CHAPTER 3

Multiple Regression

1. INTRODUCTION

In multiple regression, the response variable y_i is related to p explanatory variables $x_{i1}, \ldots, x_{ip}$ in the model

$$y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + e_i \qquad (i = 1, \ldots, n). \tag{1.1}$$

As in simple regression, the least squares (LS) technique for estimating the unknown parameters $\theta_1, \ldots, \theta_p$ is quite sensitive to the presence of outlying points. The identification of such points becomes more difficult, because it is no longer possible to spot the influential points in a scatterplot. Therefore, it is important to have a tool for identifying such points.

In the last few decades, several statisticians have given consideration to robust regression, whereas others have directed their attention to regression diagnostics (see Chapter 6). Both approaches are closely related by two important common aims, namely, identifying outliers and pointing out inadequacies of the model. However, they proceed in a different way. Regression diagnostics first attempt to identify points that have to be deleted from the data set, before applying a regression method. Robust regression tackles these problems in the inverse order, by designing estimators that dampen the impact of points that would be highly influential otherwise. A robust procedure tries to accommodate the majority of the data. Bad points, lying far away from the pattern formed by the good ones, will consequently possess large residuals from the robust fit. So in addition to insensitivity to outliers, a robust regression estimator makes the detection of these points an easy job. Of course, the residuals from LS cannot be used for this purpose, because the outliers

may possess very small LS residuals as the LS fit is pulled too much in the direction of these deviating points.

Let us look at some examples to illustrate the need for a robust alternative to LS. The first example is the well-known stackloss data set presented by Brownlee (1965). We have selected this example because it is a set of real data and it has been examined by a great number of statisticians (Draper and Smith 1966, Daniel and Wood 1971, Andrews 1974, Andrews and Pregibon 1978, Cook 1979, Dempster and Gasko-Green 1981, Atkinson 1982, Carroll and Ruppert 1985, Li 1985, and many others) by means of several methods. The data describe the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 four-dimensional observations (listed in Table 1). 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). Summarizing the findings cited in the literature, it can be said that most

Table 1. Stackloss Data

Index (i)	Rate (x_1)	Temperature (x_2)	Acid Concentration (x_3)	Stackloss (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Source: Brownlee (1965).

people concluded that observations 1, 3, 4, and 21 were outliers. According to some people, observation 2 is reported as an outlier too. Least squares regression yields the equation

$$\hat{y} = 0.716x_1 + 1.295x_2 - 0.152x_3 - 39.9 .$$

The LS index plot is shown in Figure 1. The standardization of the residuals is performed by the division of the raw residuals ($r_i = y_i - \hat{y}_i$) by the scale estimate corresponding to the fit. A horizontal band encloses the standardized residuals between -2.5 and 2.5. In Figure 1, no outliers strike the eye. From the LS index plot, one would conclude that the data set contains no outliers at all because all the standardized LS residuals fall nicely within the band. However, let us now look at Figure 2, the index plot associated with the least median of squares (LMS) fit

$$\hat{y} = 0.714x_1 + 0.357x_2 + 0.000x_3 - 34.5 .$$

This plot is based on a robust fit, and does indeed reveal the presence of harmful points. From this index plot it becomes immediately clear that

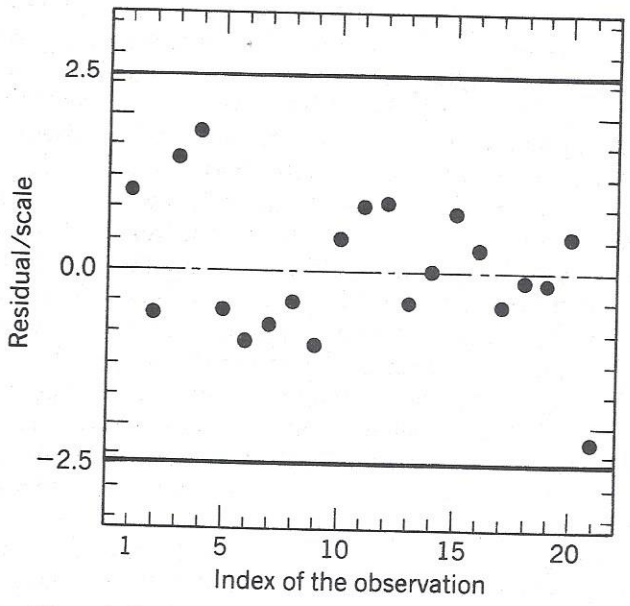


Figure 1. Stackloss data: Index plot associated with LS.

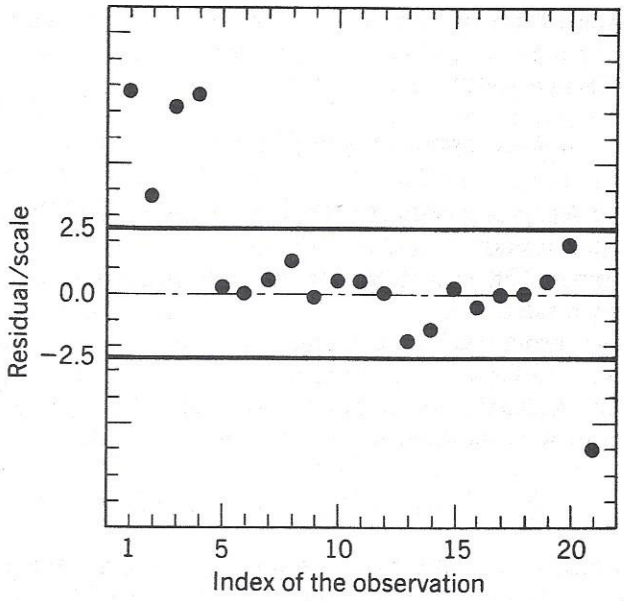


Figure 2. Stackloss data: Index plot associated with LMS.

the observations 1, 3, 4, and 21 are the most outlying, and that case 2 is intermediate because it is on the verge of the area containing the outliers. This shows how our robust regression technique is able to analyze 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 once more illustrates the danger of merely looking at the LS residuals. We would like to repeat that it is necessary to compare the standardized residuals of both the LS *and* the robust method in each regression analysis. If the results of the two procedures are in substantial agreement, then the LS can be trusted. If they differ, the robust technique can be used as a reliable tool for identifying 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 or cross-product terms and/or transforming the response variable). In this way, Atkinson (1985, pp. 129–136) analyzes the stackloss data by setting up models explaining $\log(y)$ by means of x_1, x_2, x_1x_2, and x_1^2.

The following example comes from the social sciences. The data set contains information on 20 schools from the Mid-Atlantic and New England states, drawn from a population studied by Coleman et al.

(1966). Mosteller and Tukey (1977) analyze this sample consisting of measurements on six different variables, one of which will be treated as response. They can be described as follows:

x_1 = staff salaries per pupil
x_2 = percent of white-collar fathers
x_3 = socioeconomic status composite deviation: means for family size, family intactness, father's education, mother's education, and home items
x_4 = mean teacher's verbal test score
x_5 = mean mother's educational level (one unit is equal to two school years)
y = verbal mean test score (all sixth graders).

The data set itself is exhibited in Table 2.

Table 2. Coleman Data Set, Containing Information on 20 Schools from the Mid-Atlantic and New England States

Index	x_1	x_2	x_3	x_4	x_5	y
1	3.83	28.87	7.20	26.60	6.19	37.01
2	2.89	20.10	−11.71	24.40	5.17	26.51
3	2.86	69.05	12.32	25.70	7.04	36.51
4	2.92	65.40	14.28	25.70	7.10	40.70
5	3.06	29.59	6.31	25.40	6.15	37.10
6	2.07	44.82	6.16	21.60	6.41	33.90
7	2.52	77.37	12.70	24.90	6.86	41.80
8	2.45	24.67	−0.17	25.01	5.78	33.40
9	3.13	65.01	9.85	26.60	6.51	41.01
10	2.44	9.99	−0.05	28.01	5.57	37.20
11	2.09	12.20	−12.86	23.51	5.62	23.30
12	2.52	22.55	0.92	23.60	5.34	35.20
13	2.22	14.30	4.77	24.51	5.80	34.90
14	2.67	31.79	−0.96	25.80	6.19	33.10
15	2.71	11.60	−16.04	25.20	5.62	22.70
16	3.14	68.47	10.62	25.01	6.94	39.70
17	3.54	42.64	2.66	25.01	6.33	31.80
18	2.52	16.70	−10.99	24.80	6.01	31.70
19	2.68	86.27	15.03	25.51	7.51	43.10
20	2.37	76.73	12.77	24.51	6.96	41.01

Source: Mosteller and Tukey (1977).

The ordinary LS regression for all 20 schools is given as

$$\hat{y} = -1.79x_1 + 0.044x_2 + 0.556x_3 + 1.11x_4 - 1.81x_5 + 19.9 .$$

Least median of squares regression yields the fit

$$\hat{y} = 0.580x_1 + 0.058x_2 + 0.637x_3 + 0.740x_4 - 2.32x_5 + 25.1 .$$

The left side of Table 3 lists the LS estimates and the associated residuals. These results reveal that the LS equation slightly underestimates the response for schools 3 and 11 and overestimates it for school 18. By only examining the LS results, the conclusion would be that schools 3, 11, and 18 are furthest away from the linear model. But from the right side of Table 3, it appears that school 11 does not deviate at all from the robust fit.

The robust regression spots schools 3, 17, and 18 as outliers by assigning large standardized residuals to them. Afterwards these standar-

Table 3. Coleman Data: Estimated Verbal Mean Test Score and the Associated Residuals for the LS Fit and the LMS Fit

	LS Results		LMS Results		
Index	Estimated Response	Standardized Residuals	Estimated Response	Standardized Residuals	Weights
1	36.661	0.17	38.883	−1.59	1
2	26.860	−0.17	26.527	−0.01	1
3	40.460	−1.90	41.273	−4.04	0
4	41.174	−0.23	42.205	−1.28	1
5	36.319	0.38	37.117	−0.01	1
6	33.986	−0.04	33.917	−0.01	1
7	41.081	0.35	41.628	0.15	1
8	33.834	−0.21	32.922	0.41	1
9	40.386	0.30	41.519	−0.43	1
10	36.990	0.10	34.847	2.00	1
11	25.508	−1.06	23.169	0.11	1
12	33.454	0.84	33.514	1.43	1
13	35.949	−0.51	34.917	−0.01	1
14	33.446	−0.17	32.591	0.43	1
15	24.479	−0.86	22.717	−0.01	1
16	38.403	0.63	40.041	−0.29	1
17	33.240	−0.69	35.121	−2.82	0
18	26.698	2.41	24.918	5.76	0
19	41.977	0.54	42.662	0.37	1
20	40.747	0.13	41.027	−0.01	1

dized residuals are used for computing weights, by giving a weight of 0 to all observations that have an absolute standardized residual larger than 2.5. Doing this for Coleman's data set gives rise to the last column in Table 3. These weights are then employed for a reweighted least squares (RLS) analysis. This amounts to the same thing as performing LS regression on the reduced data set containing only the 17 points with a weight of 1.

