interactive input. The program then picks out the response and the explanatory variables for the analysis.

```
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? KEY
```

Moreover, PROGRESS enables the user to store the data (in case KEY has been answered) by means of the following dialogue:

```
DO YOU WANT TO SAVE YOUR DATA IN A FILE?
PLEASE ANSWER YES OR NO: YES

IN WHICH FILE DO YOU WANT TO SAVE YOUR DATA?
(WARNING: IF THERE ALREADY EXISTS A FILE WITH THE SAME NAME,
          THEN THE OLD FILE WILL BE OVERWRITTEN.)
TYPE e.g. B:SAVE.DAT : B:PILOT.DAT
```

The whole data set will be stored, even those variables that are not used right now. This enables the user to perform another analysis afterwards, with a different combination of variables.

Depending on the answer to the following question, the output provided by PROGRESS will be written on the screen, on paper, or in a file.

```
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,
          THEN THE OLD FILE WILL BE OVERWRITTEN.)
WHAT DO YOU CHOOSE? PRN
```

We would like to give the user a warning concerning the latter two questions. The name of a DOS file is unique. This means that if the user enters a name of a file that already exists on the diskette, the old file will be overwritten by the new file.

The plots constructed by PROGRESS are intended for a printer using 8 lines per inch. (Consequently, on the screen these plots are slightly stretched out.) It is therefore recommended to adapt the printer to 8 lines per inch. For instance, this can be achieved by typing the DOS command "MODE LPT1:80,8" before running PROGRESS.

Next, PROGRESS requests a title, which will be reproduced on the output. This title should consist of at most 60 characters. When the user

enters more characters, only the first 60 will be read:

```
PLEASE ENTER A TITLE FOR THE OUTPUT (AT MOST 60 CHARACTERS):
 ------------------------------------------------------------
PILOT-PLANT DATA SET WITH ONE LEVERAGE POINT
```

The answer to the following question tells PROGRESS the way in which the data has to be read. Two possibilities are available. The first consists of reading the data in free format, which is performed by answering YES to:

```
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
```

In order to use the free format, it suffices that the variables for each case be separated by at least one blank. On the other hand, when the user answers NO to the above question, PROGRESS requests the FORTRAN format to be used to input the data. The program expects the format necessary for reading the total number of variables of the data set (in this case 5). The program will then select the variables for actual use by means of the positions chosen above. The FORTRAN format has to be set between brackets, and it should be described in at most 60 characters (including the brackets). The observations are to be processed as real numbers, so they should be read in F-formats and/or E-formats. The formats X and / are also allowed.

Because the execution time for large data sets may be quite long, the user has the option of choosing a faster version of the algorithm in that case. In other cases it is recommended to use the extensive search version because of its greater precision. (More details about the algorithm will be provided in Chapter 5.)

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

PROGRESS also allows the user to deal with missing values. However, we shall postpone the discussion of these options until Section 2 of Chapter 3.

```
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: 0
```

Finally, PROGRESS gives a survey of the options that were selected.

```
************************************************
*  PROGRESS WILL PERFORM A REGRESSION WITH CONSTANT TERM  *
************************************************

THE NUMBER OF CASES EQUALS                    20
THE NUMBER OF EXPLANATORY VARIABLES EQUALS     1
TITRATION IS THE RESPONSE VARIABLE.
THE DATA WILL BE READ FROM THE KEYBOARD.
THE DATA WILL BE SAVED IN FILE: B:PILOT.DAT
TITLE FOR OUTPUT: PILOT-PLANT DATA SET WITH ONE LEVERAGE POINT
THE DATA WILL BE READ IN FREE FORMAT.
LARGE OUTPUT IS WANTED.
NO RESIDUAL PLOTS ARE WANTED.
THE EXTENSIVE SEARCH VERSION WILL BE USED.
THERE ARE NO MISSING VALUES.
YOUR OUTPUT WILL BE WRITTEN ON: PRN

ARE ALL THESE OPTIONS OK? YES OR NO: YES
```

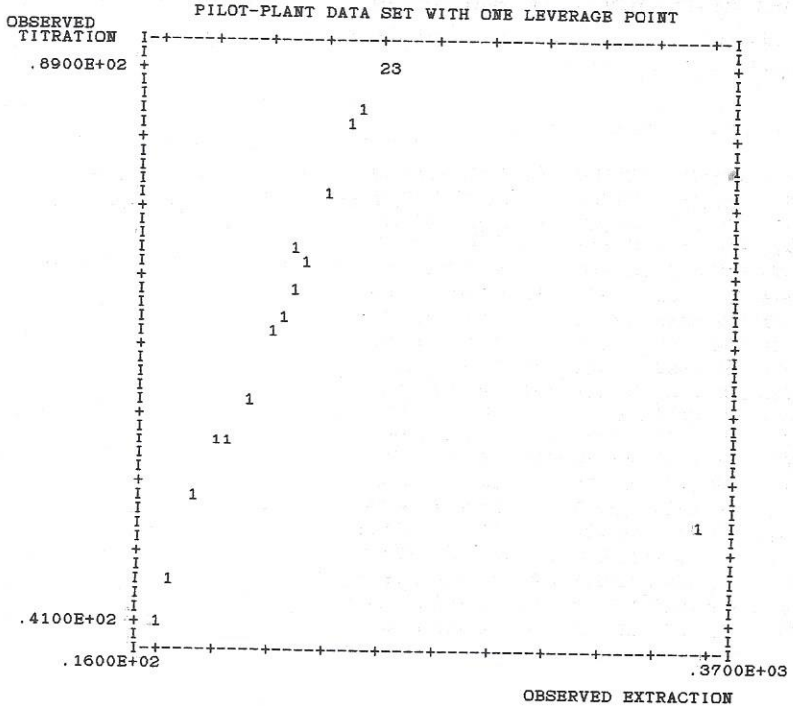
When the data have to be read from the keyboard, the user has to type the measurements for each case. For the example we are working with, this would look as follows:

```
ENTER YOUR DATA FOR EACH CASE.

THE DATA FOR CASE NUMBER  1:  1 123 0 76 28
THE DATA FOR CASE NUMBER  2:  2 109 0 70 23
THE DATA FOR CASE NUMBER  3:  3 62 1 55 29
THE DATA FOR CASE NUMBER  4:  4 104 1 71 28
THE DATA FOR CASE NUMBER  5:  5 57 0 55 27
THE DATA FOR CASE NUMBER  6:  6 370 0 48 35
THE DATA FOR CASE NUMBER  7:  7 44 1 50 24
THE DATA FOR CASE NUMBER  8:  8 100 1 66 23
THE DATA FOR CASE NUMBER  9:  9 16 0 41 27
THE DATA FOR CASE NUMBER 10:  10 28 1 43 29
THE DATA FOR CASE NUMBER 11:  11 138 0 82 21
THE DATA FOR CASE NUMBER 12:  12 105 0 68 28
THE DATA FOR CASE NUMBER 13:  13 159 1 88 24
THE DATA FOR CASE NUMBER 14:  14 75 1 58 26
THE DATA FOR CASE NUMBER 15:  15 88 0 64 26
THE DATA FOR CASE NUMBER 16:  16 164 0 88 26
THE DATA FOR CASE NUMBER 17:  17 169 1 89 23
THE DATA FOR CASE NUMBER 18:  18 167 1 88 36
THE DATA FOR CASE NUMBER 19:  19 149 0 84 24
THE DATA FOR CASE NUMBER 20:  20 167 1 88 21
```

Out of these five variables, we only use two. (In fact, in this example the other three variables are artificial, because they were added to illustrate the program.)

The output file corresponding to this example is printed below, and is followed by a discussion of the results in Section 3.

```
*******************************************
* ROBUST REGRESSION WITH A CONSTANT TERM. *
*******************************************

NUMBER OF CASES =     20
NUMBER OF COEFFICIENTS (INCLUDING CONSTANT TERM) =      2

THE EXTENSIVE SEARCH VERSION WILL BE USED.
DATA SET =          PILOT-PLANT DATA SET WITH ONE LEVERAGE POINT

THERE ARE NO MISSING VALUES.

LARGE OUTPUT IS WANTED.

THE OBSERVATIONS ARE:

      EXTRACTION  TITRATION
   1    123.0000    76.0000
   2    109.0000    70.0000
   3     62.0000    55.0000
   4    104.0000    71.0000
   5     57.0000    55.0000
   6    370.0000    48.0000
   7     44.0000    50.0000
   8    100.0000    66.0000
   9     16.0000    41.0000
  10     28.0000    43.0000
  11    138.0000    82.0000
  12    105.0000    68.0000
  13    159.0000    88.0000
  14     75.0000    58.0000
  15     88.0000    64.0000
  16    164.0000    88.0000
  17    169.0000    89.0000
  18    167.0000    88.0000
  19    149.0000    84.0000
  20    167.0000    88.0000
```



MEDIANS =

EXTRACTION	TITRATION
107.0000	69.0000

DISPERSIONS =

EXTRACTION	TITRATION
70.4235	21.4977

THE STANDARDIZED OBSERVATIONS ARE:

	EXTRACTION	TITRATION
1	.2272	.3256
2	.0284	.0465
3	-.6390	-.6512
4	-.0426	.0930
5	-.7100	-.6512
6	3.7345	-.9768
7	-.8946	-.8838
8	-.0994	-.1395
9	-1.2922	-1.3025
10	-1.1218	-1.2094
11	.4402	.6047
12	-.0284	-.0465
13	.7384	.8838
14	-.4544	-.5117
15	-.2698	-.2326
16	.8094	.8838
17	.8804	.9303
18	.8520	.8838
19	.5964	.6977
20	.8520	.8838

PEARSON CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(TITRATION IS THE RESPONSE VARIABLE)

EXTRACTION	1.00	
TITRATION	.38	1.00

SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(TITRATION IS THE RESPONSE VARIABLE)

EXTRACTION	1.00	
TITRATION	.76	1.00

****

LEAST SQUARES REGRESSION
****

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
EXTRACTION	.08071	.04695	1.71914	.10274
CONSTANT	58.93883	6.61420	8.91096	.00000

SUM OF SQUARES = 4379.69300

DEGREES OF FREEDOM = 18

SCALE ESTIMATE = 15.59860

VARIANCE - COVARIANCE MATRIX =

.2204D-02	
-.2638D+00	.4375D+02

COEFFICIENT OF DETERMINATION (R SQUARED) = .14103

THE F-VALUE = 2.955 (WITH 1 AND 18 DF) P - VALUE = .10274

OBSERVED TITRATION	ESTIMATED TITRATION	RESIDUAL	NO	RES/SC
76.00000	68.86635	7.13365	1	.46
70.00000	67.73638	2.26362	2	.15
55.00000	63.94294	-8.94294	3	-.57
71.00000	67.33282	3.66718	4	.24
55.00000	63.53938	-8.53938	5	-.55
48.00000	88.80209	-40.80209	6	-2.62
50.00000	62.49014	-12.49014	7	-.80
66.00000	67.00998	-1.00998	8	-.06
41.00000	60.23021	-19.23021	9	-1.23
43.00000	61.19875	-18.19875	10	-1.17
82.00000	70.07702	11.92298	11	.76
68.00000	67.41354	.58646	12	.04
88.00000	71.77196	16.22804	13	1.04
58.00000	64.99220	-6.99220	14	-.45
64.00000	66.04144	-2.04144	15	-.13
88.00000	72.17552	15.82448	16	1.01
89.00000	72.57908	16.42092	17	1.05
88.00000	72.41766	15.58234	18	1.00
84.00000	70.96484	13.03516	19	.84
88.00000	72.41766	15.58234	20	1.00

