4·39	45	4·55	5·54
14	4·01	4·05	30	3·48	6·05	46	4·45	4·98
15	4·29	4·26	31	4·38	4·42	47	4·42	4·50
16	4·42	4·58	32	4·56	5·10			

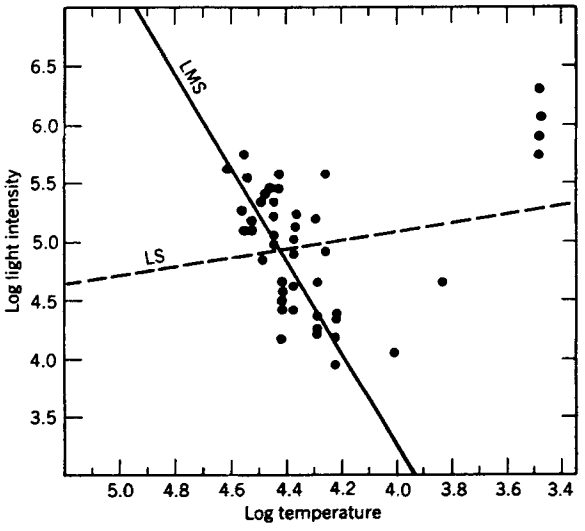

Figure 7. Hertzsprung–Russell diagram of the star cluster CYG OB1 with the LS fit (dashed line) and the LMS fit (solid line)

(which it does not fit well either). These are outliers but not mistakes. It would be more appropriate to say that the data come from two different populations. The two groups can easily be distinguished on the basis of their 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.

Several other examples from different fields can be found in Reference 18. We also computed LMS regression lines in analytical chemistry examples.[20]

3.2. Multiple regression

By extending the simple regression framework (14) to several explanatory variables, we obtain the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{20}$$

Each data point is of the form $(x_{i1}, \ldots, x_{ip}, y_i)$, containing p regressors and one response variable. The classical LS estimator of $\beta_0, \beta_1, \ldots, \beta_p$ corresponds to

$$\underset{\hat{\beta}_0, \ldots, \hat{\beta}_p}{\text{minimize}} \sum_{i=1}^{n} r_i^2 \tag{21}$$

in which the residuals r_i are given by

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} \tag{22}$$

As before, the LS estimator is very vulnerable to outliers in the response variable and to outliers in the regressors. Our main handicap is that the observations $(x_{i1}, \ldots, x_{ip}, y_i)$ have $p + 1$ dimensions so they cannot be plotted, and hence we cannot spot the outliers by eye. The worst problem is with observational studies, because they may contain leverage points; that is, points for which $(x_{k1}, \ldots, x_{kp})$ lies far from the bulk of the $(x_{i1}, \ldots, x_{ip})$ in the data. This is a major concern because

(1) least squares is affected more by leverage points than by vertical outliers, as was already seen in Figure 4(b)
(2) there are p regressors as opposed to only one response variable, so leverage points are more likely to occur than vertical outliers
(3) it becomes very difficult to *identify* leverage points owing to the higher dimensionality.

An illustration of the third aspect is given in Figure 8, which plots x_{i2} versus x_{i1} 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 $\{x_{12}, x_{22}, \ldots, x_{n2}\}$.) When there are more than two regressors, even more complicated configurations are possible. In general, it is not sufficient to look at each variable separately or at scatterplots of pairs of variables.

We are mostly concerned with regression outliers; that is, cases for which $(x_{i1}, \ldots, x_{ip}, y_i)$ violates the linear relation followed by the majority of the data, taking into account the response variable as well as the regressors. Already in simple regression we saw that the outliers cannot always be identified by means of their LS residuals. In order to identify the outliers, we need to know what the bulk of the data are like. This calls for a high-breakdown estimator such as LMS, which is defined analogously by[17]

$$\underset{\hat{\beta}_0, \ldots, \hat{\beta}_p}{\text{minimize}} \ \underset{i=1, \ldots, n}{\text{median}} \ r_i^2 \tag{23}$$

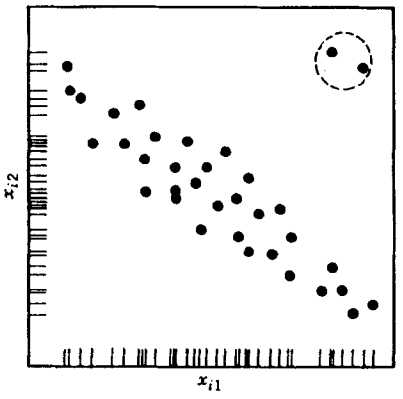

Figure 8. Plot of the explanatory variables x_{i1} and x_{i2} of a regression data set. There are two leverage points (indicated by the dashed circle) which are not outlying in either of the co-ordinates