Apart from the fitted equation and the associated residuals, outliers also affect the t-based significance levels. This is important for the construction of confidence intervals and for hypothesis testing about regression coefficients (see also Section 3 of Chapter 2). For the Coleman data set, the significance of the regression coefficients turns out to be quite different in the LS fit and the RLS fit. The t-values in Table 4 test the null hypothesis H_0: $\theta_j = 0$ against the alternative H_1: $\theta_j \neq 0$ for the LS estimates. From this table it is seen that only the variables x_3 and x_4 have LS regression coefficients that are significantly different from zero for $\alpha = 5\%$, because their t-values exceed the critical value 2.1448 of the Student distribution with 14 $(= n - p)$ degrees of freedom, and hence their p-values are below 0.05.

Let us now analyze these data with the RLS using the weights of Table 3. This gives rise to the coefficients and t-values on the right side of Table 4. It is striking that for the cleaned data set, all the explanatory variables now have regression coefficients significantly different from zero for $\alpha = 5\%$ (because the 97.5% quantile of a t-distribution with 11 degrees of freedom equals 2.2010, so all p-values are less than 0.05).

As in Chapter 2, it must be noted that the opposite can also happen. In many examples, the "significance" of certain LS regression coefficients is only caused by an outlier, and then the corresponding RLS coefficients may no longer be significantly different from zero.

The ordinary LS method is not immune to the masking effect. This means that after the deletion of one or more influential points, another

Table 4. Coleman Data: t-Values Associated with the LS and the RLS Fit

	LS Results			RLS Reults		
Variable	Coefficient	t-Value	p-Value	Coefficient	t-Value	p-Value
x_1	−1.793	−1.454	0.1680	−1.203	−2.539	0.0275
x_2	0.044	0.819	0.4267	0.082	4.471	0.0009
x_3	0.556	5.979	0.0000	0.659	19.422	0.0000
x_4	1.110	2.559	0.0227	1.098	7.289	0.0000
x_5	−1.810	−0.893	0.3868	−3.898	−5.177	0.0003
Constant	19.949	1.464	0.1652	29.750	6.095	0.0001

observation may emerge as extremely influential, which was not visible at first. Therefore, the use of a high-breakdown regression method (such as the LMS) for the determination of the weights is indispensable. As an illustration, let us consider the "Salinity data" (Table 5) that were listed by Ruppert and Carroll (1980). It is a set of measurements of water salinity (i.e., its salt concentration) and river discharge taken in North Carolina's Pamlico Sound. We will fit a linear model where the salinity is regressed against salinity lagged by two weeks (x_1), the trend, that is, the number of biweekly periods elapsed since the beginning of the spring

Table 5. Salinity Data

Index (i)	Lagged Salinity (x_1)	Trend (x_2)	Discharge (x_3)	Salinity (y)
1	8.2	4	23.005	7.6
2	7.6	5	23.873	7.7
3	4.6	0	26.417	4.3
4	4.3	1	24.868	5.9
5	5.9	2	29.895	5.0
6	5.0	3	24.200	6.5
7	6.5	4	23.215	8.3
8	8.3	5	21.862	8.2
9	10.1	0	22.274	13.2
10	13.2	1	23.830	12.6
11	12.6	2	25.144	10.4
12	10.4	3	22.430	10.8
13	10.8	4	21.785	13.1
14	13.1	5	22.380	12.3
15	13.3	0	23.927	10.4
16	10.4	1	33.443	10.5
17	10.5	2	24.859	7.7
18	7.7	3	22.686	9.5
19	10.0	0	21.789	12.0
20	12.0	1	22.041	12.6
21	12.1	4	21.033	13.6
22	13.6	5	21.005	14.1
23	15.0	0	25.865	13.5
24	13.5	1	26.290	11.5
25	11.5	2	22.932	12.0
26	12.0	3	21.313	13.0
27	13.0	4	20.769	14.1
28	14.1	5	21.393	15.1

Source: Ruppert and Carroll (1980).

season (x_2); and the volume of river discharge into the sound (x_3). Carroll and Ruppert (1985) describe the physical background of the data. They indicated that cases 5 and 16 correspond to periods of very heavy discharge. Their analysis showed that the third and sixteenth observations conspire to hide the discrepant number 5. In fact, observation 5 can be recognized as influential only after the deletion of cases 3 and 16. This is a prime example of the masking effect. On the other hand, the LMS is not affected by this phenomenon and identifies 5 and 16 in a single blow. The LS fit is given by

$$\hat{y} = 0.777x_1 - 0.026x_2 - 0.295x_3 + 9.59\,,$$

whereas the LMS yields the equation

$$\hat{y} = 0.356x_1 - 0.073x_2 - 1.30x_3 + 36.7\,.$$

A residual plot associated with the LS fit is presented in Figure 3. In this figure the standardized residuals are plotted against the estimated response. In Figure 4 such a residual plot is given for the LMS regression. From Figure 3, it appears that there is nothing wrong with the fit. However, one has to keep in mind that leverage points tend to produce small LS residuals simply by virtue of their leverage. On the other hand,

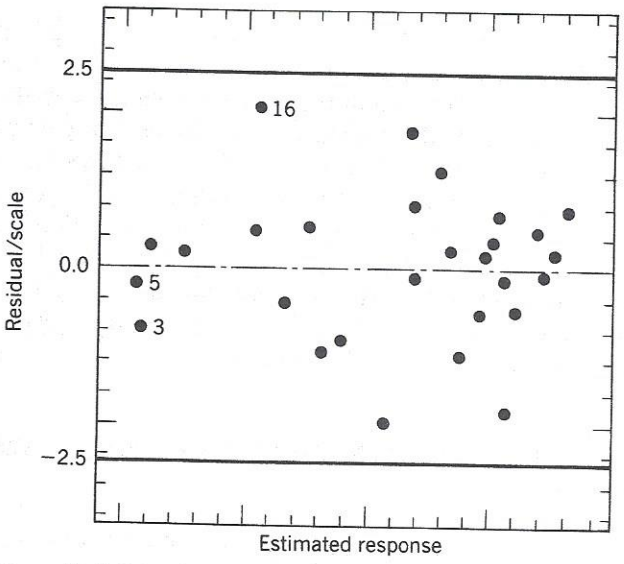


Figure 3. Salinity data: Residual plot associated with the LS fit.

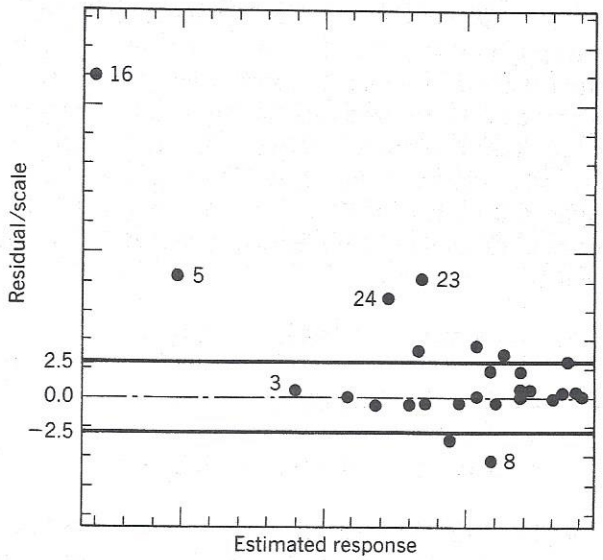


Figure 4. Salinity data: Residual plot associated with the LMS fit.

Figure 4 gives evidence of the presence of outlying observations, because some points fall far from the band. In this example, the LS residual plot cannot be trusted because it differs too much from the one associated with a robust fit.

As we said before, residual plots can also indicate possible defects in the model's functional form in the direction of the fitted values. A residual plot may, for example, display a variance pattern that is a monotone function of the response. If the functional part of the model is not misspecified, then plotting the standardized LMS residuals versus $\hat{y}_i$ gives rise to a horizontal band of points that look "structureless." Also, anomalies in the pattern might suggest a transformation of the variables in the model. For example, a curved pattern in a residual plot may lead to replacing the observed y_i by some function of y_i (log y_i or y_i raised to some power). In example 2 of Section 3 we will illustrate the use of residual plots to remedy model failures.

2. COMPUTATION OF LEAST MEDIAN OF SQUARES MULTIPLE REGRESSION

The question "How do we run program PROGRESS?" has for the most part been answered in Section 2 of Chapter 2, in the context of simple

regression. The interactive session accompanying a multiple regression analysis is completely identical.

The special treatment of data sets with missing values has not yet been discussed. In that case the interactive input becomes a little bit longer. We will illustrate this situation for part of the "Air Quality" data that originated with the New York State Department of Conservation and the National Weather Service; these data were reported in Chambers et al. (1983). The whole data set consists of daily readings of air quality values from May 1, 1973 to September 30, 1973. We will use only the values for May in our example. The variables are the mean ozone concentration (in parts per billion) from 1300 to 1500 hours at Roosevelt Island (OZONE ppb), solar radiation in Longleys in the frequency band 4000–7700 Å from 0800 to 1200 hours at Central Park (SOLAR RADI), average wind speed (in miles per hour) between 0700 and 1000 hours at La Guardia Airport (WINDSPEED), and maximum daily temperature (in degrees Fahrenheit) at La Guardia Airport (TEMPERATUR). The data are exhibited in Table 6.

The aim of the analysis is to explain the ozone concentration by means of the other variables. In Table 6, one can observe that the measurements for OZONE ppb and/or SOLAR RADI are not registered for some days. PROGRESS provides two methods for handling such an incomplete data set. One or the other can be chosen by answering 1 or 2 to the question

```
CHOOSE AN OPTION FOR THE TREATMENT OF MISSING VALUES
 ----------------------------------------
0=THERE ARE NO MISSING VALUES IN THE DATA
1=ELIMINATION OF THE CASES FOR WHICH AT LEAST ONE VARIABLE IS MISSING
2=ESTIMATES ARE FILLED IN FOR UNOBSERVED VALUES
ENTER YOUR CHOICE:
```

When the option for missing values is not equal to zero, PROGRESS needs additional information on the missing value codes for each variable. First of all, the question

```
IS THERE A UNIQUE VALUE WHICH IS TO BE INTERPRETED
AS A MISSING MEASUREMENT FOR ANY VARIABLE?
ANSWER YES OR NO:
```

must be answered. When the answer is YES, the statement

```
PLEASE ENTER THIS VALUE:
```

appears, where the user has to give a value that will be interpreted as the

Table 6. Air Quality Data Set for May 1973[a]

Index (i)	SOLAR RADI (x_1)	WINDSPEED (x_2)	TEMPERATUR (x_3)	OZONE ppb (y)
1	190	7.4	67	41
2	118	8.0	72	36
3	149	12.6	74	12
4	313	11.5	62	18
5	9999	14.3	56	999
6	9999	14.9	66	28
7	299	8.6	65	23
8	99	13.8	59	19
9	19	20.1	61	8
10	194	8.6	69	999
11	9999	6.9	74	7
12	256	9.7	69	16
13	290	9.2	66	11
14	274	10.9	68	14
15	65	13.2	58	18
16	334	11.5	64	14
17	307	12.0	66	34
18	78	18.4	57	6
19	322	11.5	68	30
20	44	9.7	62	11
21	8	9.7	59	1
22	320	16.6	73	11
23	25	9.7	61	4
24	92	12.0	61	32
25	66	16.6	57	999
26	266	14.9	58	999
27	9999	8.0	57	999
28	13	12.0	67	23
29	252	14.9	81	45
30	223	5.7	79	115
31	279	7.4	76	37

[a]The values 9999 of x_1 (solar radiation) indicate missing measurements. Also the numbers 999 of y (ozone ppb) correspond to missing values.