****

LEAST MEDIAN OF SQUARES REGRESSION
****

VARIABLE	COEFFICIENT
EXTRACTION	.31429
CONSTANT	36.34286

FINAL SCALE ESTIMATE = 1.33279

COEFFICIENT OF DETERMINATION = .99641

OBSERVED TITRATION	ESTIMATED TITRATION	RESIDUAL	NO	RES/SC
76.00000	75.00000	1.00000	1	.75
70.00000	70.60000	-.60000	2	-.45
55.00000	55.82857	-.82857	3	-.62
71.00000	69.02856	1.97144	4	1.48
55.00000	54.25714	.74286	5	.56
48.00000	152.62860	-104.62860	6	-78.50
50.00000	50.17143	-.17143	7	-.13
66.00000	67.77142	-1.77142	8	-1.33
41.00000	41.37143	-.37143	9	-.28
43.00000	45.14286	-2.14286	10	-1.61
82.00000	79.71428	2.28572	11	1.71
68.00000	69.34285	-1.34285	12	-1.01
88.00000	86.31429	1.68571	13	1.26
58.00000	59.91428	-1.91428	14	-1.44
64.00000	64.00000	.00000	15	.00
88.00000	87.88571	.11429	16	.09
89.00000	89.45714	-.45714	17	-.34
88.00000	88.82857	-.82857	18	-.62
84.00000	83.17142	.82858	19	.62
88.00000	88.82857	-.82857	20	-.62

****

REWEIGHTED LEAST SQUARES BASED ON THE LMS
****

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
EXTRACTION	.32261	.00595	54.21467	.00000
CONSTANT	35.31744	.69617	50.73091	.00000

WEIGHTED SUM OF SQUARES = 26.75224

DEGREES OF FREEDOM = 17

SCALE ESTIMATE = 1.25446

VARIANCE - COVARIANCE MATRIX =

.3541D-04
-.3772D-02 .4847D+00

COEFFICIENT OF DETERMINATION (R SQUARED) = .99425

THE F-VALUE = 2939.231 (WITH 1 AND 17 DF) P - VALUE = .00000

THERE ARE 19 POINTS WITH NON-ZERO WEIGHT.

AVERAGE WEIGHT = .95000

OBSERVED TITRATION	ESTIMATED TITRATION	RESIDUAL	NO	RES/SC	WEIGHT
76.00000	74.99884	1.00116	1	.80	1.0
70.00000	70.48225	-.48225	2	-.38	1.0
55.00000	55.31945	-.31945	3	-.25	1.0
71.00000	68.86919	2.13081	4	1.70	1.0
55.00000	53.70638	1.29362	5	1.03	1.0
48.00000	154.68420	-106.68420	6	-85.04	1.0
50.00000	49.51241	.48759	7	.39	1.0
66.00000	67.57874	-1.57874	8	-1.26	1.0
41.00000	40.47925	.52075	9	.42	1.0
43.00000	44.35061	-1.35061	10	-1.08	1.0
82.00000	79.83803	2.16197	11	1.72	1.0
68.00000	69.19180	-1.19180	12	-.95	1.0
88.00000	86.61290	1.38710	13	1.11	1.0
58.00000	59.51341	-1.51341	14	-1.21	1.0
64.00000	63.70738	.29262	15	.23	1.0
88.00000	88.22597	-.22597	16	-.18	1.0
89.00000	89.83904	-.83904	17	-.67	1.0
88.00000	89.19380	-1.19380	18	-.95	1.0
84.00000	83.38677	.61323	19	.49	1.0
88.00000	89.19380	-1.19380	20	-.95	1.0

3. INTERPRETATION OF THE RESULTS

The output provided by PROGRESS starts with some general information about the data set. In the above example, the data were two-dimensional so they could be plotted. A point in the scattergram is represented by a digit. This digit corresponds to the number of points having approximately the same coordinates. When more than nine points coincide, an asterisk (*) is printed in that position. In simple regression, such a plot reveals immediately which points may exert a strong influence on the LS estimates.

The program then prints the median m_j of each variable j. In the Pilot-Plant output displayed above, the median extraction value is 107 and the median titration value equals 69. On the next line, the dispersion s_j of each variable is given, which can be considered as a robust version of its standard deviation (the exact definition will be given in Section 1 of Chapter 4). When large output has been requested, the program then provides a list of the standardized observations, in which each measurement x_{ij} is replaced by

$$\frac{x_{ij} - m_j}{s_j} \tag{3.1}$$

(the response variable is standardized in the same way). The columns of the resulting table each have a median of 0 and a dispersion of 1. This enables us to identify outliers in any single variable. Indeed, if the absolute value of a standardized observation is large (say, larger than 2.5) then it is an outlier for that particular variable. In the Pilot-Plant output we discover in this way that case 6 has an unusually large extraction measurement, because its standardized value is 3.73. (Sometimes the standardized values can tell us that the distribution in a column is skewed, thereby suggesting data transformation.) The standardization is performed differently when there is no constant term in the regression, as we shall see in Section 6.

Next, PROGRESS provides the Pearson (product-moment) correlation coefficients between the variables, as well as the nonparametric Spearman correlation coefficients.

Before giving the robust estimates, PROGRESS starts with the classical LS results. A table is printed with the estimated coefficients, along with their standard error. The standard error of the jth LS regression coefficient is the square root of the jth diagonal element of the variance–covariance matrix of the LS regression coefficients, given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$,

where the matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

The ith row of $\mathbf{X}$ equals the vector $\mathbf{x}_i$ consisting of the p explanatory variables of the ith observation. The unknown σ^2 is estimated by

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2 , \tag{3.2}$$

where $r_i = y_i - \hat{y}_i$ is the ith residual. The estimated variance–covariance matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is also contained in the output. (Because this matrix is symmetric, only its lower triangular part is given, including the main diagonal.) For simple regression, $p = 1$ when there is no intercept term, and $p = 2$ otherwise. In the output, the quantity indicated by SCALE ESTIMATE is $s = \sqrt{s^2}$.

To construct confidence intervals for the parameters θ_j, one has to assume that the errors e_i are independently normally distributed with mean zero and variance σ^2. Under these conditions, it is well known that each of the quantities

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{jj}}} , \qquad j = 1, \ldots, p \tag{3.3}$$

has a Student distribution with $n - p$ degrees of freedom. Let us denote the $1 - \alpha/2$ quantile of this distribution by $t_{n-p,1-\alpha/2}$. Then a $(1-\alpha) \times 100\%$ confidence interval for θ_j is given by

$$[\hat{\theta}_j - t_{n-p,1-\alpha/2}\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{jj}},\ \hat{\theta}_j + t_{n-p,1-\alpha/2}\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{jj}}] \tag{3.4}$$

for each $j = 1, \ldots, p$. The same result can also be used for testing the significance of any regression coefficient, such as $\hat{\theta}_j$. The hypotheses are

$$\begin{aligned} H_0&: \quad \theta_j = 0 \quad \text{(null hypothesis)} \\ H_1&: \quad \theta_j \neq 0 \quad \text{(alternative hypothesis)} . \end{aligned} \tag{3.5}$$

Such a test may be helpful in determining if the jth variable might be deleted from the model. If the null hypothesis in (3.5) is accepted (for a certain value of α), then this indicates that the jth explanatory variable does not contribute much to the explanation of the response variable. The test statistic for this hypothesis is

$$\frac{\hat{\theta}_j}{\sqrt{s^2((\mathbf{X}^t\mathbf{X})^{-1})_{jj}}}, \qquad j = 1, \ldots, p. \tag{3.6}$$

This ratio corresponds to "T-VALUE" in the output of PROGRESS. The null hypothesis will be rejected at the level α when the absolute value of "T-VALUE" is larger than $t_{n-p,1-\alpha/2}$ (in this case we will say that the jth regression coefficient is significantly different from zero).

In the present example, $n = 20$ and $p = 2$. The 97.5% quantile (since we use $\alpha = 5\%$) of the Student distribution with $n - p = 18$ degrees of freedom equals 2.101. Therefore, one can conclude that the LS slope is not significantly different from zero, because its associated t-value equals 1.719. On the other hand, the intercept t-value equals 8.911, which is quite significant.

Next to each t-value, PROGRESS also prints the corresponding (two-sided) "p-value" or "significance level." This is the probability that a Student-distributed random variable with $n - p$ degrees of freedom becomes larger in absolute value than the t-value that was actually obtained. In order to compute this probability, we used the exact formulas of Lackritz (1984) as they were implemented by van Soest and van Zomeren (1986, personal communication). When the p-value is smaller than 0.05, then the corresponding regression coefficient is significant at the 5% level. In the above output, the p-value of the LS slope equals 0.10274 so it is not significant, whereas the p-value of the LS intercept is given as 0.00000, which makes it significant (even at the 0.1% level). The printed p-values make hypothesis testing very easy, because probability tables are no longer necessary.

However, (3.6) has to be used with caution because it is not robust at all. The distribution theory of this statistic only holds when the errors really follow a Gaussian distribution, which is rarely fulfilled in practice. It is therefore advisable to look first at the robust fit in order to be aware of the possible presence of outliers. In Section 4, we will give some guidelines based on the residuals resulting from the robust fit for identifying the harmful observations. When it appears that the data set contains influential points, then one has to resort to the t-values of the RLS solution, which will be presented below.

In order to obtain an idea of the strength of the linear relationship between the response variable and the explanatory variable(s), the coefficient of determination (R^2) is displayed in the output. R^2 measures the proportion of total variability explained by the regression. For the exact formula one has to distinguish between regression with and without a constant term. For LS regression, R^2 can be calculated as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \quad \text{in the model without constant term}$$

and as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}_m} \quad \text{in the model with constant term,}$$

where

$$\text{SSE} = \text{residual error sum of squares}$$
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 ,$$
$$\text{SST} = \text{total sum of squares}$$
$$= \sum_{i=1}^{n} y_i^2 ,$$

and

$$\text{SST}_m = \text{total sum of squares corrected for the mean}$$
$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 ,$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

In the case of simple regression with constant term, the coefficient of determination equals the square of the Pearson correlation coefficient between x and y, which explains the notation R^2.

For the Pilot-Plant data (with outlier), the R^2 value corresponding to LS equals 0.141. This means that only 14.1% of the variability in y is explained by the simple regression model.