The LMS has a breakdown point of 50% even in multiple regression; only its computation time goes up. Rousseeuw[21] also proposed the *least trimmed squares* (LTS) estimator given by

$$\underset{\beta_0, \ldots, \beta_p}{\text{minimize}} \sum_{i=1}^{h} r^2_{(i)} \tag{24}$$

where $r^2_{(1)} \leqslant r^2_{(2)} \leqslant \cdots \leqslant r^2_{(n)}$ are the ordered squared residuals (note that the residuals are first squared and then ordered) and $h \approx n/2$. The LTS has the same high breakdown point.

In order to decide whether a residual from a robust regression is unusually large, we have to compare it to something. For this purpose we need to estimate the spread of the ε_i. In classical theory the ε_i have a Gaussian distribution with an unknown scale parameter denoted by σ. When using LS, σ is estimated by

$$\hat{\sigma} = \sqrt{\left(\frac{1}{n-p-1} \sum_{i=1}^{n} r_i^2\right)} \tag{25}$$

where the r_i are the LS residuals. For LMS regression the corresponding scale estimate becomes

$$\hat{\sigma} = 1 \cdot 483 \sqrt{\left(\underset{i=1,\ldots,n}{\text{median}}\ r_i^2\right)} \tag{26}$$

where the r_i are the residuals from the LMS fit and $1 \cdot 483$ is the appropriate constant to make $\hat{\sigma}$ a consistent estimator of σ. The LMS scale estimate has itself a 50% breakdown point, whereas the LS scale has 0% breakdown because it explodes easily.

By analogy to the one-dimensional z-scores (7), we consider the *standardized residuals*

$$\frac{r_i}{\hat{\sigma}} \tag{27}$$

If $|r_i/\hat{\sigma}|$ is larger than $2 \cdot 5$, we will consider the ith case to be a regression outlier. Inserting the LS residuals for r_i and the LS scale for $\hat{\sigma}$ makes no sense, because leverage points often have small LS residuals, and even when some LS residuals are large, they will blow up $\hat{\sigma}$ and hence be divided by a large denominator in (27). As in one dimension, outliers can be identified by inserting robust estimators, in this case by considering the residuals from the LMS fit and dividing them by the LMS scale.

Let us look at a real data example to illustrate the need for a robust regression method. The

Table 2. Stackloss data with standardized residuals

					Standarized residual	
i	x_{i1}	x_{i2}	x_{i3}	y_i	LS	LMS
1	80	27	89	42	1·00	7·70
2	80	27	88	37	−0·59	3·74
3	75	25	90	37	1·40	7·14
4	62	24	87	28	1·76	7·64
5	62	22	87	18	−0·53	0·28
6	62	23	87	18	−0·93	0·00
7	62	24	93	19	−0·74	0·51
8	62	24	93	20	−0·43	1·30
9	58	23	87	15	−0·97	−0·11
10	58	18	80	14	0·39	0·51
11	58	18	89	14	0·81	0·51
12	58	17	88	13	0·86	0·00
13	58	18	82	11	−0·44	−1·87
14	58	19	93	12	−0·02	−1·36
15	50	18	89	8	0·73	0·28
16	50	18	86	7	0·28	−0·51
17	50	19	72	8	−0·47	0·00
18	50	19	79	8	−0·14	0·00
19	50	20	80	9	−0·18	0·51
20	56	20	82	15	0·44	1·87
21	70	20	91	15	−2·23	−6·06

well-known stackloss data of Brownlee[22] describe the operation of a chemical plant for the oxidation of ammonia to nitric acid. The data consist of 21 four-dimensional observations, listed in Table 2. The stackloss (y) has to be explained by the rate of operation (x_1), the cooling water inlet temperature (x_2) and the acid concentration (x_3). Applying LS regression yields the equation

$$\hat{y} = 0{\cdot}716x_1 + 1{\cdot}295x_2 - 0{\cdot}152x_3 - 39{\cdot}9$$

with corresponding scale estimate $\hat{\sigma} = 3{\cdot}24$.