Source: Chambers et al. (1983).

missing value code for all the variables in the analysis. Otherwise, the user has to answer the following question for each variable:

```
DOES VARIABLE.....CONTAIN MISSING VALUE(S)?
ANSWER YES OR NO:
```

(Instead of the dots, the actual label of the variable concerned will be printed.) If the answer to this question is YES, then the user has to enter the missing value code for this variable:

```
ENTER THE VALUE OF THIS VARIABLE WHICH HAS TO BE INTERPRETED AS
THE MISSING VALUE CODE:
```

In both missing value options, the program makes an inventory of the cases with incomplete data, for each variable for which missing values were announced. When there are variables for which more than 80% of the cases have a missing value, the program will terminate (after giving a message). When analyzing the same data again, the user should no longer include these partially observed variables because they do not contain enough information.

Let us now look at the two missing value options provided by PROGRESS. The first avenue open to the researcher is to eliminate the cases for which at least one variable is missing. This can be achieved by setting the missing value option equal to 1. When the print option differs from 0, a complete table is given, where for each case the user of the program can see which variables were missing. The regression analysis is then performed on the remaining cases. (Each case keeps its original number.)

In fact, this option can also be used for another purpose. It may happen that one wishes to perform the analysis only on a part of the data set. For example, in a sample containing attributes of many people, one may want to fit a linear model to the women only. Then one can use the missing value option 1 for the variable associated with the sex of each person. The value corresponding to men then has to be taken as the "missing" value code, in order to eliminate the observations for men.

An alternative treatment, which corresponds to option 2, consists of filling in guesses for unobserved values. This can be necessary in circumstances where deleting all partially observed cases would result in an extremely small sample (which might even be empty). Nevertheless, even for this option, cases for which the response y is lacking will first be dropped from the data set. Also, cases for which *all* explanatory variables are missing will be removed. On this reduced data set, missing data will

be replaced by the median of the corresponding variable. On the resulting data set, the previous methods of estimation can be applied. When the print option is not 0, the output of PROGRESS delivers a complete inventory of the deleted cases and of the missing values that are replaced by medians. Also here, the cases retain the numbering of the original data set.

For the Air Quality data, we encoded the missing values of variable SOLAR RADI by 9999, and those of OZONE ppb by 999. These codes are acceptable because these values have not been observed. Let us now look at a listing of the interactive input for this data set.

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

ENTER THE NUMBER OF CASES PLEASE : 31

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

WHAT IS THE TOTAL NUMBER OF VARIABLES IN YOUR DATA SET?
------------------------------------------------------
PLEASE GIVE A NUMBER BETWEEN 1 AND 50 :4

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

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

HOW MANY EXPLANATORY VARIABLES DO YOU WANT TO USE IN THE ANALYSIS?
------------------------------------------------------------------
(AT MOST    3 ) : 3

EXPLANATORY VARIABLES       :   POSITION   LABEL (AT MOST 10 CHARACTERS)
-----------------------------------↓↓↓↓----↓↓↓↓↓↓↓↓↓↓--------------------
NUMBER     1                :         1    SOLAR RADI
NUMBER     2                :         2    WINDSPEED
NUMBER     3                :         3    TEMPERATUR

HOW MUCH OUTPUT DO YOU WANT?
----------------------------
0 = SMALL OUTPUT       : LIMITED TO BASIC RESULTS
1 = MEDIUM-SIZED OUTPUT: ALSO INCLUDES 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

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 : 3

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

GIVE THE NAME OF THE FILE CONTAINING THE DATA (e.g. TYPE  A:EXAMPLE.DAT ),
or TYPE  KEY  IF YOU PREFER TO ENTER THE DATA BY KEYBOARD.
WHAT DO YOU CHOOSE ?  B:AIRMAY.DAT
```

```
WHERE DO YOU WANT YOUR OUTPUT?
------------------------------
   TYPE  CON  IF YOU WANT IT ON THE SCREEN
or TYPE  PRN  IF YOU WANT IT ON THE PRINTER
or TYPE THE NAME OF A FILE (e.g. B:EXAMPLE.OUT)
(WARNING : IF THERE ALREADY EXISTS A FILE WITH THE SAME NAME
           THE OLD FILE WILL BE OVERWRITTEN.)
WHAT DO YOU CHOOSE ?  B:AIRMAY.RES

PLEASE ENTER A TITLE FOR THE OUTPUT (AT MOST 60 CHARACTERS):
------------------------------------------------------------
 AIR QUALITY MEASUREMENTS FOR NEW YORK

DO YOU WANT TO READ THE DATA IN FREE FORMAT?
-------------------------------------------
THIS MEANS THAT YOU ONLY HAVE TO INSERT BLANK(S) BETWEEN NUMBERS.
(WE ADVISE USERS WITHOUT KNOWLEDGE OF FORTRAN FORMATS TO ANSWER YES.)
MAKE YOUR CHOICE (YES/NO): YES


WHICH VERSION OF THE ALGORITHM WOULD YOU LIKE TO USE?
----------------------------------------------------
Q = QUICK VERSION
E = EXTENSIVE SEARCH
ENTER YOUR CHOICE PLEASE (Q OR E) : E

CHOOSE AN OPTION FOR THE TREATMENT OF MISSING VALUES
---------------------------------------------------
0 = THERE ARE NO MISSING VALUES IN THE DATA
1 = ELIMINATION OF THE CASES FOR WHICH AT LEAST ONE VARIABLE IS MISSING
2 = ESTIMATES ARE FILLED IN FOR UNOBSERVED VALUES
ENTER YOUR CHOICE : 1



      ********************************************************************
      * P R O G R E S S  WILL PERFORM A REGRESSION WITH CONSTANT TERM *
      ********************************************************************

THE NUMBER OF CASES EQUALS                          31
THE NUMBER OF EXPLANATORY VARIABLES EQUALS           3
OZONE ppb  IS THE RESPONSE VARIABLE.
YOUR DATA RESIDE ON FILE       : B:AIRMAY.DAT
TITLE FOR OUTPUT : AIR QUALITY MEASUREMENTS FOR NEW YORK
THE DATA WILL BE READ IN FREE FORMAT.
LARGE OUTPUT IS WANTED.
BOTH TYPES OF RESIDUAL PLOTS ARE WANTED.
THE EXTENSIVE SEARCH ALGORITHM WILL BE USED.
TREATMENT OF MISSING VALUES IN OPTION 1: THIS MEANS THAT A CASE WITH A
MISSING VALUE FOR AT LEAST ONE VARIABLE WILL BE DELETED.
YOUR OUTPUT WILL BE WRITTEN ON : B:AIRMAY.RES

ARE ALL THESE OPTIONS OK ?  YES OR NO : YES

IS THERE A UNIQUE VALUE WHICH IS TO BE INTERPRETED
AS A MISSING MEASUREMENT FOR ANY VARIABLE?
ANSWER YES OR NO : NO
DOES VARIABLE SOLAR RADI CONTAIN MISSING VALUE(S)?
ANSWER YES OR NO : YES
ENTER THE VALUE OF THIS VARIABLE WHICH HAS TO BE INTERPRETED AS
THE MISSING VALUE CODE : 9999
DOES VARIABLE WINDSPEED  CONTAIN MISSING VALUE(S)?
ANSWER YES OR NO : NO
DOES VARIABLE TEMPERATUR CONTAIN MISSING VALUE(S)?
ANSWER YES OR NO : NO
DOES THE RESPONSE VARIABLE CONTAIN MISSING VALUE(S)?
ANSWER YES OR NO : YES
ENTER THE VALUE OF THIS VARIABLE WHICH HAS TO BE INTERPRETED AS
THE MISSING VALUE CODE : 999
```

The treatment of the missing values for this data set appears on the output of PROGRESS as follows:

```
TREATMENT OF MISSING VALUES IN OPTION 1: THIS MEANS THAT A CASE WITH A
MISSING VALUE FOR AT LEAST ONE VARIABLE WILL BE DELETED.

YOUR DATA RESIDE ON FILE        : B:AIRMAY.DAT
VARIABLE SOLAR RADI HAS A MISSING VALUE FOR   4 CASES.
VARIABLE OZONE ppb  HAS A MISSING VALUE FOR   5 CASES.

CASE HAS A MISSING VALUE FOR VARIABLES (VARIABLE NUMBER  5 IS THE RESPONSE)
----                           ---------
   5                           1   5
   6                           1
  10                           5
  11                           1
  25                           5
  26                           5
  27                           1   5

THERE ARE   24 CASES STAYING IN THE ANALYSIS.
THE OBSERVATIONS, AFTER TREATMENT OF MISSING VALUES :