One can also consider testing the hypothesis that R^2 equals zero. Formally,

$$\begin{array}{lll} H_0: & R^2 = 0 & \text{(null hypothesis)} \\ H_1: & R^2 \neq 0 & \text{(alternative hypothesis)}. \end{array} \tag{3.7}$$

These hypotheses are equivalent to testing whether the whole vector of regression coefficients (except for the constant term if the model has one) equals the zero vector, that is, (3.7) is equivalent to

$$\begin{array}{ll} H_0: & \text{All nonintercept } \theta_j\text{'s are together equal to zero} \\ H_1: & H_0 \text{ is not true}. \end{array} \tag{3.8}$$

This is quite different from the above t-test on an individual regression coefficient, because here the coefficients are considered together. If the e_i are normally distributed, then the following statistics have an F-distribution:

$$\frac{R^2/(p-1)}{(1-R^2)/(n-p)} = \frac{(\mathrm{SST}_m - \mathrm{SSE})/(p-1)}{\mathrm{SSE}/(n-p)} \sim F_{p-1,n-p}$$

for regression with a constant, and

$$\frac{R^2/p}{(1-R^2)/(n-p)} = \frac{(\mathrm{SST} - \mathrm{SSE})/p}{\mathrm{SSE}/(n-p)} \sim F_{p,n-p} \quad \text{otherwise}.$$

If the calculated value of the appropriate statistic is less than the $(1-\alpha)$th quantile of the associated F-distribution, then H_0 can be accepted. If not, H_0 may be rejected. To facilitate this test, PROGRESS prints the p-value of the F-statistic. This p-value (again computed according to Lackritz 1984) is the probability that an F-distributed random variable with the proper degrees of freedom exceeds the actual F-value.

For the contaminated Pilot-Plant data we obtain an F-value of 2.955, with 1 and 18 degrees of freedom. The corresponding p-value is 0.103, so we have no significance at the 5% level. Consequently, one can say that it appears from the LS estimates that the explanatory variable does not really "explain" the response in a significant way, since one cannot reject H_0.

The interpretation of t-test, R^2, and F-test is still valid for the multidimensional data sets that will be considered in Chapter 3.

When intermediate or large output is requested, the program continues by listing, for each observation, the actual value of the response

(y_i), the estimated response ($\hat{y}_i$), the residual ($r_i = y_i - \hat{y}_i$), and the standardized residual (r_i/s). This table allows us to identify outlying residuals. However, because case 6 has attracted the LS line so strongly (see Figure 2), and has also blown up s, its standardized residual is merely -2.62, which is perhaps slightly conspicuous but in no way dramatic.

The results for LMS regression are printed below the output for LS. First the estimates of the regression parameters are given, together with a corresponding scale estimate. This scale estimate is also defined in a robust way. For that purpose a preliminary scale estimate s^0 is calculated, based on the value of the objective function, multiplied by a finite sample correction factor dependent on n and p:

$$s^0 = 1.4826\left(1 + \frac{5}{n-p}\right)\sqrt{\underset{i}{\text{med}}\, r_i^2}\,.$$

With this scale estimate, the standardized residuals r_i/s^0 are computed and used to determine a weight w_i for the ith observation as follows:

$$w_i = \begin{cases} 1 & \text{if } |r_i/s^0| \le 2.5 \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.9}$$

The scale estimate for the LMS regression is then given by

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i - p}}\,. \tag{3.10}$$

Note that σ^* also has a 50% breakdown point, which means that it does not explode ($\sigma^* \to \infty$) or implode ($\sigma^* \to 0$) for less than 50% of contamination. More details on this scale estimate will be given in Chapter 5. In the present example, $\sigma^* = 1.33$, which is much smaller than the LS scale $s = 15.6$ computed in the beginning.

The LMS also possesses a measure to determine how well the fitted model explains the observed variability in y. In analogy to the classical one, we called it also R^2 or coefficient of determination. In the case of regression with constant term, it is defined by

$$R^2 = 1 - \left(\frac{\text{med}\,|r_i|}{\text{mad}\,(y_i)}\right)^2 \tag{3.11}$$

and by

$$R^2 = 1 - \left(\frac{\text{med}\,|r_i|}{\text{med}\,|y_i|} \right)^2 \tag{3.12}$$

when the model has no intercept term. Here, the abbreviation "mad" stands for *median absolute deviation*, defined as

$$\text{mad}\,(y_i) = \underset{i}{\text{med}}\,\{|y_i - \underset{j}{\text{med}}\, y_j|\}\,.$$

Independent of our work, formula (3.11) was also proposed by Kvalseth (1985). In the Pilot-Plant output the robust coefficient of determination equals 0.996, which means that the majority of the data fits a linear model quite nicely.

Also for the LMS, a table with observed y_i, estimated $\hat{y}_i$, residual r_i, and standardized residual r_i/σ^* is given. It now shows clearly that case 6 is an outlier, because r_6/σ^* equals the enormous value of -78.50.

The last part of the output is about reweighted least squares (RLS) regression. This corresponds to minimizing the sum of the squared residuals multiplied by a weight w_i:

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} w_i r_i^2\,. \tag{3.13}$$

The weights w_i are determined from the LMS solution as in (3.9), but with the final scale estimate σ^* (3.10) instead of s^0. The effect of the weights, which can only take the values 0 or 1, is the same as deleting the cases for which w_i equals zero. The scale estimate associated with (3.13) is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i - p}}\,. \tag{3.14}$$

(One finds again the scale estimate for ordinary LS when putting $w_i = 1$ for all cases.) Therefore, the RLS can be seen as ordinary LS on a "reduced" data set, consisting of only those observations that received a nonzero weight. Because this reduced data set does not contain regression outliers anymore, the statistics and inferences are more trustworthy than those associated with LS on the whole data set. The underlying

distribution theory is no longer entirely exact (because the weights depend on the data in a complicated way), but is still useful as a good approximation, as was confirmed by means of some Monte Carlo trials.

In the present example all w_i are equal to 1, except for case 6, which indeed was the outlier we had produced. The regression coefficients obtained by RLS strongly resemble those determined in the first step by LMS. Note that, without the outlier, the slope estimate becomes significantly different from zero.

The determination coefficient for the RLS is defined in an analogous way as for LS, but all terms are now multiplied by their weight w_i. In this example it is highly significant, because the F-value becomes very large and hence the corresponding p-value is close to zero.

To end this section, we would like to warn the reader about a common misunderstanding. When the LS and the RLS results are substantially different, the right thing to do is to identify the outliers (by means of the RLS residuals) and to study them. Instead, some people are inclined to think they have to *choose* between the LS and the RLS output, and typically they will prefer the estimates with the most significant t-values or F-value, often assuming that the highest R^2 corresponds to the "best" regression. This makes no sense, because the LS inference is very sensitive to outliers, which may affect R^2 in *both* directions. Indeed, the least squares R^2 of any data set can be made arbitrarily close to 1 by means of one or more leverage points. It often happens that RLS discards the outliers that were responsible for a high R^2 and correctly comes to the conclusion that the R^2 of the majority is not so high at all (or it may find that some θ_j are no longer significantly different from zero).

4. EXAMPLES

In this section we will further explain the results provided by PROGRESS, by means of some real-data examples appearing in the literature.

Example 1: First Word—Gesell Adaptive Score Data

This two-dimensional data set comes from Mickey et al. (1967) and has been widely cited. The explanatory variable is the age (in months) at which a child utters its first word, and the response variable is its Gesell adaptive score. These data (for 21 children) appear in Table 4, and they are plotted in Figure 5.

Mickey et al. (1967) decided that observation 19 is an outlier, by means of a sequential approach to detect outliers via stepwise regression. Andrews and Pregibon (1978), Draper and John (1981), and Paul (1983)

Table 4. First Word—Gesell Adaptive Score Data

Child (i)	Age in Months (x_i)	Gesell Score (y_i)
1	15	95
2	26	71
3	10	83
4	9	91
5	15	102
6	20	87
7	18	93
8	11	100
9	8	104
10	20	94
11	7	113
12	9	96
13	10	83
14	11	84
15	11	102
16	10	100
17	12	105
18	42	57
19	17	121
20	11	86
21	10	100

Source: Mickey et al. (1967).

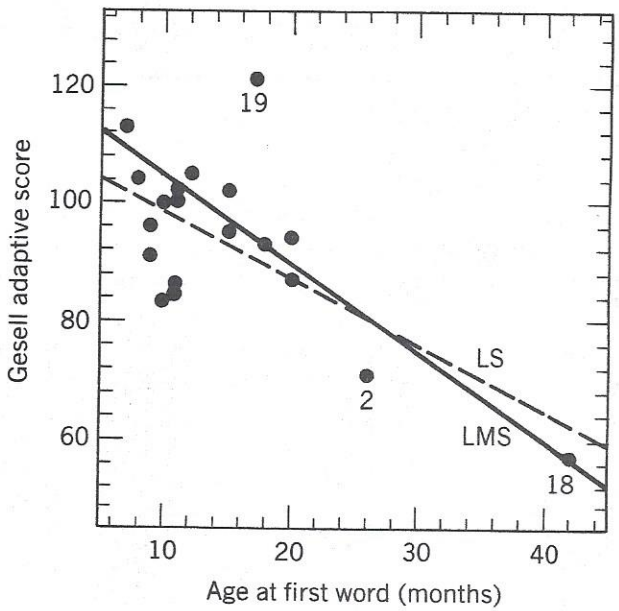


Figure 5. Scatterplot of Gesell adaptive score versus age at first word.

applied outlier diagnostics to this data set. (The use of such diagnostics will be discussed in Chapter 6.) The most important results of our own analysis are given below.

```
LEAST SQUARES REGRESSION
************************
```

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
AGE(MONTH)	-1.12699	.31017	-3.63343	.00177
CONSTANT	109.87380	5.06780	21.68077	.00000

```
SUM OF SQUARES            =      2308.58600

DEGREES OF FREEDOM        =        19

SCALE ESTIMATE            =        11.02291

VARIANCE - COVARIANCE MATRIX =

     .9621D-01
    -.1384D+01       .2568D+02

COEFFICIENT OF DETERMINATION (R SQUARED) =          .40997

THE F-VALUE =       13.202 (WITH   1 AND   19 DF)   P - VALUE =   .00177
```

OBSERVED GESELL SC.	ESTIMATED GESELL SC.	RESIDUAL	NO	RES/SC
95.00000	92.96901	2.03099	1	.18
71.00000	80.57213	-9.57213	2	-.87
83.00000	98.60395	-15.60395	3	-1.42
91.00000	99.73094	-8.73094	4	-.79
102.00000	92.96901	9.03099	5	.82
87.00000	87.33406	-.33406	6	-.03
93.00000	89.58804	3.41196	7	.31
100.00000	97.47696	2.52304	8	.23
104.00000	100.85790	3.14207	9	.29
94.00000	87.33406	6.66594	10	.60
113.00000	101.98490	11.01508	11	1.00
96.00000	99.73094	-3.73094	12	-.34
83.00000	98.60395	-15.60395	13	-1.42
84.00000	97.47696	-13.47696	14	-1.22
102.00000	97.47696	4.52304	15	.41
100.00000	98.60395	1.39605	16	.13
105.00000	96.34998	8.65002	17	.78
57.00000	62.54031	-5.54031	18	-.50
121.00000	90.71503	30.28497	19	2.75
86.00000	97.47696	-11.47696	20	-1.04
100.00000	98.60395	1.39605	21	.13

```
*******************************************************************************

LEAST MEDIAN OF SQUARES REGRESSION
**********************************
```

VARIABLE	COEFFICIENT
AGE(MONTH)	-1.50000
CONSTANT	119.75000

```
FINAL SCALE ESTIMATE          =          8.83928

COEFFICIENT OF DETERMINATION =           .44460
```