We cannot plot the data because they are four-dimensional, but we can look at the standardized residuals (27). The sixth column of Table 2 lists the standardized residuals obtained from LS. From these results one would conclude that the data set does not contain any outliers, because all the standardized residuals are below 2·5. However, let us now consider the LMS fit

$$\hat{y} = 0{\cdot}714x_1 + 0{\cdot}357x_2 + 0{\cdot}000x_3 - 34{\cdot}5$$

The equation itself is quite different, especially the coefficients of x_2 and x_3. The LMS scale estimate becomes 1·26, which is much smaller than the LS scale. As to the search for outliers, we now look at the standardized residuals based on the LMS, which are given in the last column of the table. They indicate that the observations 1, 3, 4 and 21 are outlying, because their $|r_i/\hat{\sigma}|$ is much larger than 2·5. Observation 2 is a borderline case. These conclusions correspond to the findings cited in the literature. Our robust regression technique has analysed these data in a single blow, which should be contrasted to some of the earlier analyses of the same data set, which were long and laborious.

This example illustrates the danger of merely looking at the LS residuals. We have seen that it is better to use a robust technique to identify the outliers, which may then be thoroughly investigated and perhaps corrected (if one has access to the original measurements) or deleted. Another possibility is to change the model, e.g. by adding squared regressors and/or transforming the response variable.

The LMS regression is more stable than the LS fit. Indeed, if we carry out LS on the 'cleaned' data set without the observations 1, 2, 3, 4 and 21, then we find an equation and a scale estimate that are close to the LMS. Such a 'reweighted' analysis (where each point has a weight of zero or one, depending on whether or not LMS has classified it as an outlier) is useful because it is still robust and at the same time yields all the customary output, such as t-values, confidence intervals and a coefficient of determination. For this reason the program PROGRESS described by Rousseeuw and Leroy[18] computes the LS fit, the LMS fit and the reweighted analysis. In the same book many examples are treated.

It should be stressed that the LMS does *not* 'throw away' 50% of the data. Rather, it finds a fit to the majority of the points, which can then be used to identify the actual outliers (of which there may be many, a few, or none at all). Also, note that we did not need any symmetry assumption anywhere.

3.3. Alternative approaches

Many other robust regression estimators have been proposed. Huber[23] introduced *M-estimators* for regression. They replace the r_i^2 in (21) by another function of the residuals, yielding

$$\underset{\hat{\beta}_0, \ldots, \hat{\beta}_p}{\text{minimize}} \sum_{i=1}^{n} \rho(r_i)$$

where ρ is an even function (i.e. $\rho(-t) = \rho(t)$ for all t) with a unique minimum at zero. M-estimators can be computed by means of iteratively reweighted LS or with Newton–Raphson-type algorithms. Unfortunately, their breakdown point is again $1/n$ because of the effect of outlying $(x_{11}, \ldots, x_{ip})$. Because of this vulnerability to leverage points, *generalized M-estimators* (GM-estimators) were introduced, with the basic purpose of bounding the influence of outlying $(x_{i1}, \ldots, x_{ip})$ by means of some weight function. For this reason they are often called bounded influence estimators. Particular types of GM-estimators were studied by Mallows[24] and Schweppe (see Reference 25); for a recent survey see Chap. 6 of Reference 2. It turns out, however, that the breakdown point of all GM-estimators decreases when the dimension increases. This is unsatisfactory, because it means that the breakdown point diminishes when there are more regressors and hence more opportunities for outliers to occur.

Various other estimators have been proposed, such as that of Brown and Mood,[26] the median of pairwise slopes,[27] R-estimators,[28,29] L-estimators[30,31] and the technique of Andrews.[32] Unfortunately, even in simple regression none of these methods achieves a breakdown point of 30%.

The first robust regression method with a 50% breakdown point was the *repeated median* proposed by Siegel.[33] For any $p+1$ observations with indices $\{i_1, \ldots, i_{p+1}\}$, he computes the coefficients $\beta_0(i_1, \ldots, i_{p+1}), \ldots, \beta_p(i_1, \ldots, i_{p+1})$ such that the corresponding surface fits these $p+1$ points exactly. The jth coefficient of the repeated median regression is then defined as

$$\hat{\beta}_j = \underset{i_1}{\text{median}}\left(\underset{i_2}{\text{median}}\left(\ldots\left(\underset{i_{p+1}}{\text{median}}\ \beta_j(i_1, \ldots, i_{p+1})\right)\ldots\right)\right)$$

where the outermost median is over all choices of i_1, the next is over all choices of $i_2 \neq i_1$, and so on. This estimator can be computed explicitly but requires consideration of all subsets of $p+1$ points, which may cost a lot of time. It has been successfully applied to problems with small p. However, unlike other regression estimators, the repeated median is not equivariant for linear transformations of the $(x_1, \ldots, x_p)$, which is due to its co-ordinatewise construction.