     SOLAR RADI WINDSPEED  TEMPERATUR OZONE ppb
 1    190.0000    7.4000    67.0000    41.0000
 2    118.0000    8.0000    72.0000    36.0000
 3    149.0000   12.6000    74.0000    12.0000
 4    313.0000   11.5000    62.0000    18.0000
 7    299.0000    8.6000    65.0000    23.0000
 8     99.0000   13.8000    59.0000    19.0000
 9     19.0000   20.1000    61.0000     8.0000
12    256.0000    9.7000    69.0000    16.0000
13    290.0000    9.2000    66.0000    11.0000
14    274.0000   10.9000    68.0000    14.0000
15     65.0000   13.2000    58.0000    18.0000
16    334.0000   11.5000    64.0000    14.0000
17    307.0000   12.0000    66.0000    34.0000
18     78.0000   18.4000    57.0000     6.0000
19    322.0000   11.5000    68.0000    30.0000
20     44.0000    9.7000    62.0000    11.0000
21      8.0000    9.7000    59.0000     1.0000
22    320.0000   16.6000    73.0000    11.0000
23     25.0000    9.7000    61.0000     4.0000
24     92.0000   12.0000    61.0000    32.0000
28     13.0000   12.0000    67.0000    23.0000
29    252.0000   14.9000    81.0000    45.0000
30    223.0000    5.7000    79.0000   115.0000
31    279.0000    7.4000    76.0000    37.0000
```

The least squares analysis of the reduced data set is printed in Table 7. Some of the fitted OZONE ppb values (cases 9 and 18) are negative, which is physically impossible. The strange behavior of this fit is easy to understand when comparing it to the robust analysis in Table 8. First of all, the equations of both fits differ substantially from each other. In the

Table 7. Air Quality Data: LS Fit with Estimated Response and Standardized Residuals

Variable	Coefficient	Standard Error	*t*-Value
SOLAR RADI	−0.01868	0.03628	−0.51502
WINDSPEED	−1.99577	1.14092	−1.74926
TEMPERATUR	1.96332	0.66368	2.95823
Constant	−79.99270	46.81654	−1.70864

Index	Estimated "OZONE ppb"	Standardized Residuals
1	33.231	0.43
2	43.195	−0.40
3	37.362	−1.41
4	12.933	0.28
7	24.873	−0.10
8	6.452	0.70
9	−0.700	0.48
12	31.334	−0.85
13	25.807	−0.82
14	26.639	−0.70
15	6.321	0.65
16	16.468	−0.14
17	19.901	0.78
18	−6.263	0.68
19	24.545	0.30
20	21.552	−0.59
21	16.335	−0.85
22	24.221	−0.73
23	19.944	−0.89
24	14.101	0.99
28	27.357	−0.24
29	44.591	0.02
30	59.567	3.08
31	49.238	−0.68

column comprising the standardized residuals in Table 8, case 30 emerges as an outlier. This single outlier is the cause of the bad LS fit because it has tilted the LS hyperplane in its direction. Because of this, the other points are not well fitted anymore by LS.

Of course, negative predicted values are to be expected whenever a linear equation is fitted. Indeed, when the regression surface is not horizontal, there always exist vectors $\mathbf{x}$ for which the corresponding

Table 8. Air Quality Data: RLS Fit Based on LMS

Variable	Coefficient	Standard Error	t-Value
SOLAR RADI	0.00559	0.02213	0.25255
WINDSPEED	−0.74884	0.71492	−1.04745
TEMPERATUR	0.99352	0.42928	2.31438
Constant	−37.51613	28.95417	−1.29571

Index	Estimated "OZONE ppb"	Standardized Residuals
1	24.571	1.52
2	28.686	0.68
3	27.402	−1.42
4	17.220	0.07
7	22.294	0.07
8	11.321	0.71
9	8.143	−0.01
12	25.204	−0.85
13	22.788	−1.09
14	23.413	−0.87
15	10.587	0.69
16	19.325	−0.49
17	20.786	1.22
18	5.772	0.02
19	23.232	0.63
20	17.065	−0.56
21	13.883	−1.19
22	24.369	−1.24
23	15.965	−1.11
24	14.617	1.61
28	20.137	0.26
29	33.210	1.09
30	37.950	7.13
31	34.010	0.28

predicted $\hat{y}$ is negative. This is still true for robust regression, where one can easily encounter a negative $\hat{y}_i$ in a leverage point $\mathbf{x}_i$. However, in the above example the LS predictions are negative in *good* observations!

3. EXAMPLES

In the preceding sections we have seen that the high-breakdown fits provide much to think about. In particular, the standardized residuals

associated with a robust fit yield powerful diagnostic tools. For example, they can be displayed in residual plots. These graphics make it easy to detect outlying values and call attention to model failures. Also, the residuals can be used to determine a weight for each point. Such weights make it possible to bound the effect of the outliers by using them for RLS. The fit resulting from this reweighting describes the trend followed by the majority of the data. The statistics associated with this fit, like t- and F-values, are more trustworthy than those calculated from the ordinary LS regression.

Example 1: Hawkins–Bradu–Kass Data

We shall use the data generated by Hawkins, Bradu, and Kass (1984) for illustrating some of the merits of a robust technique. Such artificial data offer the advantage that at least the position of the bad points is known exactly, which avoids some of the controversies that are inherent in the analysis of real data. In this way, the effectiveness of the technique can be measured. The data set is listed in Table 9 and consists of 75 observations in four dimensions (one response and three explanatory variables). The first 10 observations are bad leverage points, and the next four points are good leverage points (i.e., their $\mathbf{x}_i$ are outlying, but the corresponding y_i fit the model quite well). We will compare the LS with our robust regression. In Hawkins et al. (1984), it is mentioned that M-estimators do not produce the expected results, because the outliers (the bad leverage points) are masked and the four good leverage points appear outlying because they possess large residuals from those fits. This should not surprise us, because M-estimators break down early in the presence of leverage points. A certain version of the "elemental sets" diagnostic of Hawkins et al. locates the outliers, but this technique would not have coped with a larger fraction of contamination. (More details about outlier diagnostics and their breakdown points will be provided in Chapter 6.)

Let us now restrict the discussion to LS and LMS. From the index plot associated with LS (see Figure 5), it appears that observations 11, 12, and 13 are outliers because they fall outside the ± 2.5 band. Unfortunately, from the generation one knows that these are good observations. The bad leverage points have tilted the LS fit totally in their direction. Therefore the first 10 points have small standardized LS residuals. (The index plots associated with M-estimators are very similar to that of LS.)

On the other hand, the index plot of the LMS (Figure 6) identifies the first 10 points as the influential observations. The four good leverage points fall in the neighborhood of the dashed line through 0. This means that these points are well accommodated by the LMS fit. Clearly, the

Table 9. Artificial Data Set of Hawkins, Bradu, and Kass (1984)

Index	x_1	x_2	x_3	y	Index	x_1	x_2	x_3	y
1	10.1	19.6	28.3	9.7	39	2.1	0.0	1.2	−0.7
2	9.5	20.5	28.9	10.1	40	0.5	2.0	1.2	−0.5
3	10.7	20.2	31.0	10.3	41	3.4	1.6	2.9	−0.1
4	9.9	21.5	31.7	9.5	42	0.3	1.0	2.7	−0.7
5	10.3	21.1	31.1	10.0	43	0.1	3.3	0.9	0.6
6	10.8	20.4	29.2	10.0	44	1.8	0.5	3.2	−0.7
7	10.5	20.9	29.1	10.8	45	1.9	0.1	0.6	−0.5
8	9.9	19.6	28.8	10.3	46	1.8	0.5	3.0	−0.4
9	9.7	20.7	31.0	9.6	47	3.0	0.1	0.8	−0.9
10	9.3	19.7	30.3	9.9	48	3.1	1.6	3.0	0.1
11	11.0	24.0	35.0	−0.2	49	3.1	2.5	1.9	0.9
12	12.0	23.0	37.0	−0.4	50	2.1	2.8	2.9	−0.4
13	12.0	26.0	34.0	0.7	51	2.3	1.5	0.4	0.7
14	11.0	34.0	34.0	0.1	52	3.3	0.6	1.2	−0.5
15	3.4	2.9	2.1	−0.4	53	0.3	0.4	3.3	0.7
16	3.1	2.2	0.3	0.6	54	1.1	3.0	0.3	0.7
17	0.0	1.6	0.2	−0.2	55	0.5	2.4	0.9	0.0
18	2.3	1.6	2.0	0.0	56	1.8	3.2	0.9	0.1
19	0.8	2.9	1.6	0.1	57	1.8	0.7	0.7	0.7
20	3.1	3.4	2.2	0.4	58	2.4	3.4	1.5	−0.1
21	2.6	2.2	1.9	0.9	59	1.6	2.1	3.0	−0.3
22	0.4	3.2	1.9	0.3	60	0.3	1.5	3.3	−0.9
23	2.0	2.3	0.8	−0.8	61	0.4	3.4	3.0	−0.3
24	1.3	2.3	0.5	0.7	62	0.9	0.1	0.3	0.6
25	1.0	0.0	0.4	−0.3	63	1.1	2.7	0.2	−0.3
26	0.9	3.3	2.5	−0.8	64	2.8	3.0	2.9	−0.5
27	3.3	2.5	2.9	−0.7	65	2.0	0.7	2.7	0.6
28	1.8	0.8	2.0	0.3	66	0.2	1.8	0.8	−0.9
29	1.2	0.9	0.8	0.3	67	1.6	2.0	1.2	−0.7
30	1.2	0.7	3.4	−0.3	68	0.1	0.0	1.1	0.6
31	3.1	1.4	1.0	0.0	69	2.0	0.6	0.3	0.2
32	0.5	2.4	0.3	−0.4	70	1.0	2.2	2.9	0.7
33	1.5	3.1	1.5	−0.6	71	2.2	2.5	2.3	0.2
34	0.4	0.0	0.7	−0.7	72	0.6	2.0	1.5	−0.2
35	3.1	2.4	3.0	0.3	73	0.3	1.7	2.2	0.4
36	1.1	2.2	2.7	−1.0	74	0.0	2.2	1.6	−0.9
37	0.1	3.0	2.6	−0.6	75	0.3	0.4	2.6	0.2
38	1.5	1.2	0.2	0.9					

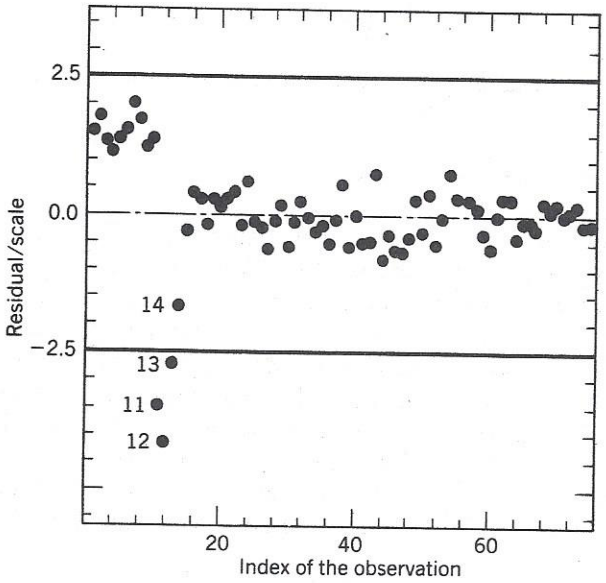


Figure 5. Hawkins–Bradu–Kass data: Index plot associated with LS regression.

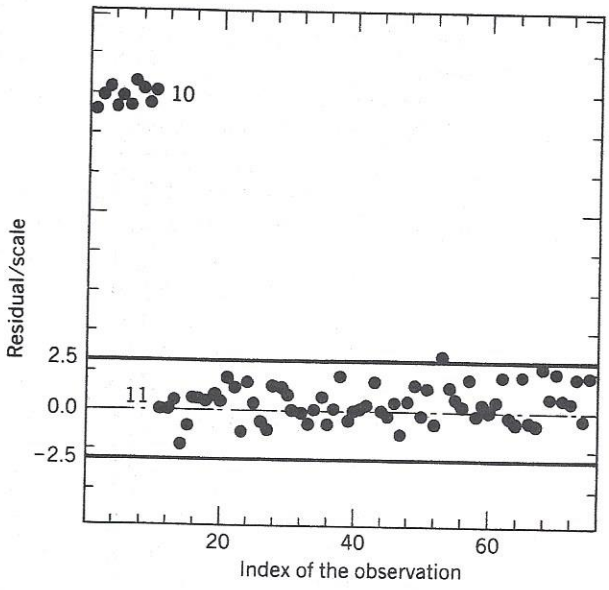


Figure 6. Hawkins–Bradu–Kass data: Index plot associated with LMS regression.

conclusions drawn from the LMS index plot agree with the construction of the data.

The following example illustrates the use of residual plots for model specification.

Example 2: Cloud Point Data

Table 10 shows a set of measurements concerning the cloud point of a liquid (Draper and Smith 1966, p. 162). The cloud point is a measure of the degree of crystallization in a stock and can be measured by the refractive index. The purpose is to construct a model where the percentage of I-8 in the base stock can be used as a predictor for the cloud point. Because the data contain only two variables, it is possible to explore the relation between these variables in a scatterplot. The scatterplot associated with Table 10 can be found in Figure 7.