OBSERVED GESELL SC.	ESTIMATED GESELL SC.	RESIDUAL	NO	RES/SC
95.00000	97.25000	-2.25000	1	-.25
71.00000	80.75000	-9.75000	2	-1.10
83.00000	104.75000	-21.75000	3	-2.46
91.00000	106.25000	-15.25000	4	-1.73
102.00000	97.25000	4.75000	5	.54
87.00000	89.75000	-2.75000	6	-.31
93.00000	92.75000	.25000	7	.03
100.00000	103.25000	-3.25000	8	-.37
104.00000	107.75000	-3.75000	9	-.42
94.00000	89.75000	4.25000	10	.48
113.00000	109.25000	3.75000	11	.42
96.00000	106.25000	-10.25000	12	-1.16
83.00000	104.75000	-21.75000	13	-2.46
84.00000	103.25000	-19.25000	14	-2.18
102.00000	103.25000	-1.25000	15	-.14
100.00000	104.75000	-4.75000	16	-.54
105.00000	101.75000	3.25000	17	.37
57.00000	56.75000	.25000	18	.03
121.00000	94.25000	26.75000	19	3.03
86.00000	103.25000	-17.25000	20	-1.95
100.00000	104.75000	-4.75000	21	-.54

```
*******************************************************************************
```

```
REWEIGHTED LEAST SQUARES BASED ON THE LMS
*****************************************

  VARIABLE        COEFFICIENT     STAND. ERROR      T - VALUE      P - VALUE
 ------------------------------------------------------------------------
 AGE(MONTH)        -1.19331          .24348         -4.90100        .00012
   CONSTANT       109.30470         3.96997         27.53290        .00000

 WEIGHTED SUM OF SQUARES =        1340.02400
 DEGREES OF FREEDOM      =          18
 SCALE ESTIMATE          =           8.62820

 VARIANCE - COVARIANCE MATRIX =

      .5928D-01
     -.8448D+00        .1576D+02

 COEFFICIENT OF DETERMINATION (R SQUARED) =          .57163
 THE F-VALUE =        24.020 (WITH    1 AND    18 DF)    P - VALUE =   .00012
 THERE ARE     20 POINTS WITH NON-ZERO WEIGHT.
 AVERAGE WEIGHT          =            .95238

   OBSERVED        ESTIMATED
  GESELL SC.      GESELL SC.       RESIDUAL    NO   RES/SC  WEIGHT
    95.00000        91.40502         3.59498     1      .42    1.0
    71.00000        78.27860        -7.27860     2     -.84    1.0
    83.00000        97.37157       -14.37157     3    -1.67    1.0
    91.00000        98.56488        -7.56488     4     -.88    1.0
   102.00000        91.40502        10.59498     5     1.23    1.0
    87.00000        85.43846         1.56154     6      .18    1.0
    93.00000        87.82509         5.17491     7      .60    1.0
   100.00000        96.17826         3.82174     8      .44    1.0
   104.00000        99.75819         4.24181     9      .49    1.0
    94.00000        85.43846         8.56154    10      .99    1.0
   113.00000       100.95150        12.04849    11     1.40    1.0
    96.00000        98.56488        -2.56488    12     -.30    1.0
    83.00000        97.37157       -14.37157    13    -1.67    1.0
    84.00000        96.17826       -12.17826    14    -1.41    1.0
   102.00000        96.17826         5.82174    15      .67    1.0
   100.00000        97.37157         2.62843    16      .30    1.0
   105.00000        94.98495        10.01505    17     1.16    1.0
    57.00000        59.18563        -2.18563    18     -.25    1.0
   121.00000        89.01840        31.98160    19     3.71    1.0
    86.00000        96.17826       -10.17826    20    -1.18    1.0
   100.00000        97.37157         2.62843    21      .30    1.0
```

Because medium-sized output was requested, PROGRESS gives for each estimator a table with the observed response variable (y_i), its estimated value ($\hat{y}_i$), the residual (r_i), and the standardized residual (denoted by "RES/SC"). The standardized residuals for each regression are obtained by dividing the raw residuals by the scale estimate of the fit. A supplementary column with weights is added to the table for RLS regression. (These weights are determined from the LMS solution, as described in Section 3.) In this example, the case with index 19 received a zero weight. Indeed, this case has been identified as outlying because it has a large residual from the LMS fit. The equations of the LS and the RLS do not differ very much for this data set. The pair of points 2 and 18 has pulled the LS in the "good" direction. These points are good leverage points and possess small LMS residuals as well as small LS residuals. (The deletion of one or both of these points would have a considerable effect on the size of the confidence intervals.) One might even say that this data set is not a very good example of linear regression because deleting the leverage points (2 and 18) would not leave much of a linear

relationship between x and y. (We will return to this at the end of Section 6.)

Example 2: Number of Fires in 1976–1980

This data set (listed in Table 5) shows the trend from 1976 to 1980 of the number of reported claims of Belgian fire-insurance companies (from the annual report of the Belgian Association of Insurance Companies). It is included here to have an example with very few points.

When looking at the scatterplot in Figure 6, one notices a slight upward trend over the years. However, the number for 1976 is extraordinarily high. The reason lies in the fact that in that year the summer was extremely hot and dry (compared to Belgian standards), causing trees and

Table 5. Number of Fire Claims in Belgium from 1976 to 1980

Year (x_i)	Number of Fires (y_i)
76	16,694
77	12,271
78	12,904
79	14,036
80	13,874

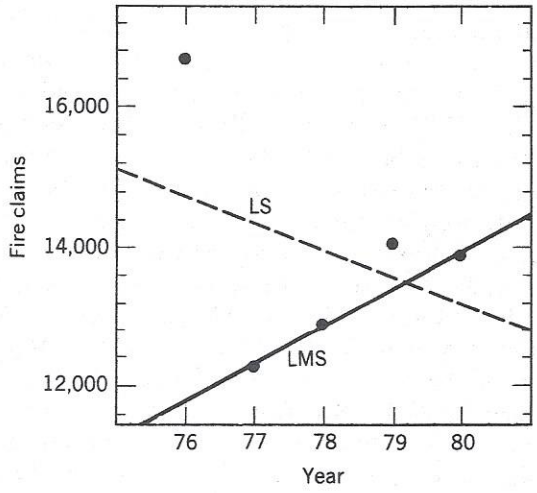


Figure 6. Number of fire claims in Belgium for the years 1976–1980.

bushes to catch fire spontaneously. It is striking that the LS $\hat{y} = -387.5x + 44180.8$ (dashed line) and the LMS $\hat{y} = 534.3x - 28823.3$ (solid line) are very different. The LS fits the data with a decreasing trend, whereas the LMS line increases. The outlier lies on the outside of the x-range and causes LS to grossly misbehave. It does not even possess the largest LS residual. This example shows again that examination of the LS residuals is not sufficient to identify the outlier(s).

Of course, in such a small data set one cannot draw any strong statistical conclusions, but it does show that one should think carefully about the data whenever LS and LMS yield substantially different results.

Example 3: Annual Rates of Growth of Prices in China

Table 6 contains the annual rates of growth of the average prices in the main cities of Free China from 1940 to 1948 (Simkin 1978). For instance, in 1940 prices went up 1.62% as compared to the previous year. In 1948 a huge jump occurred as a result of enormous government spending, the budget deficit, and the war, leading to what is called *hyperinflation*.

The LMS regression equation is given by

$$\hat{y} = 0.102x - 2.468\,,$$

whereas the LS estimate corresponds to

$$\hat{y} = 24.845x - 1049.468\,,$$

Table 6. Annual Rates of Growth of Average Prices in the Main Cities of Free China from 1940 to 1948

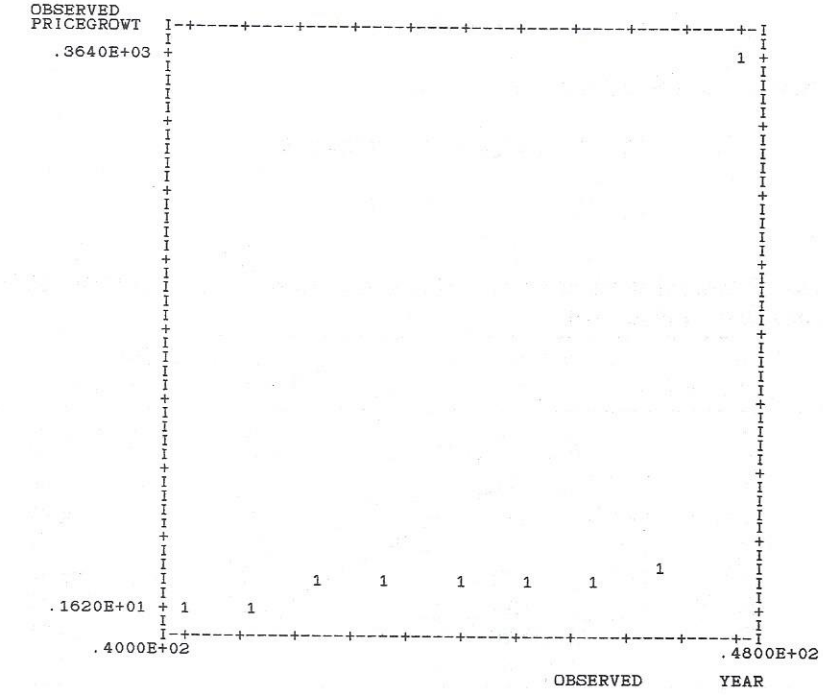
Year (x_i)	Growth of Prices (y_i)	Estimated Growth By LMS	Estimated Growth By LS
40	1.62	1.61	−55.67
41	1.63	1.71	−30.82
42	1.90	1.82	−5.98
43	2.64	1.92	18.87
44	2.05	2.02	43.71
45	2.13	2.12	68.56
46	1.94	2.22	93.40
47	15.50	2.33	118.25
48	364.00	2.43	143.09

Source: Simkin (1978).

which is totally different. To show which of these lines yields the better fit, Table 6 lists the estimated values by both methods. The LMS provides a fair approximation to the majority of the data, except of course for the last two years, where the observed y_i go astray. On the other hand, the LS fit is bad everywhere: The estimated $\hat{y}_i$ is even negative for the first three years, after which it becomes much too large, except for the 1948 value, which it cannot match either. Least squares smears out the effect (of nonlinearity of the original data) over the whole column, whereas LMS fits the majority of the data (where it is indeed linear) and allows the discrepancy to show up in those two years where actually something went wrong. Applying PROGRESS to this data set yields (among other things) the following output:

```
*******************************************
* ROBUST REGRESSION WITH A CONSTANT TERM. *
*******************************************

NUMBER OF CASES =      9
NUMBER OF COEFFICIENTS (INCLUDING CONSTANT TERM) =      2

THE EXTENSIVE SEARCH VERSION WILL BE USED.
DATA SET = ANNUAL RATES OF GROWTH OF PRICES IN CHINA
THERE ARE NO MISSING VALUES.
```

ANNUAL RATES OF GROWTH OF PRICES IN CHINA



MEDIANS =

YEAR	PRICEGROWT
44.0000	2.0500

DISPERSIONS =

YEAR	PRICEGROWT
2.9652	.6227

THE STANDARDIZED OBSERVATIONS ARE:

	YEAR	PRICEGROWT
1	-1.3490	-.6906
2	-1.0117	-.6745
3	-.6745	-.2409
4	-.3372	.9475
5	.0000	.0000
6	.3372	.1285
7	.6745	-.1767
8	1.0117	21.5998
9	1.3490	581.2665

PEARSON CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(PRICEGROWT IS THE RESPONSE VARIABLE)

YEAR	1.00	
PRICEGROWT	.57	1.00

SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
(PRICEGROWT IS THE RESPONSE VARIABLE)

YEAR	1.00	
PRICEGROWT	.85	1.00

****

LEAST SQUARES REGRESSION
****

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE	P - VALUE
YEAR	24.84500	13.67404	1.81695	.11207
CONSTANT	-1049.46800	602.69280	-1.74130	.12517

SUM OF SQUARES = 78531.34000

DEGREES OF FREEDOM = 7

SCALE ESTIMATE = 105.91870

VARIANCE - COVARIANCE MATRIX =

.1870D+03	
-.8227D+04	.3632D+06

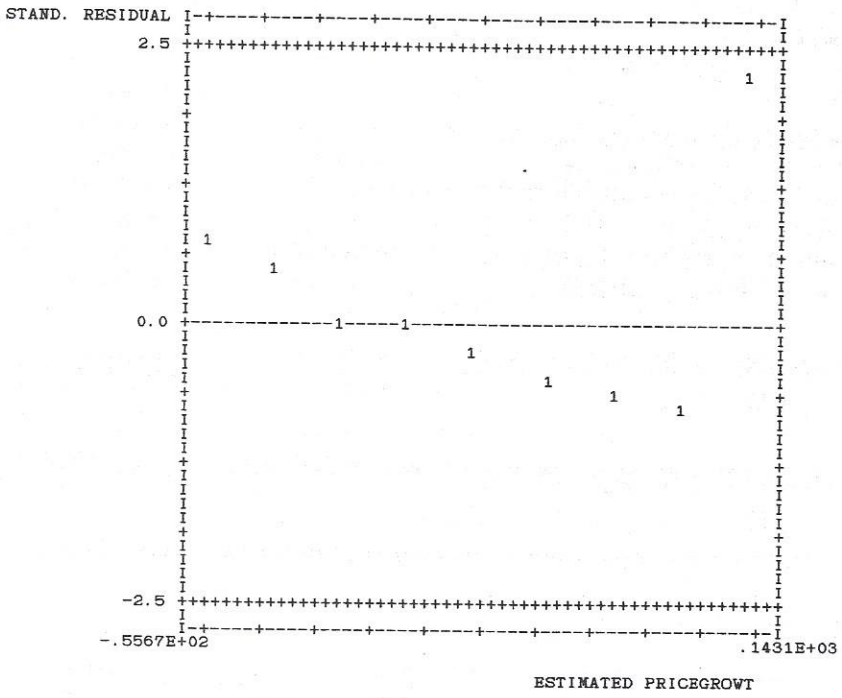
COEFFICIENT OF DETERMINATION (R SQUARED) = .32047

THE F-VALUE = 3.301 (WITH 1 AND 7 DF) P - VALUE = .11207

OBSERVED PRICEGROWT	ESTIMATED PRICEGROWT	RESIDUAL	NO	RES/SC
1.62000	-55.66766	57.28766	1	.54
1.63000	-30.82269	32.45269	2	.31
1.90000	-5.97766	7.87766	3	.07
2.64000	18.86731	-16.22731	4	-.15
2.05000	43.71228	-41.66228	5	-.39
2.13000	68.55737	-66.42738	6	-.63
1.94000	93.40234	-91.46234	7	-.86
15.50000	118.24730	-102.74730	8	-.97
364.00000	143.09230	220.90770	9	2.09

ANNUAL RATES OF GROWTH OF PRICES IN CHINA

--- L E A S T S Q U A R E S ---



****

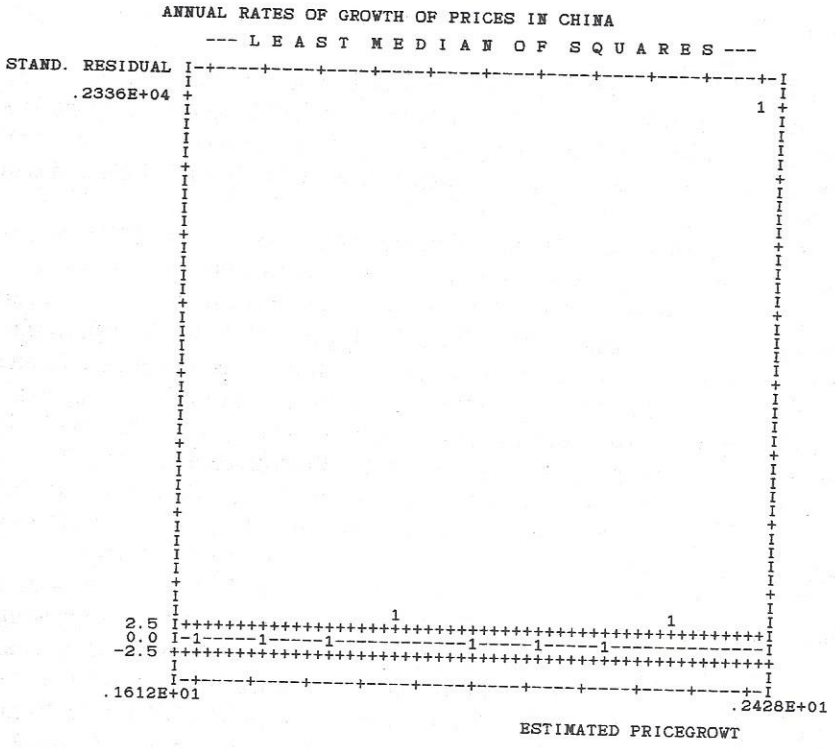
LEAST MEDIAN OF SQUARES REGRESSION
****

VARIABLE	COEFFICIENT
YEAR	.10200
CONSTANT	-2.46800

FINAL SCALE ESTIMATE = .15475

COEFFICIENT OF DETERMINATION = .93824

OBSERVED PRICEGROWT	ESTIMATED PRICEGROWT	RESIDUAL	NO	RES/SC
1.62000	1.61200	.00800	1	.05
1.63000	1.71400	-.08400	2	-.54
1.90000	1.81600	.08400	3	.54
2.64000	1.91800	.72200	4	4.67
2.05000	2.02000	.03000	5	.19
2.13000	2.12200	.00800	6	.05
1.94000	2.22400	-.28400	7	-1.84
15.50000	2.32600	13.17400	8	85.13
364.00000	2.42800	361.57200	9	2336.42



Note that observations 8 and 9 are outliers in the y-direction, which show up in the second column of the standardized data. Also, the Spearman rank correlation is much higher than the Pearson correlation because the outliers have disturbed the linear relation, whereas they obey the near-monotone relation between both variables.

For this example, we will discuss the residual plots (printed above) provided by PROGRESS. They display the standardized residual versus the estimated value of the response. An index plot is similar, except it contains the standardized residual plotted versus the index of the observation. (This makes it easier to find which observation corresponds to a given point in the plot. The index plot is also useful because the indices often reflect the order in which the observations were measured, so it may reveal sudden changes in the data recording process.) In both plots, a dashed line is drawn through zero, and a horizontal band on the interval $[-2.5, 2.5]$ is marked. These lines facilitate the interpretation of the

results. When the observed value y_i equals the estimated value $\hat{y}_i$, then the resulting residual becomes zero. Points in the neighborhood of this zero line are best fitted by the model. If the residuals are normally distributed, then one can expect that roughly 98% of the standardized residuals will lie in the interval $[-2.5, 2.5]$. Thus, observations for which the standardized residuals are situated far from the horizontal band can be identified as outlying.

The first residual plot in the above output shows how the LS fit masks the bad point. The LS has been pulled away by this outlier, and its scale estimate has exploded. The LS residual associated with the outlier even lies within the horizontal band. Because of this effect, the interpretation of a residual plot corresponding to the LS estimator is dangerous. In the residual plot of the LMS, the outlier is very far away from the band. Residual plots corresponding to robust estimators are even more useful in problems with several variables, as will be illustrated in Chapter 3.

These graphical tools are very convenient for spotting the outlying observations. The LS result can be trusted only if the residual plots of both the robust and nonrobust regression methods agree closely.

Besides the identification of outliers, the residual plot provides a diagnostic tool for assessing the adequacy of the fit and for suggesting transformations. An ideal pattern of the residual plot, which indicates an adequate model and well-behaved data, is a horizontal cloud of points with constant vertical scatter. Anomalies in the pattern of residuals can lead to several courses of action. For instance, the plot may indicate that the variance of the residuals increases or decreases with increasing estimated y or with another variable. (See the education expenditure data in Section 3 of Chapter 3 for an illustration.) This is called *heteroscedasticity*, in contrast to the classical model where the errors have a constant variance, in which case we speak of *homoscedasticity*. This problem may be approached by applying a suitable transformation on either an explanatory or the response variable. If this heteroscedasticity appears in an index plot, then one should turn back to the origin of the data in order to look for the cause of the phenomenon. For example, it could be that a time-related variable has to be included in the model. Also, other model failures may be visible in residual plots. A pattern resembling a horseshoe may be caused by nonlinearity. (See the cloud point data in Section 3 of Chapter 3 for an example.) In such a situation, a transformation on an explanatory or on the response variable, or an additional squared term or a cross-product term in the model, or the addition of an extra explanatory variable may be required. (In many applications of regression, there is a substantial amount of prior knowledge that can be useful in choosing between these possibilities.)

Example 4: Brain and Weight Data

Table 7 presents the brain weight (in grams) and the body weight (in kilograms) of 28 animals. (This sample was taken from larger data sets in Weisberg 1980 and Jerison 1973.) It is to be investigated whether a larger brain is required to govern a heavier body.

A clear picture of the relationship between the logarithms (to the base 10) of these measurements is shown in Figure 7. This logarithmic

Table 7. Body and Brain Weight for 28 Animals

Index (i)	Species	Body Weight[a] (x_i)	Brain Weight[b] (y_i)
1	Mountain beaver	1.350	8.100
2	Cow	465.000	423.000
3	Gray wolf	36.330	119.500
4	Goat	27.660	115.000
5	Guinea pig	1.040	5.500
6	Diplodocus	11700.000	50.000
7	Asian elephant	2547.000	4603.000
8	Donkey	187.100	419.000
9	Horse	521.000	655.000
10	Potar monkey	10.000	115.000
11	Cat	3.300	25.600
12	Giraffe	529.000	680.000
13	Gorilla	207.000	406.000
14	Human	62.000	1320.000
15	African elephant	6654.000	5712.000
16	Triceratops	9400.000	70.000
17	Rhesus monkey	6.800	179.000
18	Kangaroo	35.000	56.000
19	Hamster	0.120	1.000
20	Mouse	0.023	0.400
21	Rabbit	2.500	12.100
22	Sheep	55.500	175.000
23	Jaguar	100.000	157.000
24	Chimpanzee	52.160	440.000
25	Brachiosaurus	87000.000	154.500
26	Rat	0.280	1.900
27	Mole	0.122	3.000
28	Pig	192.000	180.000

[a]In kilograms.
[b]In grams.

