doi.org/10.26398/IJAS.0029-005

A REACTION TO A CHALLENGING EXAMPLE IN MULTIPLE REGRESSION ANALYSIS

Ettore Marubini, Annalisa Orenti[1]

Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Epidemiology and Biometry G. A. Maccacaro, University of Milan, Milan, Italy

***Abstract**. In a very stimulating paper, Preece gives an artificial dataset useful to illustrate the hazard of multiple regression and challenges the reader to spot the simple inbuilt features of these data. The present note aims at finding how Preece generated the whole set of data. First of all OLS regression model is fitted to the data; after checking for model assumptions some doubts arise on the validity of OLS regression; thus robust regression estimators are considered as a proper alternative. The latter give discordant coefficient estimates, but after a deep analysis, they agree in highlighting the presence of two subsets within the dataset: 9 cases being generated by one model, and the remaining 8 cases being generated by a second model. This particular pattern of the data is recognized by the mixture model as well.*

***Keywords**: Linear regression analysis, Regression illustrative example, Hazards in regression analysis, Robust regression, Mixture models.*

1. INTRODUCTION

In a very stimulating paper Preece (1986) discusses the features of illustrative examples used in statistical papers and textbooks and presents critical considerations on their appropriateness. With regard to regression analysis, the Author criticizes two examples (one of which is the well known "stackloss dataset" (Brownlee, 1960, p. 491)), because they are lacking the information crucial to the conduct of a good analysis and a proper interpretation of its results. However, in concluding the section, the Author states that: "…the carefully restricted use of specifically devised sets of artificial data has a place

[1] Corresponding author, e-mail: annalisa.orenti@guest.unimi.it

in the teaching of regression...” (Preece, 1986, p.39). Then he gives a table of artificial data to illustrate the hazard of multiple regression and challenges the reader to spot the simple inbuilt features of these data, but he does not give the solution. He says: “This feature may be unlikely to arise in practice, but the difficulty of recognizing it is a clear warning of some of the problems of multiple regression.“ (Preece, 1986, p. 40). As far as we know Huber and Ronchetti (2009, p. 153) give the solution and state that it can be find immediately by resorting to projection pursuit (Hastie et al., 2009, pp. 389-392). However in our opinion, this approach does not seem in accord with Preece’s provocative proposal, which instead appears to be referred to the use of statistical routine tools for regression analysis.

Finding a solution to the Preece’s problem was proposed as an exercise to students attending the “Scuola di Specializzazione in Statistica Sanitaria e Biometria” at the University of Milan, after their second course of Statistical Methodology, including a set of lessons on multiple regression analysis. This note aims to present the approach developed together with the students. This approach enables deeply investigating the properties of each estimator and appreciating the differences among their behaviours.

2. PRELIMINARY METHODOLOGICAL CONSIDERATIONS

When there are two carriers (see Table 1), the linear model pertinent to the i-th case is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad (1)$$

where y_i is the response variable, $\mathbf{x}_i'$ is the generic carrier vector, ε_i is the random error component with i=1, …, 17. The parameter vector $\boldsymbol{\beta}' = [\beta_0\ \beta_1\ \beta_2]$ and $\widehat{\boldsymbol{\beta}}$ indicates any estimate of $\boldsymbol{\beta}$, so that the residual is $e_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$.

The Preece’s artificial dataset can be processed by the Ordinary Least Squares (OLS) regression method, whose coefficient estimates ($\widehat{\boldsymbol{\beta}}_{OLS}$) are obtained by minimizing, with respect to $\boldsymbol{\beta}$, the objective function:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} e_i^2$$

As it is known that OLS regression method is very sensitive (or susceptible) to outliers and offers very poor performance, it is necessary to check model assumptions by means of single case diagnostics, namely assessing the diagonal elements of the hat matrix, $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ to identify

leverage points, plotting studentized residuals versus predicted values to detect y-outliers, drawing q-q plot of studentized residuals to check for normality. Furthermore plots of studentized residuals versus each carrier can be drawn to check the appropriateness of the linear component of each carrier. Possible failures in fulfilling model assumptions can lead the analyst to conjecture that the data may be heterogeneous in the error component and/or as an effect of the data generation process. The statistical tools suitable for facing such problems are known to be the robust regression estimators.

Among the many robust estimators available nowadays (for a detailed discussion of their characteristics the student is referred to Davies, 1993) we choose some of them belonging to different classes. As such they are based on different computational algorithms, but they are all commonly applied in the statistical literature and are implemented even in open-source statistical software like R (http://www.r-project.org/). Namely they are:

i) Least Absolute Deviation (LAD) estimator (rq function with option tau=0.5 in quantreg package); it was introduced by the astronomer Boskovic, in 1757 (see Birkes and Dodge, 1993, p. 57). The corresponding regression coefficient estimates ($\widehat{\boldsymbol{\beta}}_{\mathrm{LAD}}$) are obtained by minimizing, with respect to $\boldsymbol{\beta}$, the objective function:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |e_i|$$

It can be shown (see Seber and Lee, 2003, p. 79) that it is an M-estimator.

ii) Least Median of Squares (LMS) estimator (Rousseeuw, 1984) (lqs function with option method="lms", in MASS package). The corresponding regression coefficient estimates ($\widehat{\boldsymbol{\beta}}_{\mathrm{LMS}}$) are obtained by minimizing, with respect to $\boldsymbol{\beta}$, the objective function:

$$\min_{\boldsymbol{\beta}} \mathrm{M}[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2] = \min_{\boldsymbol{\beta}} \mathrm{M}\left(e_i^2\right), \quad \text{with } i = 1, \ldots, n$$

where M indicates the Median.

It is a high breakdown point estimator. We recall here that Finite Sample Breakdown Point is the largest proportion of anomalous data that can occur in a sample without entailing the possibility of arbitrary large bias (Donoho and Huber, 1983).

iii) S-estimator (Rousseeuw and Yohai, 1984) (lqs function with option method="S", in MASS package). The corresponding regression coefficient estimates ($\widehat{\boldsymbol{\beta}}_{\mathrm{S}}$) are obtained by minimizing, with respect to $\boldsymbol{\beta}$, the objective function:

$$\min_{\boldsymbol{\beta}} \mathrm{S}(\boldsymbol{\beta}),$$

which is a certain type of robust M-estimate of the scale of the residuals.

It is a high breakdown point estimator.

iv) MM-estimator (Yohai, 1987) (rlm function with options method="MM", psi.weights="biweight" in MASS package) which is a two stage procedure, introduced to improve high breakdown point estimators, toward higher efficiency. In the first stage the high breakdown point S-estimator is used to obtain regression coefficient estimates and scale estimate ($\hat{\sigma}_S$). The regression coefficient estimates are the starting point of the iterative process of the second stage and the scale estimate is kept fixed during iterations of the second stage. The objective function is now:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(e_i)$$

where $\rho(\cdot)$ is the Tukey's biweight function with tuning constant 4.685, to guarantee 95% efficiency.

In dealing with outliers, it is helpful to consider two analytical paradigms suitable when facing two different problems, namely the outlier accommodation problem and the outlier detection problem. In the first one, prediction or other inferences valid for the basic subset are the aims of the analysis. Outliers are assumed to be generated by some uncommon mechanism, alternative to the one generating the basic subset, poorly predicted by the model and not of great interest to the analyst. In the second paradigm, outliers are of primary concern: "Parameter estimates are needed merely to investigate the discordance of certain observations." (Ruppert and Simpson, 1990, p. 644).

Finding how Preece generated the whole set of 17 data reported in Table 1 is the goal of our exercise, thus the second paradigm seems appropriate to look for the solution.

Table 1: Preece's artificial dataset

ID	x_1	x_2	y
1	9.1	5.4	30.9
2	10.7	8.0	58.8
3	11.4	7.3	56.7
4	13.8	7.9	67.5
5	14.1	3.9	32.4
6	14.5	4.1	46.7
7	8.3	3.7	13.2
8	12.6	6.4	55.2
9	7.3	6.3	33.6
10	7.9	6.4	36.4
11	9.2	7.2	47.2
12	15.8	5.9	64.5
13	12.9	6.4	51.3
14	5.1	5.3	17.5
15	10.1	5.5	34.8
16	10.3	2.6	19.4
17	10.0	7.8	55.2

3. RESULTS

3.1 ORDINARY LEAST SQUARES REGRESSION

As a first step the Preece's artificial dataset, given in Table 1, was processed by the OLS method. The coefficient estimates are $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (-43.477, 3.724, 7.778)'$ and the corresponding regression ANOVA is reported in Table 2. We observe that the Sums of Squares of x_1 (Table 2.A) and $x_1|x_2$ (Table 2.B) are similar and the same happens for the Sums of Squares of x_2 (Table 2.B) and $x_2|x_1$ (Table 2.A); furthermore the correlation coefficient between x_1 and x_2 is $r(x_1, x_2)=0.008$. These findings justify the use of both carriers in the regression model. With regard to leverages, no case exceeds the empirical cut-off 2p/n (0.353), where p=3 is the number of regression parameters and n=17 is the number of cases in the dataset. Considering the plot of studentized residuals versus predicted values, reported in Figure 1, we can see that no particular pattern emerges and all studentized residuals are within the empirical thresholds (-2, 2), except case 5. Plots of studentized residuals versus each carrier, reported in Figure 2, do not enlighten any non linear effect of the carriers. An interaction term between x_1 and x_2, added to model (1) is far from being significant (ANOVA table not shown). From these findings a misspecification

of the regression model does not appear. On the other hand the studentized residual q-q plot, drawn in Figure 3, raises doubts about the presence of a unique normal distribution of the residuals and/or the homogeneity of the data in the sample; it is to explore this possibility further that we resorted to robust regression analysis.

Table 2: Regression ANOVA for model (1)

A

Source of variability	Degrees of freedom	Sum of Squares	Mean square
x_1	1	1874.60	1874.60
$x_2\|x_1$	1	2438.88	2438.88
Residual	14	131.90	9.42

B

Source of variability	Degrees of freedom	Sum of Squares	Mean square
x_2	1	2471.80	2471.80
$x_1\|x_2$	1	1841.70	1841.70
Residual	14	131.90	9.42

A and B differ for the order of carriers inclusion in the model.

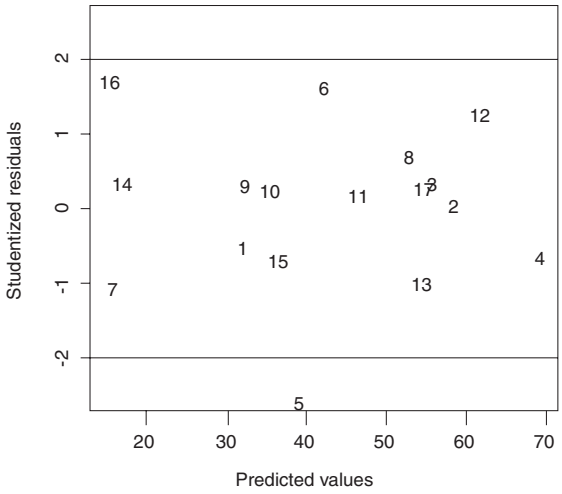

Figure 1: OLS regression model: plot of studentized residuals versus predicted values, together with pertinent threshold values

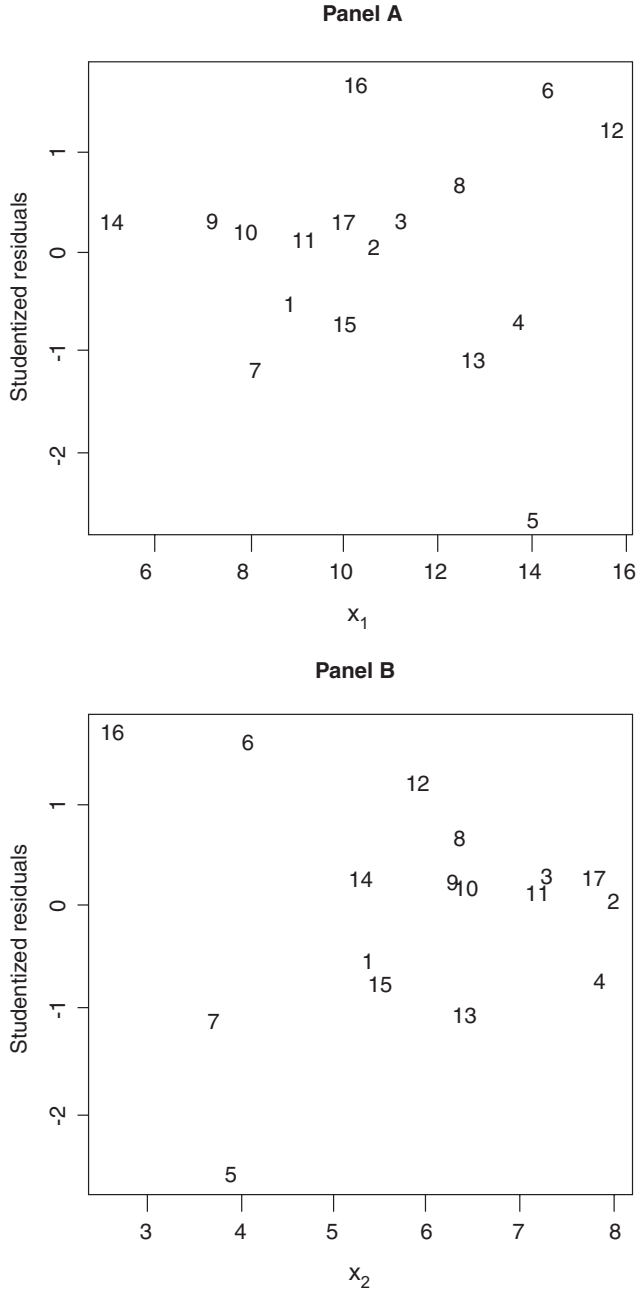

Figure 2: Panel A: plot of studentized OLS residuals versus x_1; Panel B: plot of studentized OLS residuals versus x_2

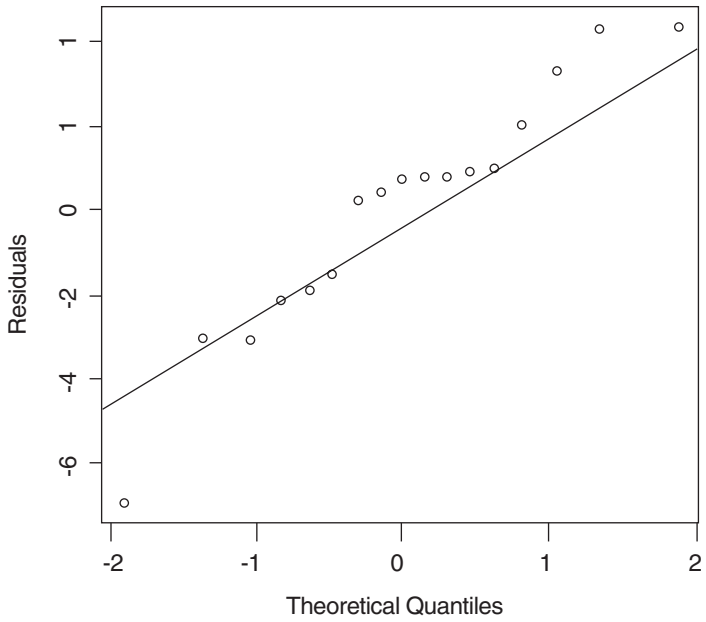

Figure 3: Studentized residual q-q-plot for OLS regression

3.2 ROBUST REGRESSIONS

The dataset was processed by the four previously mentioned robust regression methods and two different scenarios emerge, the first one supported by LAD and LMS procedures and the second one supported by S-estimator and MM-estimator.

The LAD and LMS regression coefficient estimates are $\widehat{\boldsymbol{\beta}}_{\mathrm{LAD}} = \widehat{\boldsymbol{\beta}}_{\mathrm{LMS}} = (-40, 4, 7)'$. They enable computing the 17 predicted values: $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\mathrm{LAD}}$; nine of these form the subset [1]: (2, 3, 6, 8, 10, 11, 12, 14, 16); they are equal to the corresponding y_i, thus they give null residuals e_i, whereas the remaining eight cases, forming the subset [2], do not. At this point one may argue that data in the subset [2] could have been generated according to two different alternatives, namely: i) response values could be the predicted values of the same model as in the subset [1] but with the addition of a random error term (see model (1)); otherwise ii) they could be the predicted values of a different model still to be found. To try to distinguish between the two alternatives, it appears convenient to fit the following model to the whole dataset by means of the OLS method:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_0 d_i + \gamma_1 d_i x_{i1} + \gamma_2 d_i x_{i2} + \varepsilon_i \qquad (2)$$

where d_i is a dummy variable assuming value 0 for the cases belonging to the subset [1] and value 1 for the cases belonging to the subset [2]. The quantities γ_0, γ_1 and γ_2 account for the possible differences of the regression coefficients between the model $M_1(\boldsymbol{\beta})$ fitted to the subset [1] and the model $M_2(\boldsymbol{\vartheta})$ fitted to the subset [2]. In the first alternative of data generation, previously outlined, the estimates of γ_0, γ_1 and γ_2 are expected to be null, whereas in the second alternative they are expected to be different from 0, so that the 3 components of $\boldsymbol{\vartheta}$ are estimated as $\left(\hat{\beta}_0 + \hat{\gamma}_0, \hat{\beta}_1 + \hat{\gamma}_1, \hat{\beta}_2 + \hat{\gamma}_2\right)'$.

The OLS estimates of model (2) result in: $\widehat{\boldsymbol{\beta}} = (-40, 4, 7)'$ and $\widehat{\boldsymbol{\gamma}} = (-5, -1, 2)'$, so that $\widehat{\boldsymbol{\vartheta}} = (-45, 3, 9)'$. Table 3 reports the regression ANOVA on model (2). The null Error Sum of Squares suggests that even in the subset [2] the residuals are null, thus y_i exactly equals $\hat{y}_i$, as predicted by model $M_2(\boldsymbol{\vartheta})$.

Table 3: Regression ANOVA for model (2)

Source of variability	Degrees of freedom	Sum of Squares	Mean square
x_1	1	1874.60	1874.60
$x_2 \mid x_1$	1	2438.88	2438.88
subtotal (x_1, x_2)	2	4313.48	
$d \mid x_1, x_2$	1	66.94	66.94
$d*x_1 \mid d, x_1, x_2$	1	26.07	26.07
$d*x_2 \mid d*x_1, d, x_1, x_2$	1	38.89	38.89
subtotal($d, d*x_1, d*x_2 \mid x_1, x_2$)	3	131.90	
Error	11	0	0

Note that the estimate $\widehat{\boldsymbol{\vartheta}}$ is equal or very similar to the corresponding regression coefficient estimates obtained by S-estimator $\left(\widehat{\boldsymbol{\beta}}_S = (-45, 3, 9)'\right)$ and by MM-estimator $\left(\widehat{\boldsymbol{\beta}}_{MM} = (-45.20, 3.03, 8.98)'\right)$ respectively.

It may be surprising that the LAD and LMS estimators give regression coefficient estimates equal to those of subset [1], whereas S and MM estimators give regression coefficient estimates equal to those of subset [2]. This can be justified by the differences in the objective functions to be minimized and by the use of different computational algorithms. For a detailed discussion the student is referred to Barrodale and Roberts (1974) and to Marazzi (1993).

However, under the outliers detection paradigm mentioned in section 2 (Preliminary methodological considerations), S and MM-estimators agree with LAD and LMS procedures in considering $y = -40 + 4x_1 + 7x_2$ suitable for fitting the nine cases in subset [1] and $y = -45 + 3x_1 + 9x_2$ suitable for fitting the eight cases in subset [2].

3.3 ERROR MEAN SQUARE ESTIMATES

It is now instructive to compare the regression ANOVA exhibits obtained by OLS method in fitting the artificial dataset "ignoring" the data generation process (model (1), Table 2) and "knowing" the data generation process (model (2), Table 3). A total of 3 degrees of freedom (d.f.) is spent to account for the previously mentioned increase of information; in fact 14 d.f. of the Residual Sum of Squares in Table 2 reduce to 11 d.f. of the Error Sum of Squares in Table 3. Correspondingly the Residual Sum of Squares = 131.9 (Table 2) is split in the three sources of variability (d, $d*x_1$, $d*x_2$), whose sum of Sum of Squares equals 131.9. Consequently no Error Sum of Squares is available and this agrees with the fact that in this artificial dataset the responses (y) equal the values predicted ($\hat{y}$) by one model for subset [1] and by an alternative model for subset [2]. Actually the Residual Sum of Squares in model (1) is not due to the error component but to the fact that the model is not "saturated", i.e. not all the sources of variability have been explained.

In view of these comments we reconsider the studentized residuals computed at the beginning of the analysis ignoring the dataset generation process. The Residual Mean Square $\frac{131.9}{14} = 9.42$ is actually a biased estimate of the true error variance $\sigma^2=0$, as OLS assumption of data homogeneity is not fulfilled in this dataset. This drawback is emphasized in this case by the absence of error variance, but it may happen even in the presence of $\sigma^2\neq0$, provided that the difference between the models generating the two subsets of the dataset is sufficiently large.

3.4 MIXTURE MODEL VIA EM

One further available tool is obtained by applying the Expectation Maximization (EM) procedure (Dempster et al., 1977) to face the problem of mixture models (flexmix function in flexmix package), as suggested by Aitkin and Wilson (1980).

The EM algorithm is a general iterative method to find the maximum-likelihood estimate of the parameters of an underlying probability density for a given data set. In its simplest form the probability model for the data is given by the mixture of two Normal densities:

$$f(y) = (1 - \delta)f_1(y) + \delta f_2(y)$$

where (1-δ) is the proportion of data belonging to the first subset with probability density $f_1(y)$, and $f_2(y)$ refers to the probability density of the second subset.

Resorting to robust method one assumes that the main part (100%-δ, 0<δ<50%) of the data (basic subset or bulk) is concordant with a single model, say $M_1(\beta)$, whereas the remaining δ is discordant from it. The latter includes different kinds of outlying observations: y-outliers, x-outliers or both.

Preece dataset was processed by mixture model using EM procedure. Two different clusters were identified: one containing observations 1, 4, 5, 7, 9, 13, 15, 17 and the other containing observations 2, 3, 6, 8, 10, 11, 12, 14, 16. These clusters overlap subset [2] and [1] respectively. OLS regression models were fitted to each of them. As expected the regression coefficient estimates were equal to those previously reported and the corresponding error variance estimates were null.

4. ISSUES DESERVING A DEEP DISCUSSION

Three aspects of this artificial dataset deserve to be discussed with the students:

i) the small sample size (n=17). Its ratio to the number of carriers (p-1=2) corresponds almost to the minimum widely accepted as a rule of thumb, 8;

ii) 9 cases belong to subset [1] and 8 cases belong to subset [2]. Is it meaningful in this context to refer to an outliers accommodation paradigm?

iii) the choice of $y_i = \hat{y}_i$, i.e. $\sigma^2 = 0$. Possible extension of the analysis can be performed after adding i.i.d. error terms.

Finally, it is to be said that the exercise was found to be appealing for enabling the students to have a systematic refreshment of the topics developed in the whole course on regression analysis.

REFERENCES

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. In *Technometrics*, 22(3): 325-331.

Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478: Solution of an overdetermined system of equations in the l1-norm. In *Communications of the Association for Computing Machinery*, 17: 319-320.

Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*. John Wiley and Sons, New York.

Brownlee, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. John Wiley and Sons, New York.

Davies, P. L. (1993). Aspects of robust linear regression. In *The annals of statistics*, 21 (4): 1843-1899. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical society, Series B*, 39: 1-38.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second edition*. Springer, New York.

Huber, P.J. and Ronchetti E. M. (2009). *Robust Statistics*. John Wiley and Sons, New York. Second edition.

Marazzi, A. (1993). *Algorithms, Routines and S Functions for Robust Statistics*. Wadsworth and Brooks/Cole.

Preece, D.A. (1986). Illustrative Examples: Illustrative of What? In *Journal of the Royal Statistical Society. Series D (The Statistician)*, 35 (1): 33-44.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. In *Journal of the American Statistical Association*, 79 (388): 871-880.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In Robust and nonlinear time series (J. Franke, W. Härdle and D. Martin, eds.). *Lecture notes in statistics* 26: 256-272. Springer, New York.

Ruppert, D. and Simpson, D. G. (1990). Unmasking Multivariate Outliers and Leverage Points: Comment. In *Journal of the American Statistical Association*, 85 (411): 644-646.

Seber, G. A. F. and Lee, A. J. (2003), *Linear Regression Analysis*. John Wiley and Sons, Hoboken, New Jersey. Second edition.

Yohai, V. J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. In *The Annals of Statistics*, 15 (2): 642-656.