The curved pattern in Figure 7 indicates that a simple linear model is not adequate. We will now examine whether the residual plot associated

Table 10. Cloud Point of a Liquid

Index (i)	Percentage of I-8 (x)	Cloud Point (y)
1	0	22.1
2	1	24.5
3	2	26.0
4	3	26.8
5	4	28.2
6	5	28.9
7	6	30.0
8	7	30.4
9	8	31.4
10	0	21.9
11	2	26.1
12	4	28.5
13	6	30.3
14	8	31.5
15	10	33.1
16	0	22.8
17	3	27.3
18	6	29.8
19	9	31.8

Source: Draper and Smith (1969, p. 162).

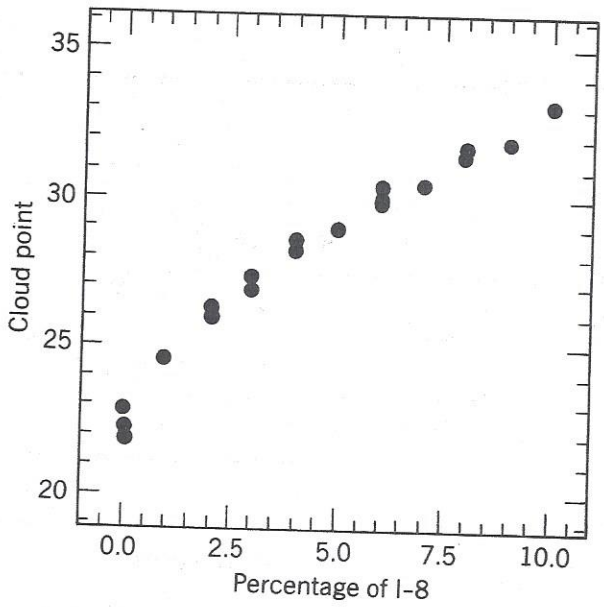


Figure 7. Cloud point data: Scatterplot.

with the linear fit would have suggested this. From the residual plot of the LS line in Figure 8, it is clear that the straight-line fit is imperfect because the residuals appear not to be randomly distributed about the zero line.

In order to be sure that this pattern is not caused by the presence of outliers, we will compare Figure 8 to the residual plot associated with the RLS line (Figure 9). The departure from linearity is magnified in this plot. The residuals inside the ± 2.5 band tend to follow a parabolic curve. Those associated with cases 1, 10, and 16 fall outside the band, which indicates that they have large residuals from the RLS line. In spite of the fact that the slopes of both the LS and RLS lines are significantly different from zero (see the t-values in Table 11), the linear model is not appropriate.

Moreover, the value of R^2, which is a measure of model adequacy, is high. For LS, R^2 equals 0.955. The RLS value of R^2 is 0.977, showing that 97.7% of the total variation in the response is accounted for by the explanatory variable. Of course, R^2 is only a single number summary. It is not able to characterize an entire distribution or to indicate all possible defects in the functional form of the model. A large R^2 does not ensure that the data have been well fitted. On the other hand, the residual plots do embody the functional part of the model. Therefore, inspection of

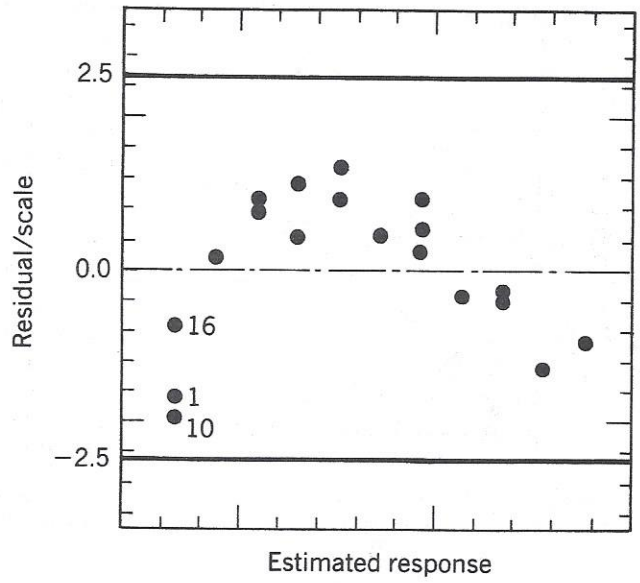


Figure 8. Cloud point data: Residual plot associated with LS.

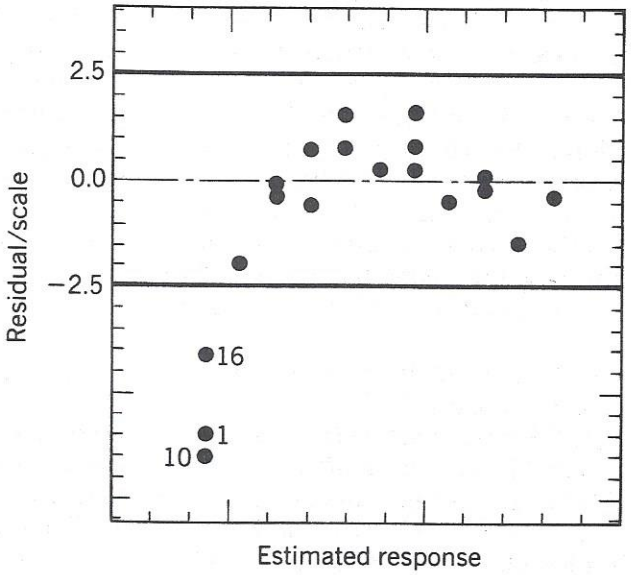


Figure 9. Cloud point data: Residual plot associated with LMS-based RLS.

Table 11. Cloud Point Data: Estimated Slope and Intercept by LS and RLS Regression, with Their t-Values

	LS Results			RLS Results		
Variable	$\hat{\theta}$	Standard Error	t-Value	$\hat{\theta}$	Standard Error	t-Value
Percentage of I-8	1.05	0.055	18.9	0.89	0.036	24.7
Constant	23.35	0.297	78.7	24.37	0.212	115.2
		$R^2 = 0.955$			$R^2 = 0.977$	

these plots is an indispensable part of regression analysis, even when R^2 is large or in the case of significant t-values. When this graphical display reveals an unexpected pattern, one has to remedy the defect by adapting the model. Depending on the anomalies in the pattern of the residuals, it may happen that an additional explanatory variable is necessary, or that some variables in the model have to be transformed. The possible improved models have to be restricted to those that are linear in the coefficients, or else we leave the domain of linear regression. However, several nonlinear functions are linearizable by using a suitable transformation. Daniel and Wood (1971) list various types of useful transformations.

A curved plot such as Figure 9 is one way to indicate nonlinearity. The usual approach to account for this apparent curvature is to consider the use of a quadratic model such as

$$y = \theta_1 x + \theta_2 x^2 + \theta_3 + e\,. \tag{3.1}$$

The LS and RLS estimates for this model are given in Table 12, along with some summary statistics.

The R^2 values for both the LS and the RLS have increased a little bit, whereas the t-values of the regression coefficients have hardly changed. For both fits, the coefficients are significantly different from zero at $\alpha = 5\%$. Let us now look at the distribution of the residuals associated with LS and RLS, in Figures 10 and 11, respectively.

Examining the LS residual plot, it would appear that the additional squared term has not been completely successful in restoring the distribution of the residuals. The pattern is not entirely neutral. Before deciding to change the model again, let us analyze the information provided by the RLS residual plot. From this plot, some outliers strike the eye. The observations 1, 10, and 15, which have obtained a zero weight from the LMS fit, fall outside the horizontal band. (Notwithstanding their zero

Table 12. Cloud Point Data: LS and RLS Estimates for the Quadratic Model along with Summary Values

	LS Results			RLS Results		
Variable	$\hat{\theta}$	Standard Error	t-Value	$\hat{\theta}$	Standard Error	t-Value
Percentage of I-8	1.67	0.099	16.9	1.57	0.084	18.8
(Percentage of I-8)^2	−0.07	0.010	−6.6	−0.07	0.009	−7.5
Constant	22.56	0.198	113.7	22.99	0.172	133.7
	$R^2 = 0.988$			$R^2 = 0.993$		

weights, the outliers are still indicated in the plot.) The presence of these points has affected the LS estimates because LS tries to make all the residuals small, even those associated with outliers, at the cost of an increased estimate of θ_1. On the other hand, the outlying observations do not act upon the RLS estimates, at which they obtain a large residual. The residuals inside the band in Figure 11, which correspond to the other observations, display no systematic pattern of variation. Summarizing the findings from Figures 10 and 11, one can conclude that the quadratic equation describes a more appropriate fit than the

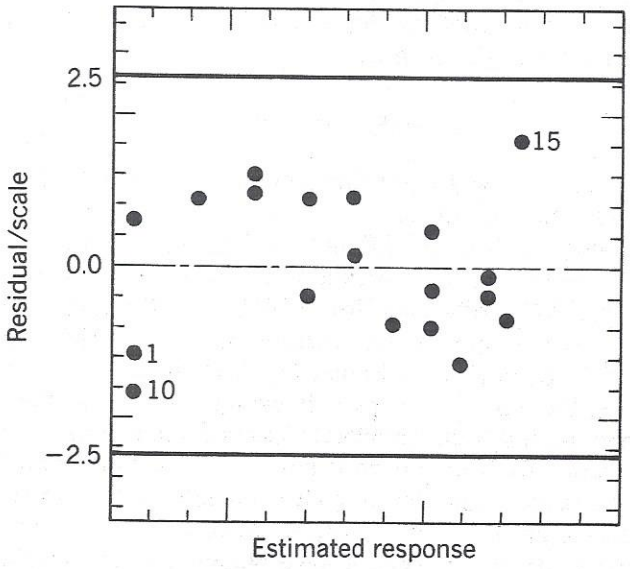


Figure 10. Cloud point data: LS residual plot for the quadratic model.

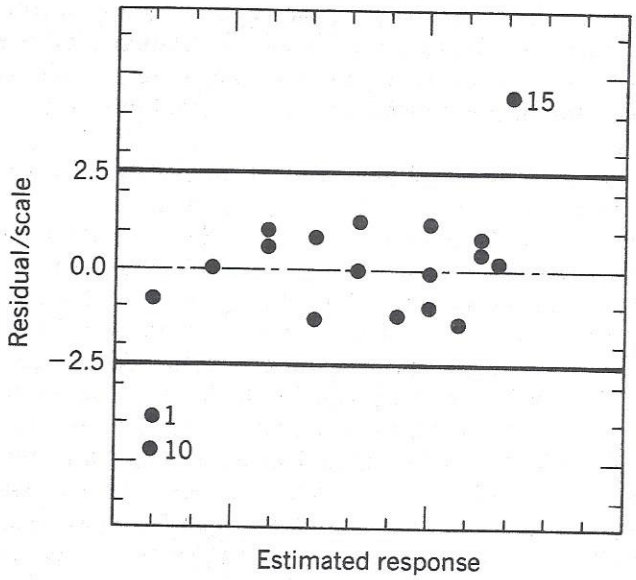


Figure 11. Cloud point data: RLS residual plot for the quadratic model.

simple line. Moreover, the residual plot of the RLS quadratic fit locates three observations that were responsible for the deformed pattern in the residual plot of the LS quadratic fit.

The presence of outliers or the choice of an appropriate model are not the only problems in regression analysis. A nearly linear dependence between two or more explanatory variables can also seriously disturb the estimated regression surface or make the regression coefficients uninterpretable. This phenomenon is called *multicollinearity*. (The terms *collinearity* and *ill-conditioning* are also employed in the literature.) The ideal situation would be that there is no relation among the variables in the factor space. In that case, it is easy to interpret a regression coefficient as the amount of change in the response when the corresponding explanatory variable grows with one unit, while the other explanatory variable(s) remain fixed. When collinearity is present in the data, the contribution of a single explanatory variable to the regression equation is hard to estimate.

The detection of multicollinearity may be very complicated. When there are only two explanatory variables, then collinearity leads to a high absolute value of the Pearson correlation or the alternative (more robust) Spearman rank correlation coefficient. For higher-dimensional factor spaces, the two-by-two correlation coefficients are not always sufficient

for discovering collinearity, because the collinearity may involve several variables. Therefore, the squared multiple correlation coefficients R_j^2 of the regression of x_j on the remaining x-variables are possible diagnostics for measuring the degree to which any x_j is related to the other explanatory variables.

Chatterjee and Price (1977), Belsley et al. (1980), Weisberg (1980), and Hocking (1983), among others, report some tools for identifying collinearity. Most of these tools are based on the correlation matrix or its inverse. For example, the so-called Variance Inflation Factor (VIF) is based on the estimated variance of the ith regression coefficient (obtained from LS). There are some debates whether or not the data should be standardized first, because this may have a large effect on the resulting collinearity diagnostics (see the article by Belsley 1984 and its discussion). Another approach to deal with collinearity is ridge regression (Hoerl and Kennard 1970, 1981), based on the principle of using a little bit of all the variables rather than all of some variables and none of the remaining ones (Marquardt and Snee 1975). We will not describe these techniques in further detail here. Unfortunately, most of them are not immune to the presence of contamination, as was also noted by Mason and Gunst (1985) and Ercil (1986). For instance, consider the nonrobustness of the Pearson correlation coefficient, which may be affected a great deal by outliers. This means that the correlation coefficient can be close to zero because of the presence of a single outlier disturbing an otherwise linear relationship, thereby hiding collinearity. On the other hand, the correlation coefficient can also be carried arbitrarily close to 1 by means of a far outlier, which appears to create collinearity. Therefore, the identification of linear dependencies in factor space, combined with the detection of outliers, is an important problem of regression analysis. Indeed, collinearity inflates the variance of the regression coefficients, may be responsible for a wrong sign of the coefficients, and may affect statistical inference in general. Therefore, in cases where the presence of collinearity is anticipated, we recommend the use of the classical collinearity diagnostics on both the original data set and on the reweighted one based on LMS, in which the outliers have been removed.