Source: Weisberg (1980) and Jerison (1973).

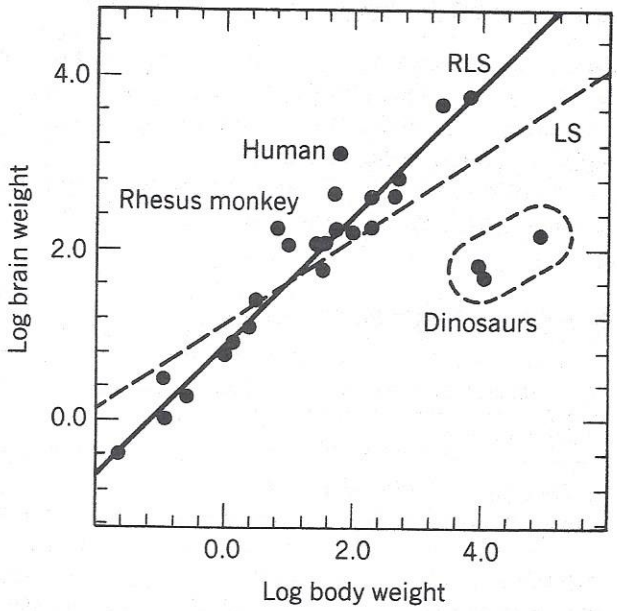


Figure 7. Logarithmic brain weight versus logarithmic body weight for 28 animals with LS (dashed line) and RLS fit (solid line).

transformation was necessary because plotting the original measurements would fail to represent either the smaller or the larger measurements. Indeed, both original variables range over several orders of magnitude. A linear fit to this transformed data would be equivalent to a relationship of the form

$$\hat{y} = \hat{\theta}_2' x^{\hat{\theta}_1}$$

between brain weight (y) and body weight (x). Looking at Figure 7, it seems that this transformation makes things more linear. Another important advantage of the log scale is that the heteroscedasticity disappears.

The LS fit is given by

$$\log \hat{y} = 0.49601 \log x + 1.10957$$

(dashed line in Figure 7). The standard error associated with the slope equals 0.0782, and that of the intercept term is 0.1794. In Section 3, we explained how to construct a confidence interval for the unknown regression parameters. For the present example, $n = 28$ and $p = 2$, so one has to

use the 97.5% quantile of the t-distribution with 26 degrees of freedom, which equals 2.0555. Using the LS results, a 95% confidence interval for the slope is given by [0.3353; 0.6567]. The RLS yields the solid line in Figure 7, which is a fit with a steeper slope:

$$\log \hat{y} = 0.75092 \log x + 0.86914 .$$

The slope estimated by the RLS technique even falls outside the 95% confidence interval associated with the LS fit! The standard error of the regression coefficients in RLS is reduced remarkably as compared with

Table 8. Standardized LS and RLS Residuals for the Brain and Body Weight Data

Index	Species	Standardized LS Residuals	Standardized RLS Residuals	w_i
1	Mountain beaver	−0.40	−0.27	1
2	Cow	0.29	−1.13	1
3	Gray wolf	0.29	0.17	1
4	Goat	0.36	0.50	1
5	Guinea pig	−0.57	−0.65	1
6	Diplodocus	−2.15	−10.19	0
7	Asian elephant	1.30	1.08	1
8	Donkey	0.58	0.21	1
9	Horse	0.54	−0.43	1
10	Potar monkey	0.68	2.02	1
11	Cat	0.06	0.68	1
12	Giraffe	0.56	−0.37	1
13	Gorilla	0.53	0.00	1
14	Human	1.69	4.15	0
15	African elephant	1.13	0.08	1
16	Triceratops	−1.86	−9.20	0
17	Rhesus monkey	1.10	3.47	0
18	Kangaroo	−0.19	−1.29	1
19	Hamster	−0.98	−0.81	1
20	Mouse	−1.04	−0.17	1
21	Rabbit	−0.34	−0.39	1
22	Sheep	0.40	0.29	1
23	Jaguar	0.14	−0.80	1
24	Chimpanzee	1.02	2.22	1
25	Brachiosaurus	−2.06	−10.94	0
26	Rat	−0.84	−0.80	1
27	Mole	−0.27	1.35	1
28	Pig	0.02	−1.50	1

the LS, namely 0.0318 for the slope and 0.0618 for the constant term. A 95% confidence interval for the unknown slope is now given by [0.6848; 0.8171], which is narrower than the interval coming from LS. The t-values associated with the RLS regression coefficients are very large, which implies that the slope and intercept are significantly different from zero. Moreover, the determination coefficient R^2, which is a summary measure for overall goodness of fit, increases from 0.608 for LS to 0.964 for RLS. This example shows that not only the LS regression coefficients, but also the whole LS inference, may become doubtful in the presence of outliers.

Table 8 lists the standardized LS and RLS residuals and the w_i determined on the basis of the LMS. From the RLS, it is easy to detect unusual observations and to give them special consideration. Indeed, looking at the five cases with zero w_i, one can easily understand why they have to be considered as outlying. The most severe (and highly negative) RLS residuals are those of cases 6, 16, and 25, which are responsible for the low slope of the LS fit. These are three dinosaurs, each of which possessed a small brain as compared with a heavy body. In this respect they contrast with the mammals which make up the rest of the data set. The LMS regression also produced a zero weight for cases 14 and 17, namely the human and the rhesus monkey. For them, the actual brain weight is higher than that predicted by the linear model. Unlike the dinosaurs, their residuals are therefore positive. Concluding, one could say that dinosaurs, humans, and rhesus monkeys do not obey the same trend as the one followed by the majority of the data.

5. AN ILLUSTRATION OF THE EXACT FIT PROPERTY

The phrase "exact fit" stands for situations where a large percentage of the observations fits some linear equation exactly. For example, in simple regression this happens when the majority of the data lie exactly on a straight line. In such a case a robust regression method should recover that line. At an Oberwolfach Meeting, Donoho (1984) called this the *exact fit property*. For instance, the repeated median satisfies this property (Siegel 1982), as well as the LMS (Rousseeuw 1984). When at least $n - [n/2] + 1$ of the observations lie on the same line, then the equation of this line will be the LMS solution. More details on the exact fit property and its relation to the breakdown point can be found in Section 4 of Chapter 3.

The data in Table 9 come from Emerson and Hoaglin (1983, p. 139). They were devised by A. Siegel as a counterexample for the resistant line estimator (which will be briefly discussed in Section 7). Looking at the

Table 9. Siegel's Data Set

i	x_i	y_i
1	−4	0
2	−3	0
3	−2	0
4	−1	0
5	0	0
6	1	0
7	2	−5
8	3	5
9	12	1

Source: Emerson and Hoaglin (1983).

scatterplot of these data (Figure 8), Emerson and Hoaglin suggest that a line with slope 0 would be a reasonable summary. Indeed, six out of the nine points actually lie *on* the line with zero slope and zero intercept. By running PROGRESS we see that least median of squares yields this line exactly, unlike least squares.

Exact fit situations also occur in real data. One example has to do with the use of an electron microscope in crystallography. Several discrete variables (such as the number of edges and certain symmetry properties) are observed for a large number of "cells" (consisting of molecules) in a

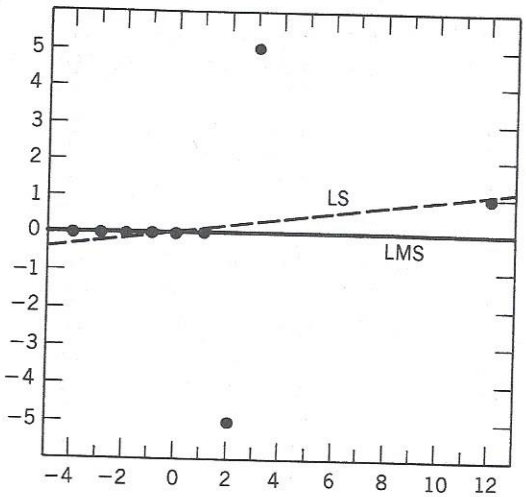


Figure 8. Example of exact fit: Scatterplot of Siegel's data set with LS (dashed line) and LMS fit (solid line).

regular lattice. In practice, most of these cells are good (and hence fit exactly), whereas a fraction are damaged by the radiation of the microscope itself.

6. SIMPLE REGRESSION THROUGH THE ORIGIN

When it is known in advance that the intercept term is zero, then one has to impose this on the model. This leads to equation (1.2) in Section 1 of this chapter.

Afifi and Azen (1979, p. 125) give an example (copied in Table 10) where such a model is suitable. The data concern the calibration of an instrument that measures lactic acid concentration in blood. One compares the true concentration x_i with the measured value y_i. The explanatory variable in this data set is designed. This means that its values are fixed in advance. Consequently, such a data set does not contain leverage points. The scatterplot in Figure 9 displays a roughly linear relationship

Table 10. Data on the Calibration of an Instrument that Measures Lactic Acid Concentration in Blood

Index (i)	True Concentration (x_i)	Instrument (y_i)
1	1.0	1.1
2	1.0	0.7
3	1.0	1.8
4	1.0	0.4
5	3.0	3.0
6	3.0	1.4
7	3.0	4.9
8	3.0	4.4
9	3.0	4.5
10	5.0	7.3
11	5.0	8.2
12	5.0	6.2
13	10.0	12.0
14	10.0	13.1
15	10.0	12.6
16	10.0	13.2
17	15.0	18.7
18	15.0	19.7
19	15.0	17.4
20	15.0	17.1

Source: Afifi and Azen (1979).

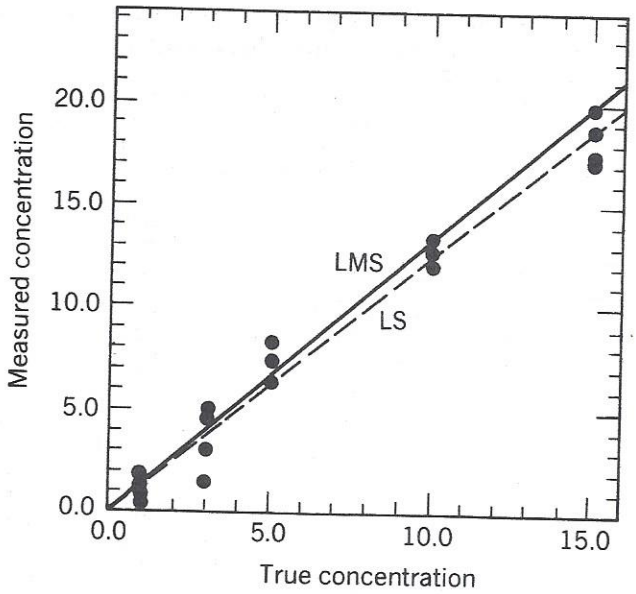


Figure 9. Scatterplot of the data in Table 10. A model without intercept term is used. The LS estimator corresponds to the dashed line and the LMS estimator corresponds to the solid line.

between both variables. Indeed, for this example the LS (dashed line) and the LMS (solid line) almost coincide. However, the literature also contains calibration data with outliers. Massart et al. (1986) apply the LMS to detect outliers and model errors in some real-data examples from analytical chemistry.