The LMS method combines equivariance with a 50% breakdown point but has a low asymptotic efficiency (which does not, however, detract from its ability to identify outliers). The LTS estimator of Rousseeuw[21] has a better asymptotic efficiency and is also equivariant, but needs somewhat more computation time than the LMS. The LTS objective (24) is similar to LS, the only difference being that the largest residuals are not used in the sum, which allows the fit to stay away from the outliers. The best robustness properties are achieved when h is approximately $n/2$, and then the breakdown point attains 50%. Another variant is the class of *S-estimators* introduced by Rousseeuw and Yohai,[19] which are also equivariant and have 50% breakdown point, while sharing some of the nice mathematical properties of Huber's M-estimators.

Another approach to the identification of aberrant points is the construction of *outlier diagnostics*. These are quantities computed from the data with the purpose of pinpointing influential points, after which these outliers are to be removed or corrected, followed by an LS analysis on the remaining cases. The idea behind these diagnostics is to look at the effect of deleting one point at a time. However, it is much more difficult to diagnose outliers when there are several of them, owing to the so-called 'masking effect' which says that one outlier may mask another. The naive extensions of classical diagnostics to such multiple outliers often give rise to extensive computations (e.g. the consideration of *all* subsets of points is an impossible task). Recent work by Rousseeuw and van Zomeren[34] indicates that one needs to use robust methods in one way or another to safely identify multiple outliers. This is because one needs to know with respect to which pattern the points are outlying.

Sometimes robust methods are confused with *non-parametric* methods. Basically, there is no relation between them. The confusion stems from the coincidence that for one-dimensional data certain non-parametric methods (e.g. the Wilcoxon rank test) also happen to be relatively robust. In regression, however, they are quite different, because non-parametric regression does *not* assume a linear model of the type (20). Therefore it does not yield an explicit equation to describe the fit. Its main purpose is to compute interpolated $\hat{y}$-values corresponding to unobserved values of the regressors $x_1, \ldots, x_p$, which is achieved by local smoothing of the data. Until now, little has been done to construct non-parametric regression methods that are also insensitive to outliers.

3.4. Robust methods for other situations

Robust methods are not only useful for estimation in one-dimensional samples or when fitting a linear regression model. The books by Hampel *et al.*[2] and Rousseeuw and Leroy[18] also cover robust tests, robust multivariate location and covariance matrices, the problem of unsuspected serial correlations in supposedly independent data, robustness in time series and robust estimation for circular data. Other topics are being studied intensively by several researchers, such as robust analysis of variance, robust non-linear regression and robust cluster analysis, to name a few. Whereas some techniques have reached the stage where they can readily be applied, research still goes on in other directions.