Example 3: Heart Catheterization Data

The "Heart Catheterization" data set of Table 13 (from Weisberg 1980, p. 218 and Chambers et al. 1983, p. 310) demonstrates some of the effects caused by collinearity.

A catheter is passed into a major vein or artery at the femoral region

Table 13. Heart Catheterization Data[a]

Index (i)	Height (x_1)	Weight (x_2)	Catheter Length (y)
1	42.8	40.0	37
2	63.5	93.5	50
3	37.5	35.5	34
4	39.5	30.0	36
5	45.5	52.0	43
6	38.5	17.0	28
7	43.0	38.5	37
8	22.5	8.5	20
9	37.0	33.0	34
10	23.5	9.5	30
11	33.0	21.0	38
12	58.0	79.0	47

[a]Patient's height is in inches, patient's weight in pounds, and catheter length is in centimeters.

Source: Weisberg (1980, p. 218).

and moved into the heart. The catheter can be maneuvered into specific regions to provide information concerning the heart function. This technique is sometimes applied to children with congenital heart defects. The proper length of the introduced catheter has to be guessed by the physician. For 12 children, the proper catheter length (y) was determined by checking with a fluoroscope that the catheter tip had reached the right position. The aim is to describe the relation between the catheter length and the patient's height (x_1) and weight (x_2). The LS computations, as well as those for RLS, for the model

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 + e$$

are given in Table 14.

For both regressions, the F-value is large enough to conclude that $\hat{\theta}_1$ and $\hat{\theta}_2$ together contribute to the prediction of the response. Looking at the t-test for the individual regression coefficients, it follows that the LS estimate for θ_1 is not significantly different from zero at $\alpha = 5\%$. The same can be said for θ_2. However, the F-statistic says that the two x-variables viewed en bloc are important. For the RLS fit, $\hat{\theta}_1$ is not yet significantly different from zero, which means that the corresponding explanatory variable contributes little to the model. Also, the sign of $\hat{\theta}_1$ makes no sense in this context. Such phenomena typically occur in

Table 14. Heart Catheterization Data: LS and RLS Results

	LS Results			RLS Results		
Variable	$\hat{\theta}$	t-Value	p-Value	$\hat{\theta}$	t-Value	p-Value
Height	0.211	0.6099	0.5570	−0.723	−2.0194	0.0900
Weight	0.191	1.2074	0.2581	0.514	3.7150	0.0099
Constant	20.38	2.4298	0.0380	48.02	4.7929	0.0030
	$F = 21.267$ ($p = 0.0004$)			$F = 32.086$ ($p = 0.0006$)		

situations where collinearity is present. Indeed, Table 15 shows the high correlations among the variables, as computed by PROGRESS.

The correlation between height and weight is so high that either variable almost completely determines the other. Moreover, the low t-value for the regression coefficients confirms that either of the explanatory variables may be left out of the model. For the heart catheterization data, we will drop the variable weight and look at the simple regression of catheter length on height. A scatterplot of this two-dimensional data set is given in Figure 12, along with the LS and RLS fits.

The LS yields the fit $\hat{y} = 0.612x + 11.48$ (dashed line in Figure 12). The RLS fit $\hat{y} = 0.614x + 11.11$ (which now has a positive $\hat{\theta}_1$) lies close to the LS. Note that cases 5, 6, 8, 10, and 11 lie relatively far from the RLS fit. Because they are nicely balanced above and below the RLS line, the LS and RLS fits do not differ visibly for this sample.

The alternative choice would be to use weight instead of height as our predictor. A plot of measurements of catheter length versus the patient's weight is shown in Figure 13. This scatterplot suggests that a linear model

Table 15. Heart Catheterization Data: Correlations Between the Variables

	Pearson Correlation Coefficients		
Height	1.00		
Weight	0.96	1.00	
Catheter length	0.89	0.90	1.00
	Height	Weight	Catheter length
	Spearman Rank Correlation Coefficients		
Height	1.00		
Weight	0.92	1.00	
Catheter length	0.80	0.86	1.00
	Height	Weight	Catheter length

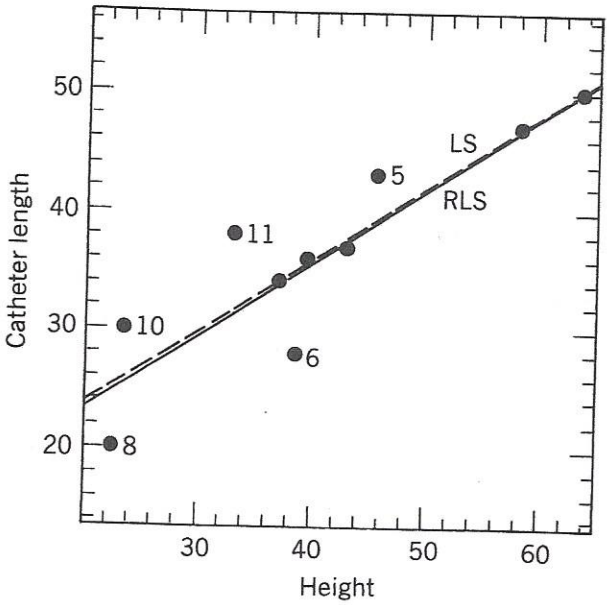


Figure 12. Heart catheterization data: Scatterplot of proper catheter length versus height for 12 children, with LS fit (dashed line) and RLS fit (solid line).

is not appropriate here. It is clear that a transformation is required. On physical grounds, one might try to use the cube root of the weight instead of the weight itself in order to obtain linearity.

Even when there is no evidence of multicollinearity, it may happen that the complete set of explanatory variables is too large for using them all in the model. In the first place, too many variables make it hard to understand the described mechanism. This should, however, be combined with the objective of explaining the variability of the response as much as possible, which leads to the consideration of more explanatory variables. But from a statistical point of view, one can sometimes say that the reduction of the number of variables improves the precision of the fit. Indeed, explanatory variables for which the associated regression coefficients are not significantly different from zero may increase the variance of the estimates. Subject-matter knowledge is sometimes sufficient to decide which of the possible equations is most appropriate. Otherwise one has to apply a statistical procedure for finding a suitable subset of the variables. This problem is called *variable selection*. It is closely linked to the problem of model specification, which we discussed above. Indeed, the problem of variable selection also includes the question, "In which form

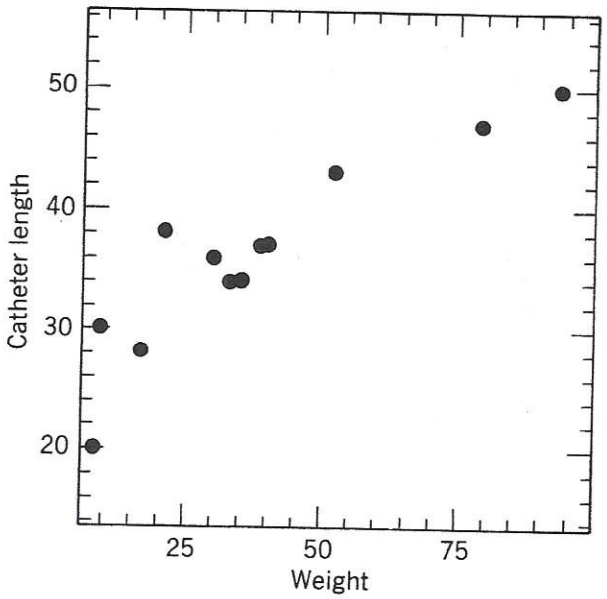


Figure 13. Heart catheterization data: Scatterplot of proper catheter length versus weight for 12 children.

should the explanatory variable enter the equation: as the original variable, or as a squared term, or as a logarithmic term, and so on?". For simplicity, one usually treats these problems separately in order to avoid the ideal, but intractable, approach consisting of a simultaneous treatment of both problems. The presence of outliers, either in y or in $\mathbf{x}$, complicates the situation even more. Therefore, we recommend starting the analysis with a high-breakdown regression on the full set of variables in order to determine the weights of the observations. Then, as a first approximation, one can use a classical technique for variable selection on the "cleaned" sample. Such variable selection techniques have been widely investigated in the literature, for instance by Aitkin (1974), Allen (1974), Diehr and Hoflin (1974), Hill and Hunter (1974), Narula and Wellington (1977, 1979), Snee (1977), Suich and Derringer (1977, 1980), Ellerton (1978), McKay (1979), Hintze (1980), Rencher and Pun (1980), Weisberg (1981), Wilkinson and Dallal (1981), Baskerville and Toogood (1982), Madsen (1982), and Young (1982). Hocking (1976) reviews the topic of variable selection. We will briefly describe the most widely used techniques.

When the set of candidate variables is not too large, one can consider

all the possible models starting from this set. This is the so-called *all-subsets regression* procedure. It leads to

$$2^{(\text{number of available explanatory variables})}$$

different models, a number that increases rapidly. Even when attention is focused on LS estimation (or our RLS, assuming that a very robust technique has been executed first), this number becomes computationally infeasible. This has motivated some people to develop more efficient algorithms, which appeal to numerical methods for calculating the LS estimates for the successive subsets. Usually, these procedures are based on either Gauss–Jordan reduction or a sweep operator (Beaton 1964, Seber 1977). Once the various fitted equations are at one's disposal, one needs a criterion for judging which subset of variables yields the "best" fit. Hocking (1976) describes different measures for this purpose, including the mean squared error, the coefficient of determination, and the C_p statistic (see also Daniel and Wood 1971 and Mallows 1973 for a thorough treatment of C_p). The Furnival and Wilson (1974) algorithm, which is available in the regression program P2R of BMDP, is a branch-and-bound type of procedure that cuts the computation time by searching only in certain promising directions.