Let us now look at data used in Hampel et al. (1986, Chapter 6), which are reproduced here in Table 11. They are measurements of water flow at two different points (Libby, Montana and Newgate, British Columbia) on the Kootenay river in January, for the years 1931–1943. The original data came from Ezekiel and Fox (1959, pp. 57–58), and Hampel et al. changed the Newgate measurement for the year 1934 to 15.7 for illustrative purposes (thereby converting a "good" leverage point into a "bad" one).

For each variable j, PROGRESS will now compute another type of dispersion,

$$s_j = 1.4826 \operatorname*{med}_i |x_{ij}| , \tag{6.1}$$

because it is logical to consider the deviations from zero (and not from some average or median value). For the same reason, the data are

Table 11. Water Flow Measurements in Libby and Newgate on the Kootenay River in January for the Years 1931–1943

Year	Libby (x_i)	Newgate (y_i)
31	27.1	19.7
32	20.9	18.0
33	33.4	26.1
34	77.6	15.7[a]
35	37.0	26.1
36	21.6	19.9
37	17.6	15.7
38	35.1	27.6
39	32.6	24.9
40	26.0	23.4
41	27.6	23.1
42	38.7	31.3
43	27.8	23.8

[a]The original value was 44.9.

Source: Ezekiel and Fox (1959).

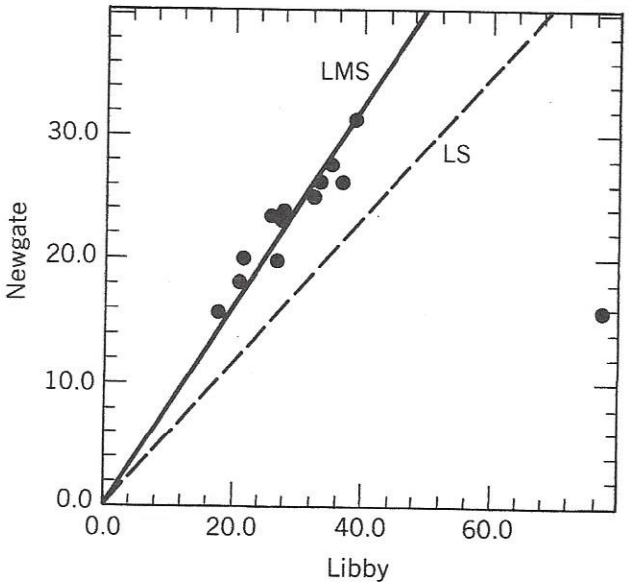


Figure 10. Scatterplot of the water flow in Libby and Newgate, with LS (dashed line) and LMS fit (solid line).

standardized by replacing each x_{ij} by

$$\frac{x_{ij}}{s_j}, \tag{6.2}$$

which is printed when large output was requested. For the Kootenay data, this yields the following output:

```
DISPERSION OF ABSOLUTE VALUES=

       41.2163       34.6928

THE STANDARDIZED OBSERVATIONS ARE:

        LIBBY        NEWGATE
 1      0.6575       0.5678
 2      0.5071       0.5188
 3      0.8104       0.7523
 4      1.8828       0.4525
 5      0.8977       0.7523
 6      0.5241       0.5736
 7      0.4270       0.4525
 8      0.8516       0.7956
 9      0.7909       0.7177
10      0.6308       0.6745
11      0.6696       0.6658
12      0.9389       0.9022
13      0.6745       0.6860
```

This shows us that case 4 (the year 1934) is a leverage point.

Estimating the unknown slope with the LS technique and the LMS technique gives rise to the dashed and the solid lines in Figure 10.

The LS line ($\hat{y} = 0.5816x$) is attracted by the leverage point associated with the year 1934. On the other hand, this contaminated case produces a large standardized residual with respect to the LMS fit given by $\hat{y} = 0.8088x$. The coefficient of determination corresponding to this fit is rather large too, namely 0.997, whereas it was only 0.798 for the LS fit. From the scatterplot in Figure 10, it appears that the LMS fits the good points very closely, whereas the LS fit fails to give a good prediction for the response variable because its slope is too small.

*7. OTHER ROBUST TECHNIQUES FOR SIMPLE REGRESSION

During the past 50 years, many techniques have been proposed for estimating slope and intercept in a simple regression model. Most of them do not attain a breakdown point of 30%, as will be shown by means of the breakdown plot discussed below.

Emerson and Hoaglin (1983) give a historical survey that contains

some explanation about the techniques of Wald (1940), Nair and Shrivastava (1942), Bartlett (1949), and Brown and Mood (1951). These regression methods are all based on the idea of splitting up the data set and then defining a summary value in each group. They testify to the concern about the dramatic lack of robustness of least squares regression, but can only be applied to simple regression.

The resistant line method of Tukey (1970/1971) is a variant of some of these first approaches to robust estimation. This "pencil-and-paper" technique for fitting a line to bivariate data starts from a partition of the data set into three parts, as nearly equal in size as possible. This allocation is performed according to the smallest, intermediate, and largest x-values. Tied x-values are assigned to the same group. Then, the resistant slope is determined such that the median of the residuals in the two outermost groups (L stands for left, and R stands for right) are equal, that is,

$$\underset{i\in L}{\text{med}}\,(y_i - \hat{\theta}_1 x_i) = \underset{i\in R}{\text{med}}\,(y_i - \hat{\theta}_1 x_i)\,, \tag{7.1}$$

and the intercept is chosen to make the median residuals of both groups zero. The worst-case bound on the available protection against outliers for this technique is 1/6 because one uses the median, which has a breakdown point of 1/2, in both groups. Velleman and Hoaglin (1981) provide an algorithm and a portable program for finding the resistant line. Some theoretical and Monte Carlo results on this estimator are provided by Johnstone and Velleman (1985b). The Brown and Mood (1951) technique is defined in a similar way, but with two groups instead of three. As a consequence, the breakdown point for this technique increases to 1/4.

Andrews (1974) developed another median-based method for obtaining a straight-line fit. He starts by ordering the x-values from smallest to largest. Then he eliminates a certain number of the smallest and of the largest x-values, and also a certain number of the values in the neighborhood of the median x-value. Of the two remaining subsets of x-values, he calculates the medians ($\text{med}\,x_1$ and $\text{med}\,x_2$). For the corresponding y-values, the medians are calculated too ($\text{med}\,y_1$ and $\text{med}\,y_2$). The slope of the fit is then defined as

$$\hat{\theta}_1 = \frac{\text{med}\,y_2 - \text{med}\,y_1}{\text{med}\,x_2 - \text{med}\,x_1}$$

The breakdown point of this procedure is at most 25% because half of the data in either subset can determine the fitted line. Andrews generalized

this technique to multiple regression by applying a so-called "sweep" operator. This means that at each iteration the dependence of one variable on another is removed by adjusting a variable by a multiple (determined in an earlier stage) of another. The same idea has been used to generalize the resistant line (Johnstone and Velleman 1985b, Emerson and Hoaglin 1985).

Another group of simple regression estimators uses the pairwise slopes as building stones in their definition, without splitting up the data set. Theil (1950) proposed as an estimator of $\hat{\theta}_1$ the median of all C_n^2 slopes, namely

$$\hat{\theta}_1 = \underset{1 \le i < j \le n}{\text{med}} \frac{y_j - y_i}{x_j - x_i}, \tag{7.2}$$

which possesses a high asymptotic efficiency. Sen (1968) extended this estimator to handle ties among the x_i. It turns out that the breakdown point of these techniques is about 29.3%. This value can be explained by the following reasoning: Since the proportion of "good" slopes has to be at least 1/2, one needs that $(1-\varepsilon)^2$ be at least 1/2, where ε is the fraction of outliers in the data. From this it follows that

$$\varepsilon \le 1 - (1/2)^{1/2} \approx 0.293 . \tag{7.3}$$

More recently, Siegel (1982) increased the breakdown point to 50% by means of a crucial improvement, leading to the repeated median estimator. In this method, one has to compute a two-stage median of the pairwise slopes instead of a single median such as in (7.2). The slope and intercept are then defined as

$$\begin{aligned} \hat{\theta}_1 &= \underset{i}{\text{med}} \underset{j \ne i}{\text{med}} \frac{y_j - y_i}{x_j - x_i} \\ \hat{\theta}_2 &= \underset{i}{\text{med}} \left(y_i - \hat{\theta}_1 x_i \right) . \end{aligned} \tag{7.4}$$

This list of simple regression techniques is not complete. Several other methods will be dealt with in Section 6 of Chapter 3.

In order to compare the fit obtained by different regression methods in the presence of contamination, we considered an artificial example. Thirty "good" observations were generated according to the linear relation

$$y_i = 1.0 x_i + 2.0 + e_i , \tag{7.5}$$

where x_i is uniformly distributed on [1, 4], and e_i is normally distributed with mean zero and standard deviation 0.2. Then a cluster of 20 "bad" observations was added, possessing a spherical bivariate normal distribution with mean (7, 2) and standard deviation 0.5. This yielded 40% of contamination in the pooled sample, which is very high. Actually, this amount was chosen to demonstrate what happens if one goes above the breakdown point of most estimators for simple regression. Let us now see which estimator succeeds best in describing the pattern of the majority of the data. The classical least squares method yields $\hat{\theta}_1 = -0.47$ and $\hat{\theta}_2 = 5.62$: It clearly fails because it tries to suit both the good and the bad points. Making use of the ROBETH library (Marazzi 1980), three robust estimators were applied: Huber's M-estimator [Chapter 1, equation (2.10)] with $\psi(x) = \min(1.5, \max(-1.5, x))$, Mallows' generalized M-estimator [Chapter 1, equation (2.12)] with Hampel weights, and Schweppe's generalized M-estimator [Chapter 1, equation (2.13)] with Hampel–Krasker weights (both Mallows' and Schweppe's estimators use the same Huber function). All three methods, however, gave results virtually indistinguishable from the LS solution: the four lines almost coincide in Figure 11.

The repeated median estimator (7.4) yields $\hat{\theta}_1 = 0.30$ and $\hat{\theta}_2 = 3.11$. If the cluster of "bad" points is moved further down, the repeated median line follows it a little more and then stops. Therefore, this method does not break down, although in this particular example it does not yield a

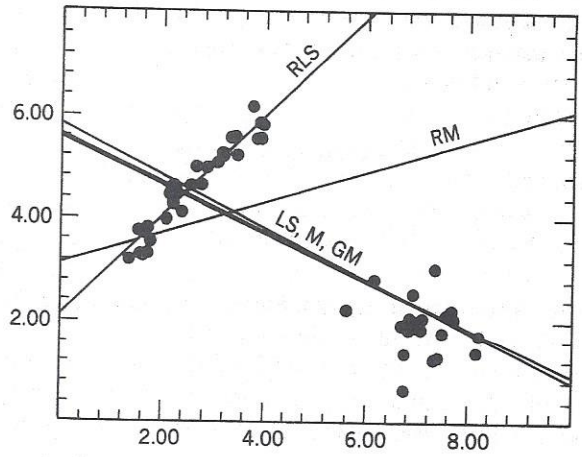


Figure 11. Regression lines for the simulated data using six methods (RLS, reweighted least squares based on LMS; LS, least squares; M, Huber's M-estimator; GM, Mallows' and Schweppe's generalized M-estimators; RM, repeated median). *Source:* Rousseeuw (1984).

very good fit. On the other hand, the LMS-based RLS yields $\hat{\theta}_1 = 0.97$ and $\hat{\theta}_2 = 2.09$, which comes quite close to the original values of θ_1 and θ_2. When the cluster of bad points is moved further away, this solution does not change any more. Moreover, the RLS method does not break down even when only 26 "good" points and 24 outliers are used.

The breakdown properties of these estimators were investigated more extensively in a larger experiment. To begin with, we generated 100 "good" observations according to the linear relation (7.5). To these data, we applied the same fitting techniques as in the above example (see Figure 11). Because the data were well behaved, all estimators yielded values of $\hat{\theta}_1$ and $\hat{\theta}_2$ which were very close to the original θ_1 and θ_2. Then we started to contaminate the data. At each step we deleted one "good" point and replaced it by a "bad" point generated according to a bivariate normal distribution with mean $(7, 2)$ and standard deviation 0.5. We repeated this until only 50 "good" points remained. The LS was immediately affected by these leverage points, so the estimated slope $\hat{\theta}_1$ became negative, moving away from the ideal value $\theta_1 = 1.0$. In Figure 12, the value of $\hat{\theta}_1$ is drawn as a function of the percentage of outliers. We call this a *breakdown plot.*

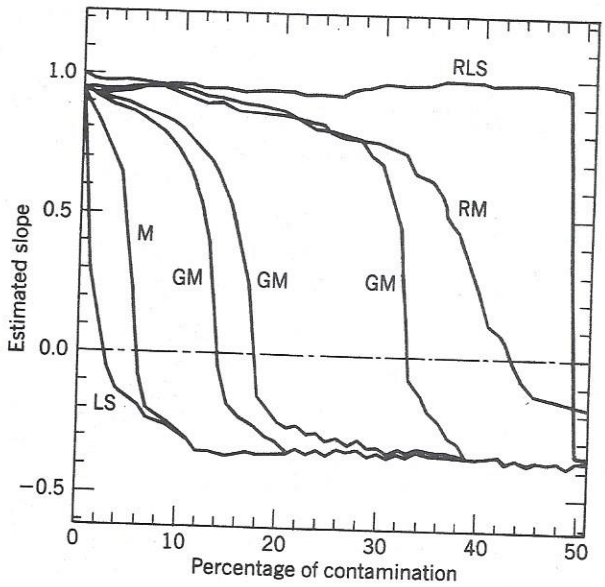


Figure 12. Breakdown plot, showing the estimated slope as a function of the percentage of contamination. The estimators are those of Figure 11, applied to a similar data configuration. *Source:* Rousseeuw et al. (1984b).

We see that LS breaks down first and is then followed by the Huber-type M-estimator and the GM-estimators. (The best of those appears to be the Mallows estimator with Hampel weights, which could tolerate slightly over 30% of outliers in this experiment. Note, however, that GM-estimators will break down much earlier in higher dimensions.) The repeated median goes down gradually and becomes negative at about 40%, whereas the RLS holds on until the very end before breaking down.

REMARK. At this point we would like to say something about the interpretation of breakdown. By definition, breakdown means that the bias $\|T(Z') - T(Z)\|$ becomes unbounded. For smaller fractions of contamination the bias remains bounded, but this does not yet imply that it is small. Although the breakdown point of LMS regression is always 50%, we have found that the effect of outliers may differ according to the quality of the "good" data. For instance, compare the data configurations in Figure 13a and b. In both situations there is a fraction $1 - \varepsilon$ of original data and a percentage ε of outliers. Because $\varepsilon < 50\%$, moving the outliers around does not make the LMS break down in either case. Nevertheless, its maximal bias becomes much larger—though not infinite!—in Figure 13b, where the majority of the data is "ball-shaped" with R^2 close to 0, than in Figure 13a, where the good data already have a strong linear structure with $R^2 \approx 1$. It seems to us that this behavior lies in the nature

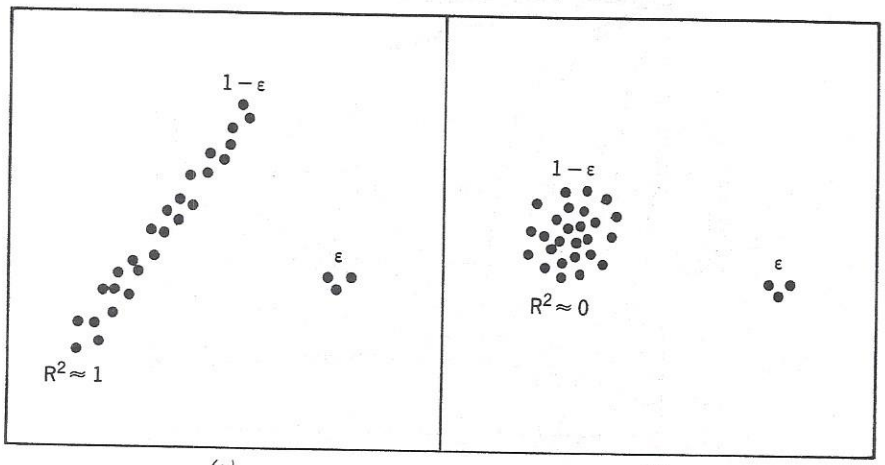


Figure 13. Sketch of two data configurations: (a) the good data possess a strong linear structure so the LMS will have only a small bias, and (b) the majority of the points have no linear structure so the effect of outliers will become much larger.

of things. For instance, the data in Figures 2, 3, 4, 6, 7, 10, and 11 are like Figure 13*a*, whereas the Mickey data (Figure 5) are not much better than Figure 13*b*.

EXERCISES AND PROBLEMS

Sections 1–4

1. Which is more robust, the Pearson product-moment correlation coefficient or the Spearman rank correlation coefficient? How does this reflect in the correlation coefficients (between extraction and titration) in the PROGRESS output reproduced at the end of Section 2?
2. In the framework of LS regression with intercept, explain why the (one-sided) p-value of the coefficient of determination equals the (one-sided) p-value of the F-statistic. If there is only one explanatory variable apart from the intercept, show that this probability is also equal to the (two-sided) p-value of the t-statistic of the slope and to the (two-sided) p-value of the Pearson correlation coefficient. What happens if we switch to RLS?
3. Making use of the definition of the LMS, can you explain heuristically why it is required that the number of cases be more than twice the number of regression coefficients (i.e., $n > 2p$)? Apart from this constraint, can you also argue why small n might lead to problems, based on the probability of linear clusters occurring merely by virtue of random fluctuations?
4. Reanalyze the Pilot-Plant data (Table 1) with an outlier in the y-direction. For instance, assume that the y-value of the 19th observation was recorded as 840 instead of 84. Apply PROGRESS to these contaminated data, and compare the effect of the outlier on LS, LMS, and RLS.
5. Find (or construct) an example where the least squares R^2 is larger with the outlier included than without it.
6. Apply PROGRESS to the monthly payments data of exercise 8 of Chapter 1. Give a detailed discussion of the results, including the standardized observations, the correlation coefficients, the p-values, the coefficients of determination, and the standardized residuals. Compare the LMS and the RLS with the eye fit. How many outliers are identified?
7. Let us look at an example on Belgian employers' liability insurance. Table 12 lists the number of reported accidents of people going to or

Table 12. Number of Accidents in 1975–1981

Year (x_i)	Number of Accidents (y_i)
75	18,031
76	18,273
77	16,660
78	15,688
79	23,805
80	15,241
81	13,295

Source: The 1981 Annual Report of the Belgian Association of Insurance Companies.

coming from work, from 1975 to 1981. Only accidents leading to at least 1 day of absence were counted. When looking at these data, we notice a downward trend over the years. However, the number for 1979 is very high. This is because during the first few months of 1979 it was extremely cold, with snow and ice on most days (one has to go back many decades to find records of a similar winter). Therefore, the number of fractures, sprains, and dislocations in January and February was overwhelming. Discuss the PROGRESS output for these data. Is 1979 an outlier in x or in y? (Look at the scatterplot.) Can the outlier be identified by means of the standardized observations? Explain the difference between the Pearson correlation and the Spearman correlation. How many outliers can be identified on the basis of the standardized residuals of the LS solution? And with the robust regressions? What is the effect of the outlier on the p-value of the LS slope?

8. Table 13 lists the total 1981 premium income of pension funds of Dutch firms, for 18 professional branches. In the next column the respective premium reserves are given. The highest amounts correspond to P.G.G.M., the pension fund for medical care workers. Its premium income is three times larger than the second largest branch (the building industry). Draw a plot of reserves versus income (e.g., by means of PROGRESS). Compare the simple regression of reserves as a function of income using LS with the LMS regression line. Is P.G.G.M. a good or a bad leverage point for this model? Does it have the largest LS residual? How does this compare with the RLS residuals, and what happens to the p-values of the coefficients? The residual plot indicates that a linear model is not very suitable here. Is it possible to transform either variable to make the relation more linear?

Table 13. Pension Funds for 18 Professional Branches

Index	Premium Income[a]	Premium Reserves[a]
1	10.4	272.2
2	15.6	212.9
3	16.2	120.7
4	17.9	163.6
5	37.8	226.1
6	46.9	622.9
7	52.4	1353.2
8	52.9	363.6
9	71.0	951.7
10	73.9	307.2
11	76.3	588.4
12	77.0	952.5
13	131.2	1157.3
14	151.7	2105.6
15	206.1	3581.4
16	314.7	3404.7
17	470.8	4095.3
18	1406.3	6802.7

[a]In millions of guilders.

Source: de Wit (1982, p. 52).

Sections 5 and 6

9. Run the exact fit example of Section 5 again as a simple regression through the origin.

10. Look for an example where a line through the origin is more appropriate (from subject-matter knowledge) than with intercept.

11. Repeat exercise 8 (on the pension funds of 18 professions) by means of simple regression without intercept. This makes sense because the RLS intercept was not significantly different from zero. Draw the corresponding lines through the origin. Which transformation do you propose to linearize the data?

Section 7

12. Apply the resistant line (7.1), Theil's estimator (7.2), and the repeated median (7.4) to the monthly payments data (Table 1 of Chapter 1).

13. In the framework of simple regression, Hampel (1975, p. 379) considered the line through $(\text{med}_j\, x_j, \text{med}_j\, y_j)$ which has equal num-

bers of positive residuals on both sides of $\text{med}_j\, x_j$. (He then dismissed this estimator by stating that it may lead to a lousy fit.) Simon (1986) proposed the "median star" line, which also goes through ($\text{med}_j\, x_j$, $\text{med}_j\, y_j$) and has the slope

$$\hat{\theta}_1 = \underset{i}{\text{med}} \left(\frac{y_i - \underset{j}{\text{med}}\, y_j}{x_i - \underset{j}{\text{med}}\, x_j} \right).$$

Show that this is the same estimator. What is its breakdown point? Write a small program and apply it to the situations of Figures 11 and 12. Repeat this for another version of the median star in which the intercept is chosen so that $\text{med}_i\, (r_i) = 0$.

14. (Research problem). Is it possible to construct a standardized version of the breakdown plot, which does not depend on random number generation?