REFERENCES

1. F. R. Hampel, 'The influence curve and its role in robust estimation', *J. Am. Stat. Assoc.* **69**, 383–393 (1974).
2. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: the Approach Based on Influence Functions*, Wiley, New York (1986).
3. D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, NJ (1972).
4. J. L. Hodges Jr., 'Efficiency in normal samples and tolerance of extreme values for some estimates of location', in *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, Vol. 1, pp. 163–168, University of California Press, Berkeley and Los Angeles, CA (1967).
5. F. R. Hampel, 'A general qualitative definition of robustness', *Ann. Math. Stat.* **42**, 1887–1896 (1971).
6. D. L. Donoho and P. J. Huber, 'The notion of breakdown point', in *A Festschrift for Erich Lehmann*, ed. by P. Bickel, K. Doksum and J. L. Hodges Jr., pp 157–184 Wadsworth, Belmont, CA (1983).
7. P. J. Rousseeuw and G. W. Bassett Jr., 'The remedian: a robust averaging method for large data sets', *J. Am. Stat. Assoc.* **85**, 97–104 (1990).
8. R. D. Martin and C. J. Masreliez, 'Robust estimation via stochastic approximation', *IEEE Trans. Information Theory*, **IT-21**, 263–271 (1975).
9. J. Pearl, 'A space-efficient on-line method of computing quantile estimates', *J. Algorithms*, **2**, 164–177 (1981).
10. L. Tierney, 'A space-efficient recursive procedure for estimating a quantile of an unknown distribution', *SIAM J. Sci. Stat. Comput.* **4**, 706–711 (1983).
11. J. W. Tukey, 'The ninther, a technique for low-effort robust (resistant) location in large samples', in *Contributions to Survey Sampling and Applied Statistics in Honor of H. O. Hartley*, ed. by H. A. David, pp. 251–257, Academic Press, New York (1978).
12. R. Trau, P. Salu, K. Wisnia, L. Kaufman, P. J. Rousseeuw and A. Pierreux, 'Simultaneous ERG–VER recording: statistical study', *Bull. Belg. Ophthalmol. Soc.* **206**, 61–67 (1983).
13. P. J. Rousseeuw, 'An application of L_1 to astronomy', in *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, ed. by Y. Dodge, pp. 405–416, North-Holland, Amsterdam (1987).
14. C. de Loore, P. Monderen and P. J. Rousseeuw, 'A new statistical method to derive radial velocity shifts from stellar spectra', *Astron. Astrophys.* **178**, 307–309 (1987).
15. F. Y. Edgeworth, 'On observations relating to several quantities', *Hermathena*, **6**, 279–285 (1887).
16. Y. Dodge, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, North-Holland, Amsterdam (1987).
17. P. J. Rousseeuw, 'Least median of squares regression', *J. Am. Stat. Assoc.* **79**, 871–880 (1984).
18. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley–Interscience, New York (1987).
19. P. J. Rousseeuw and V. Yohai, 'Robust regression by means of *S*-estimators', in *Robust and Nonlinear Time Series Analysis*, ed. by J. Franke, W. Härdle and R. D. Martin, pp. 256–272, *Lecture Notes in Statistics No. 26*, Springer, New York (1984).
20. D. Massart, L. Kaufman, P. J. Rousseeuw and A. M. Leroy, 'Least median of squares: a robust method for outlier and model error detection in regression and calibration', *Anal. Chim. Acta*, **187**, 171–179 (1986).
21. P. J. Rousseeuw, 'Regression techniques with high breakdown point', *IMS Bull.* **12**, 155 (1983).
22. K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, 2nd edn, Wiley, New York (1965).
23. P. J. Huber, 'Robust regression: asymptotics, conjectures and Monte Carlo', *Ann. Stat.* **1**, 799–821 (1973).
24. C. L. Mallows, 'On some topics in robustness', *Technical Memorandum*, Bell Telephone Laboratories, Murray Hill, NJ (1975).
25. R. W. Hill, 'Robust regression when there are outliers in the carriers', *Ph.D. Thesis*, Harvard University, Cambridge, MA (1977).
26. G. W. Brown and A. M. Mood, 'On median tests for linear hypotheses', in *Proc. 2nd Berkeley Symp. on Mathematical Statistics and Probability*, pp. 159–166, University of California Press, Berkeley and Los Angeles, CA (1951).

27. H. Theil, 'A rank-invariant method of linear and polynomial regression analysis (Parts 1–3)', *Ned. Akad. Wetenschappen Proc., Ser. A*, **53**, 386–392, 521–525, 1397–1412 (1950).
28. J. Jureckovà, 'Nonparametric estimate of regression coefficients', *Ann. Math. Stat.* **42**, 1328–1338 (1971).
29. L. A. Jaeckel, 'Estimating regression coefficients by minimizing the dispersion of residuals', *Ann. Math. Stat.* **5**, 1449–1458 (1972).
30. P. J. Bickel, 'On some analogues to linear combinations of order statistics in the linear model', *Ann. Stat.* **1**, 597–616 (1973).
31. R. Koenker and G. W. Bassett, 'Regression quantiles', *Econometrica*, **46**, 33–50 (1978).
32. D. F. Andrews, 'A robust method for multiple linear regression', *Technometrics*, **16**, 523–531 (1974).
33. A. F. Siegel, 'Robust regression using repeated medians', *Biometrika*, **69**, 242–244 (1982).
34. P. J. Rousseeuw and B. van Zomeren, 'Unmasking multivariate outliers and leverage points', *J. Am. Stat. Assoc.* **85**, 633–639 (1990).