Although it is felt that the investigation of all subsets produces the "best" set, it is not the most widely used method because of its computational cost. The so-called *stepwise procedures*, which consist of either adding or deleting one explanatory variable at a time, have been the favorite methods throughout. One distinguishes *forward selection* and *backward elimination* stepwise procedures, and a combination of both. Variations of these types have been implemented in several statistical packages, such as BMDP, SAS, and SPSS.

In *forward selection*, one starts with a simple regression model in which the explanatory variable is the variable that correlates best with the response. Then, at each subsequent step, one adds one variable to the model. At any step the selected variable x_j is the one producing the largest F_j-ratio among the candidates. This F_j-ratio is defined by

$$F_j = \frac{\text{SSE}_k - \text{SSE}_{k+(j)}}{\hat{\sigma}^2_{k+(j)}},$$

where SSE_k is the residual error sum of squares corresponding to the model with k terms, and $\text{SSE}_{k+(j)}$ is the one corresponding to the model where x_j is added. Variable x_j will be included in the equation if F_j is

larger than a prespecified value. This prespecified value is often referred to as the *stopping rule*. One can choose this value such that the procedure will run the full course. That way one obtains one subset of each size. Moreover, one can then use a criterion for selecting the "best" of these subsets.

In the *backward elimination* methods, one works in the inverse way, starting from an equation containing all the variables. At each step, one eliminates the "worst" variable. For instance, the variable x_j will be a candidate for elimination from the current model (consisting of k terms) if it produces the smallest F_j-ratio, where

$$F_j = \frac{\mathrm{SSE}_{k-(j)} - \mathrm{SSE}_k}{\hat{\sigma}_k^2}.$$

Again several stopping rules similar to those for forward selection have been suggested (see Hocking 1976). Efroymson (1960) combined both ideas: His method is basically of the forward selection type, but at each step the elimination of a variable is also possible.

Faced with a problem of variable selection, one has to be aware of the weak points of the available techniques. For instance, the stepwise procedures do not necessarily yield the "best" subset of a given size. Moreover, these techniques induce a ranking on the explanatory variables which is often misused in practice. The order of deletion or inclusion is very deceptive, because the first variable deleted in backward elimination (or similarly the first one added in forward selection) is not necessarily the worst (or the best) in an absolute sense. It may, for example, happen that the first variable entered in forward selection becomes unnecessary in the presence of other variables. Also, forward and backward stepwise techniques may lead to totally different "best" subsets of variables. Berk (1978) compares the stepwise procedures with all-subsets regression. He shows that if forward selection agrees with all-subsets regression for every subset size, then backward elimination will also agree with all-subsets regression for every subset size, and inversely. All-subsets regression, however, is not the ideal way to avoid the disadvantages of the stepwise techniques. Indeed, the evaluation of all the possible subsets also largely depends on the employed criterion. Moreover, although the all-subsets procedure produces the "best" set for each subset size, this will not necessarily be the case in the whole population. It is only the "best" in the sample. The observation made by Gorman and Toman (1966) is perhaps suitable to conclude the topic of variable selection: "It is unlikely that there is a single best subset, but rather several equally good ones."

Example 4: Education Expenditure Data

The problem of *heteroscedasticity* has already been mentioned in Section 4 of Chapter 2, in the discussion on the diagnostic power of residual plots. The following data set, described by Chatterjee and Price (1977, p. 108) provides an interesting example of heteroscedasticity. It deals with education expenditure variables for 50 U.S. states. The data are reproduced in Table 16. The y-variable in Table 16 is the per capita expenditure on public education in a state, projected for 1975. The aim is to explain y by means of the explanatory variables x_1 (number of residents per thousand residing in urban areas in 1970), x_2 (per capita personal income in 1973), and x_3(number of residents per thousand under 18 years of age in 1974).

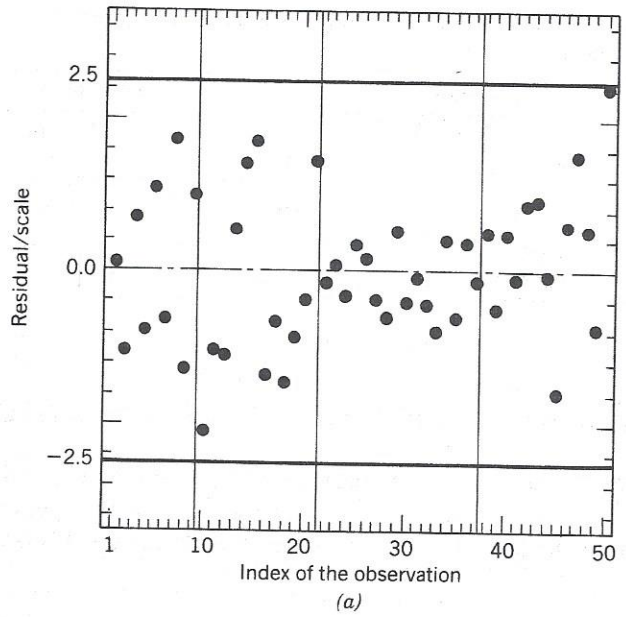
Often the index i of an observation is time-related. In that case, the pattern of an index plot may point to nonconstancy of the spread of the residuals with respect to time. However, the magnitude of the residuals may also appear to vary systematically with $\hat{y}_i$ or an explanatory variable, or with another ordering of the cases besides the time ordering. The objective for the present data set is to analyze the constancy of the relationships with regard to a spatial ordering of the cases. The data in Table 16 are grouped by geographic region. One can distinguish four groups: the northeastern states (indices 1–9), the north central states (indices 10–21), the southern states (indices 22–37), and the western states (indices 38–50).

The routine application of least squares to these data yields the coefficients in Table 17, which also contains the results of reweighted least squares based on the LMS. Let us now compare the LS index plot (Figure 14*a*) with that of RLS (Figure 14*b*). Their main difference is that the fiftieth case (Alaska) can immediately by recognized as an outlier in the RLS plot, whereas it does not stand out clearly in the LS plot. (It seems that the education expenditure in Alaska is much higher than could be expected on the basis of its population characteristics x_1, x_2, and x_3 alone.) On the other hand, both plots appear to indicate that the dispersion of the residuals changes for the different geographical regions. This phenomenon is typical for heteroscedasticity. The defect can be remedied by handling the four clusters separately, but then the number of cases becomes very limited (in this example). Chatterjee and Price (1977) analyze these data by using another type of weighted LS regression. They assign weights to each of the four regions in order to compute a weighted sum of squared residuals. These weights are estimated in a first stage by using the mean (in a certain region) of the squared residuals resulting from ordinary LS. Chatterjee and Price also considered Alaska as an outlier and decided to omit it.

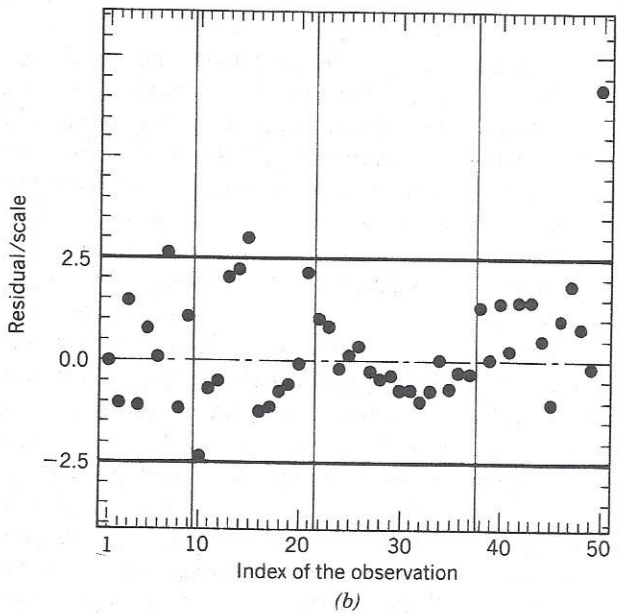
Table 16. Education Expenditure Data

Index	State		x_1	x_2	x_3	y
1	ME	Northeastern	508	3944	325	235
2	NH		564	4578	323	231
3	VT		322	4011	328	270
4	MA		846	5233	305	261
5	RI		871	4780	303	300
6	CT		774	5889	307	317
7	NY		856	5663	301	387
8	NJ		889	5759	310	285
9	PA		715	4894	300	300
10	OH	North central	753	5012	324	221
11	IN		649	4908	329	264
12	IL		830	5753	320	308
13	MI		738	5439	337	379
14	WI		659	4634	328	342
15	MN		664	4921	330	378
16	IA		572	4869	318	232
17	MO		701	4672	309	231
18	ND		443	4782	333	246
19	SD		446	4296	330	230
20	NB		615	4827	318	268
21	KS		661	5057	304	337
22	DE	Southern	722	5540	328	344
23	MD		766	5331	323	330
24	VA		631	4715	317	261
25	WV		390	3828	310	214
26	NC		450	4120	321	245
27	SC		476	3817	342	233
28	GA		603	4243	339	250
29	FL		805	4647	287	243
30	DY		523	3967	325	216
31	TN		588	3946	315	212
32	AL		584	3724	332	208
33	MS		445	3448	358	215
34	AR		500	3680	320	221
35	LA		661	3825	355	244
36	OK		680	4189	306	234
37	TX		797	4336	335	269
38	MT	Western	534	4418	335	302
39	ID		541	4323	344	268
40	WY		605	4813	331	323
41	CO		785	5046	324	304
42	NM		698	3764	366	317
43	AZ		796	4504	340	332
44	UT		804	4005	378	315
45	NV		809	5560	330	291
46	WA		726	4989	313	312
47	OR		671	4697	305	316
48	CA		909	5438	307	332
49	AK		831	5309	333	311
50	HI		484	5613	386	546

Source: Chatterjee and Price (1977).



(a)



(b)

Figure 14. Education expenditure data: (*a*) Index plot associated with LS. (*b*) Index plot associated with LMS-based RLS.

Table 17. LS and RLS Results on Education Expenditure Data: Coefficients, Standard Errors, and *t*-Values

	LS Results			RLS Results		
Variable	$\hat{\theta}$	Standard Error	*t*-Value	$\hat{\theta}$	Standard Error	*t*-Value
x_1	−0.004	0.051	−0.08	0.075	0.042	1.76
x_2	0.072	0.012	6.24	0.038	0.011	3.49
x_3	1.552	0.315	4.93	0.756	0.292	2.59
Constant	−556.6	123.2	−4.52	−197.3	117.5	−1.68

*4. PROPERTIES OF THE LMS, THE LTS, AND *S*-ESTIMATORS

This section contains some theoretical results and may be skipped by those who are only interested in the application of robust regression and not its mathematical aspects. The first part is about the LMS estimator, given by

$$\underset{\hat{\theta}}{\text{Minimize}} \underset{i}{\text{med}}\, r_i^2 . \tag{4.1}$$

The existence of this estimator will be proven, and its breakdown and exact fit properties are stated. It is also shown that it attains the maximal breakdown point among all regression equivariant estimators. Then some results on one-step *M*-estimators are presented. Finally, least trimmed squares (LTS) and *S*-estimators are covered. Most of the material in this section follows Rousseeuw (1984) and Rousseeuw and Yohai (1984).

The n observations $(\mathbf{x}_i, y_i) = (x_{i1}, \ldots, x_{ip}, y_i)$ belong to the linear space of row vectors of dimension $p+1$. The unknown parameter $\boldsymbol{\theta}$ is a p-dimensional column vector $(\theta_1, \ldots, \theta_p)^t$. The (unperturbed) linear model states that $y_i = \mathbf{x}_i\boldsymbol{\theta} + e_i$ where e_i is distributed according to $N(0, \sigma^2)$. Throughout this section it is assumed that all observations with $\mathbf{x}_i = \mathbf{0}$ have been deleted, because they give no information on $\boldsymbol{\theta}$. This condition is automatically satisfied if the model has an intercept because then the last coordinate of each $\mathbf{x}_i$ equals 1. Moreover, it is assumed that in the $(p+1)$-dimensional space of the $(\mathbf{x}_i, y_i)$, there is no vertical hyperplane through zero containing more than $[n/2]$ observations. (Such a vertical hyperplane is a p-dimensional subspace that contains $(0, \ldots, 0)$ and $(0, \ldots, 0, 1)$. We call this subspace a hyperplane because its dimension is p, which is one less than the dimension of the total space. The notation $[q]$ stands for the largest integer less than or equal to q.)

The first theorem guarantees that the minimization in (4.1) always leads to a solution.

Theorem 1. There always exists a solution to (4.1).

Proof. We work in the $(p+1)$-dimensional space E of the observations $(\mathbf{x}_i, y_i)$. The space of the $\mathbf{x}_i$ is the horizontal hyperplane through the origin, which is denoted by $(y=0)$ because the y-coordinates of all points in this plane are zero. Two cases have to be considered:

CASE A. This is really a special case, in which there exists a $(p-1)$-dimensional subspace V of $(y=0)$ going through zero and containing at least $[n/2]+1$ of the $\mathbf{x}_i$. The observations $(\mathbf{x}_i, y_i)$ corresponding to these $\mathbf{x}_i$ now generate a subspace S of E (in the sense of linear algebra), which is at most p-dimensional. Because it was assumed that E has no vertical hyperplane containing $[n/2]+1$ observations, it follows that S does not contain $(0, \ldots, 0, 1)$; hence the dimension of S is at most $p-1$. This means that there exists a nonvertical hyperplane H given by some equation $y = \mathbf{x}\boldsymbol{\theta}$ which includes S. For this value of $\boldsymbol{\theta}$, clearly $\operatorname{med}_i r_i^2 = 0$, which is the minimal value. This reasoning can be illustrated by taking the value of p equal to 2 and considering a linear model without intercept term. Figure 15 illustrates the positions of the subspaces S and V of E.

CASE B. Let us now assume that we are in the general situation in which case A does not hold. The rest of the proof will be devoted to showing that there exists a ball around the origin in the space of all $\boldsymbol{\theta}$, to which attention can be restricted for finding a minimum of $\operatorname{med}_i r_i^2(\boldsymbol{\theta})$. Because the objective function $\operatorname{med}_i r_i^2(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, this is sufficient for the existence of a minimum. Put

$$\delta = \tfrac{1}{2} \inf \{\tau > 0;\ \text{there exists a } (p-1)\text{-dimensional subspace } V \text{ of } (y=0) \text{ such that } V^\tau \text{ covers at least } [n/2]+1 \text{ of the } \mathbf{x}_i\},$$

where V^τ is the set of all $\mathbf{x}$ with distance to V not larger than τ. Case A corresponds to $\delta = 0$, but now $\delta > 0$. Denote $M := \max_i |y_i|$. Now attention may be restricted to the closed ball around the origin with radius $(\sqrt{2}+1)M/\delta$. Indeed, for any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| > (\sqrt{2}+1)M/\delta$, it will be shown that

$$\operatorname*{med}_i r_i^2(\boldsymbol{\theta}) > \operatorname*{med}_i y_i^2 = \operatorname*{med}_i r_i^2(\mathbf{0}),$$

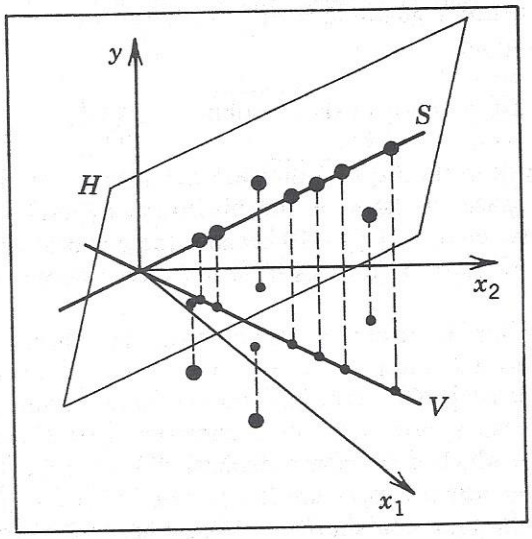


Figure 15. Illustration of the position of the subspaces S and V in the space E, as defined in the proof of Theorem 1 (Case A). There are 10 observations and $p = 2$.

so smaller objective functions cannot be found outside the ball. A geometrical construction (illustrated in Figure 16) is needed to prove this. Such a $\boldsymbol{\theta}$ determines a nonvertical hyperplane H given by $y = \mathbf{x}\boldsymbol{\theta}$. By the dimension theorem of linear algebra,

$$\begin{aligned}\dim(H \cap (y=0)) &= \dim(H) + \dim(y=0) - \dim(H + (y=0)) \\ &= p + p - (p+1) \\ &= p - 1\end{aligned}$$

because $\|\boldsymbol{\theta}\| > (\sqrt{2}+1)M/\delta$ implies that $\boldsymbol{\theta} \neq \mathbf{0}$ and hence $H \neq (y=0)$. Therefore, $(H \cap (y=0))^{\delta}$ contains at most $[n/2]$ of the $\mathbf{x}_i$. For each of the remaining observations $(\mathbf{x}_i, y_i)$, we construct the vertical two-dimensional plane P_i through $(\mathbf{x}_i, y_i)$, which is orthogonal to $(H \cap (y=0))$. (This plane does not pass through zero, so to be called vertical, it has to go through both $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_i, y_i + 1)$.) We see that

$$|r_i| = |\mathbf{x}_i\boldsymbol{\theta} - y_i| \geq \big|\,|\mathbf{x}_i\boldsymbol{\theta}| - |y_i|\,\big|$$

with $|\mathbf{x}_i\boldsymbol{\theta}| > \delta|\tan(\alpha)|$, where α is the angle in $(-\pi/2, \pi/2)$ formed by H and the horizontal line in P_i. Therefore $|\alpha|$ is the angle between the line

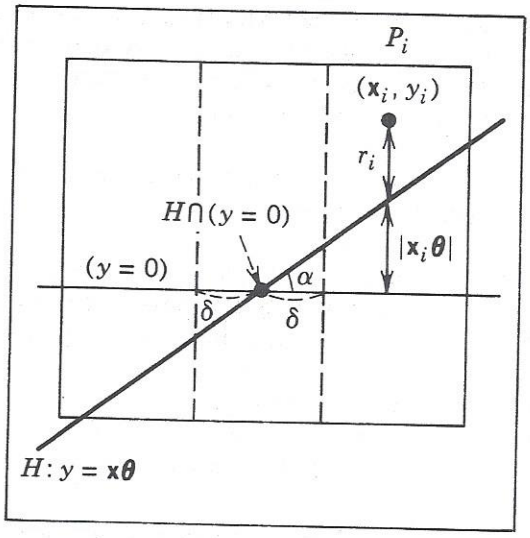


Figure 16. Illustration of a geometrical construction in the proof of Theorem 1 (Case B).

orthogonal to H and $(\mathbf{0}, 1)$, hence

$$|\alpha| = \arccos\left\{\frac{|(-\boldsymbol{\theta}, 1)(\mathbf{0}, 1)'|}{\|(-\boldsymbol{\theta}, 1)\|\|(\mathbf{0}, 1)\|}\right\} = \arccos\left\{\frac{1}{\sqrt{1 + \|\boldsymbol{\theta}\|^2}}\right\}$$

and finally $|\tan(\alpha)| = \|\boldsymbol{\theta}\|$.

Because $\|\boldsymbol{\theta}\| > (\sqrt{2} + 1)M/\delta$, it follows that

$$|\mathbf{x}_i\boldsymbol{\theta}| > \delta\|\boldsymbol{\theta}\| > M \geq |y_i|$$

so

$$|r_i(\boldsymbol{\theta})| > (\delta\|\boldsymbol{\theta}\| - |y_i|)\,.$$

But then

$$\begin{aligned} r_i^2(\boldsymbol{\theta}) &> ((\sqrt{2} + 1)M - |y_i|)^2 \\ &> ((\sqrt{2} + 1)M - M)^2 \\ &> 2M^2 \end{aligned}$$

for at least $n - [n/2]$ observations. Hence

$$\underset{i}{\text{med}}\, r_i^2(\boldsymbol{\theta}) > M^2 \geq \underset{i}{\text{med}}\,(y_i^2)\,.$$

So the objective function associated with such a $\boldsymbol{\theta}$ is larger than the one for $\boldsymbol{\theta} = \mathbf{0}$. Therefore, we only have to search for a solution $\boldsymbol{\theta}$ in the closed ball $B(\mathbf{0}, (\sqrt{2}+1)M/\delta)$. Because this set is compact and $\operatorname{med}_i r_i^2(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, the infimum is a minimum. □

REMARK. This proof is not constructive. To actually find a solution to (4.1) we use the algorithm described in Chapter 5.

Let us now discuss some equivariance properties of the LMS. For regression estimators, one can consider three types of equivariance. Ranked from higher to lower priority, there exists regression, scale, and affine equivariance.

An estimator T is called *regression equivariant* if

$$T(\{(\mathbf{x}_i,\ y_i + \mathbf{x}_i\mathbf{v});\ i = 1, \ldots, n\}) = T(\{(\mathbf{x}_i,\ y_i);\ i = 1, \ldots, n\}) + \mathbf{v}\,, \quad (4.2)$$

where $\mathbf{v}$ is any column vector. Regression equivariance is just as crucial as translation equivariance for a multivariate location estimator, but not as often formulated. It is implicit in the notion of a regression estimator. For instance, many proofs of asymptotic properties or descriptions of Monte Carlo studies begin with the phrase "without loss of generality, let $\boldsymbol{\theta} = \mathbf{0}$", which assumes that the results are valid at any parameter vector through application of (4.2). On the other hand, note that the coefficient of determination (R^2) is *not* regression invariant, because it depends on the inclination of the regression surface (Barrett 1974).

An estimator T is said to be *scale equivariant* if

$$T(\{(\mathbf{x}_i,\ cy_i);\ i = 1, \ldots, n\}) = cT(\{(\mathbf{x}_i,\ y_i);\ i = 1, \ldots, n\}) \quad (4.3)$$

for any constant c. It implies that the fit is essentially independent of the choice of measurement unit for the response variable y.

One says that T is *affine equivariant* if

$$T(\{(\mathbf{x}_i\mathbf{A},\ y_i);\ i = 1, \ldots, n\}) = \mathbf{A}^{-1}T(\{(\mathbf{x}_i,\ y_i);\ i = 1, \ldots, n\}) \quad (4.4)$$

for any nonsingular square matrix $\mathbf{A}$. In words, affine equivariance means that a linear transformation of the $\mathbf{x}_i$ should transform the estimator T accordingly, because $\hat{y}_i = \mathbf{x}_i T = (\mathbf{x}_i\mathbf{A})(\mathbf{A}^{-1}T)$. This allows us to use another coordinate system for the explanatory variables, without affecting the estimated $\hat{y}_i$.

The LMS satisfies all three equivariance properties: