Robust Regression:
A Weighted Least Squares Approach

Samprit Chatterjee
Martin Mächler

Dept. of Statistics & OR, New York University
44 West 4th Street, Room 8-67
New York, NY 10012-1126, USA
`schatter@stern.nyu.edu`

Seminar für Statistik, Swiss Federal Inst. of Technology (ETH)
CH-8092 Zurich, Switzerland
`maechler@stat.math.ethz.ch`

Key Words: outliers; leverage points; influence; M-estimator; iterative reweighting; masking.

ABSTRACT

Robust regression has not had a great impact on statistical practice, although all statisticians are convinced of its importance. The procedures for robust regression currently available are complex, and computer intensive. With a modification of the Gaussian paradigm, taking into consideration outliers and leverage points, we propose an iteratively weighted least squares method which gives robust fits. The procedure is illustrated by applying it on data sets which have been previously used to illustrate robust regression methods. It is hoped that this simple, effective and accessible method will find its use in statistical practice.

1 INTRODUCTION

Considerable work has been done on robust regression in the last 20 years. In spite of all the advances, robust regression is not widely used in practice. Several reasons have been advanced for this state of affairs, some of them quite superficial, others substantive. Among the superficial reasons are (i) robust methods have not been presented to the statistical community vigorously ("not sold or marketed well") (ii) no commercial software is available for the general user, though this picture is changing. Although there is some truth to (ii), we feel (i) does not hold water. A look through the statistical journals will convince the reader of the validity of the point we are making. Robust regression methods have concentrated primarily, though not exclusively, in guarding against outliers. Guarding against outliers is very important but there are equally troubling problems in model fitting. Are the commonly proposed methods robust against dependent errors, long term correlations, model mis-specifications? Robust regression procedures currently available are complex because they are "optimal". Can we be satisfied with near "optimal procedures" which are simple, easy to understand, and mostly work? Do we ever need methods with a 50% breakdown point? Would any statistician fit a regression model to a data set which has 50% contaminated data without a pre-screening procedure? We feel that all (most) statisticians would be satisfied with a procedure which would be robust against 20 to 25 percent contamination.

In this paper we propose a robust method of fitting a linear model which is simple, easy to understand and compute. The method is based on iterative least squares and consequently lies in mainstream statistics. After describing our algorithm, we apply it to few well-known published data sets which have been used to illustrate the effectiveness of robust regression procedures. The proposed method performs as effectively as the existing methods. Our procedure is to be preferred on grounds of simplicity, interpretability, and ease of understanding.

2 PRELIMINARY NOTATION AND TERMINOLOGY

We consider the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ $(n\times 1)$, $\mathbf{X}$ $(n\times k)$, $\boldsymbol{\beta}$ $(k\times 1)$, and $\boldsymbol{\varepsilon}$ $(n\times 1)$. $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$. The predicted values $\hat{\mathbf{y}}$ of $\mathbf{y}$ and the residuals $\mathbf{r}$ are defined as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = P\mathbf{y} && (2) \\ \mathbf{r} &= \mathbf{y} - \hat{\mathbf{y}} = (I-P)\mathbf{y} && (3) \end{aligned}$$

where

$$P = [p_{ij}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{4}$$

is the leverage matrix or projection matrix. Points which are not well fitted by the linear model will have large residuals and are called *outliers*. Points which are outlying in the x-space will be called *leverage points*. The notation and terminology is quite standard and can be found in any text on regression analysis.

3 THE GAUSSIAN PARADIGM AND ITS MODIFICATION

In standard regression analysis the linear model is fitted by the Principle of Least Squares which we call the Gaussian paradigm. The paradigm may be stated as: for a body of data, the "best" fit to a given model is obtained by minimizing the sum of squares of the residuals. The problems associated with the application of least squares to model fitting has been discussed widely and extensively (a partial list would be Atkinson, 1985; Chatterjee and Hadi, 1988; Cook and Weisberg, 1982; Hampel, Ronchetti, Rousseeuw and Stahel, 1986; Rousseeuw and Leroy, 1987). The problems from the least squares fit arise due to two main causes. These may be summarized by saying that the Gaussian paradigm, as enunciated earlier, gives too much weight to outliers and high leverage points. This feature distorts the fitted line. One can start however from this fit, and looking at various regression diagnostics, come to a valid and acceptable analysis of the data. This approach has been described in Chatterjee and Price, 1991; Daniel and Wood, 1980; Draper and Smith, 1981; Belsley, Kuh and Welsch, 1980; Weisberg, 1980.

The difficulties associated with the least squares fit can be removed by modifying the Gaussian paradigm. We propose to modify the paradigm as follows: For a given model, the "best" fit is obtained by minimizing the sum of squares of the residuals without allowing observations which do not fit the given model, or are far from the main body of the data from exercising substantial influence over the fit. The modified paradigm will generate the fitting principle for the procedure we propose.

The fitting principle we use can now be stated. The "best" fit for a data set to a given model is obtained by a fitting procedure which gives less weight to observations lying far from the main body of the data, and also to those which are not fitted (well) by the model. In the next section we translate this fitting principle into a computing algorithm employing iteratively weighted least squares.

4 THE FITTING ALGORITHM

The algorithm described below does not have the disadvantages of OLS. The algorithm is a weighted least squares (*WLS*) algorithm which gives low weight to points with (i) high leverage and (ii) large residuals. Since the weights are determined by the residuals, and as these change from iteration to iteration, the procedure is an iterative one. We start with an initial fit and iterate. We describe the algorithm formally and then comment on its rationale. Notation: We use Q^j to denote the value of Q in iteration step j, and ":=" means "*is set to*".

Step 0 **:** Given weights $w_i^0 = 1/\max(p_{ii}, \bar{p})$, where $\bar{p} = k/n$, and p_{ii}, as in (4), the leverage of the i^{th} point, compute the weighted least squares regression

$$\widehat{\boldsymbol{\beta}}^0 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i^0 \left(y_i - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2 \tag{5}$$

Step j **:** $(j = 1, 2, \ldots)$

Compute new weights from the residuals of the last fit,

$$\begin{aligned} \mathbf{r}^{j-1} &= \mathbf{y} - \hat{\mathbf{y}}^{j-1} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{j-1}, \\ w_i^j &= (1 - p_{ii})^2 \,/\max\left(\left|r_i^{j-1}\right|, \mathrm{med}_{i'}\left|r_{i'}^{j-1}\right|\right),\ i = 1, \ldots, n, \\ \text{where}\quad \mathrm{med}_i\, x_i &:= \mathrm{median}(x_1, x_2, \ldots, x_n). \end{aligned}$$

Compute the weighted least squares regression

$$\widehat{\boldsymbol{\beta}}^j = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i^j \left(y_i - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2 \tag{6}$$

This is carried on till convergence.

4.1 DISCUSSION OF THE WEIGHTS AND THE ESTIMATOR

In step 0 of the algorithm, the fit is obtained by down weighting points with high leverage. Note also that the weight of any point is bounded in the spirit of the Hampel-Krasker-Welsch estimator (see Hampel, 1978; Krasker and Welsch, 1982).

In step j, the weights are a function of the residuals as well as leverage. The actual function used comes from the sample influence of the point as determined by the empirical influence curve based on $n-1$ observations after deleting the i-th observation (Chatterjee and Hadi, 1988, p. 116). Note also that the weights are bounded in this step as in step 0. The weights in step j incorporate the ideas of

Mallows (1975), Merrill and Schweppe (1971) as well as those of Hampel, Krasker and Welsch mentioned above. A link to the proposed algorithm and the work of Coakley and Hettmansperger (1993) should also be noted. As will be shown in the next section, our algorithm gives similar numerical results as the procedures already cited.

Our algorithm computes a "Mallows-type" M-estimate of regression, since step j is nothing more than the "modified weights" algorithm of Huber (1981, p.183 ff) with an additional factor of $(1-p_{ii})^2$ which depends on $\mathbf{X}$. The weights w_i^j in step j above can be expressed (dropping the dependence on j for the moment) as

$$\begin{aligned}
w_i &= v_i^2 \tilde{w}_i \quad \text{where } v_i := 1 - p_{ii} \quad \text{and} && (7)\\
\tilde{w}_i &= 1/\max\left(|r_i|, \text{med}_{i'} |r_{i'}|\right) \\
&= \min\left(1/|r_i|, 1/(c_* \hat{\sigma})\right) && (8)\\
&= \frac{1}{c_* \hat{\sigma}} \min\left(1, \frac{c_*}{|r_i|/\hat{\sigma}}\right) && (9)\\
&= \frac{1}{c_* \hat{\sigma}}\, \psi_{c_*}^H\left(\frac{r_i}{\hat{\sigma}}\right) / \left(\frac{r_i}{\hat{\sigma}}\right), && (10)
\end{aligned}$$

where $\hat{\sigma}$ is the (scaled) MAD (median absolute deviation) estimator $\hat{\sigma} = \frac{1}{c_*}\,\text{med}_i |r_i|$, with $c_* = \Phi^{-1}(\frac{3}{4}) \approx .6745$ (for consistency of $\hat{\sigma}$), and ψ_c^H is Huber's function, $\psi_c^H(r) := \max(-c, \min(c, r)) = r \cdot \min(1, c/|r|)$.

In the appendix, we will show that the estimated standard errors as provided by standard WLS have to be multiplied by a factor of 1.38373 in order to adjust for the fact the our weights depend on the y_i and therefore are *random*.

Our estimator is *not* a bounded influence estimator in the strict sense, the factor $(1-p_{ii})^2$ however properly downweights influential observations (high leverage points) as long as there is not substantial "masking".

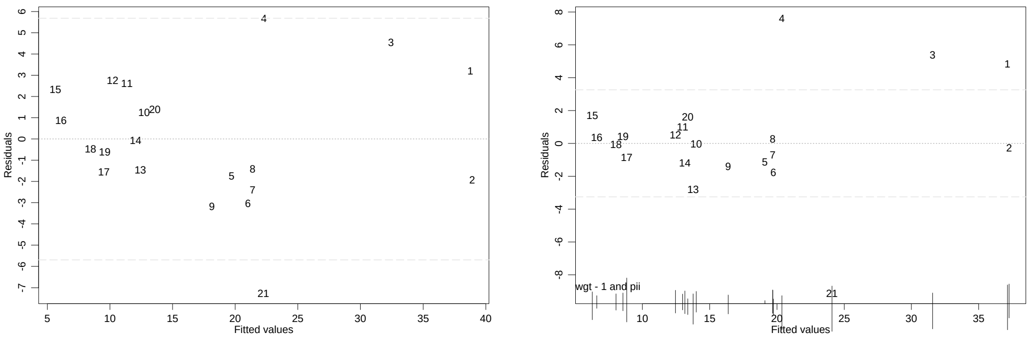
We should also point out that, although the weights are not fixed, the algorithm provides "least squares like" residuals to do diagnostic checking. In fact we can use all the graphical diagnostics (i.e. all the residual plots) to look at model violations such as lack of linearity, time dependence of errors, unusual points etc. Our fitting algorithm gives us a good start, better than the OLS fit, to look at the problems in the data and the deficiencies in the fitted model. Our fitted model does not suffer from the distorting effects of the OLS.

5 ILLUSTRATIVE EXAMPLES

We give the results from the application of our method of fitting to three sets of data. These data sets have all been used to illustrate robust regression methods.

a. Stack Loss Data

The stack loss data was originally given by Brownlee (1965) and has been analyzed extensively. It would not be incorrect to say that more papers have been written on this data than the number of observations it contains. We will not give all the references, but Daniel and Wood (1980) is a good starting source. The stackloss of ammonia in an industrial operation is to be explained by three variables, airflow, temperature and concentration. After extensive analysis it was concluded that observations 1,3,4, and 21 were outliers. A plot of the residuals against the fitted values (which will subsequently be referred to as the Tukey-Anscombe (TA) plot from the least squares fit does not reveal the outliers (Figure 1a). The TA plot from



(a) Least Squares (b) New IRLS approach.

Figure 1: Tukey-Anscombe [TA] plots of stackloss data; LS and our IRLS.

our robust fit (Figure 1b) reveals the offending points clearly. The residual plot for the remaining points shows no structure and consequently the model is well fitted. The model given by our method is

$$\begin{array}{lllll} \text{STACKLOSS} = & -37.00 & +0.84\ \text{AIRFLOW} & +0.63\ \text{TEMP} & -0.11\ \text{CONC} \\ & (10.54) & (.14) & (.34) & (.14) \end{array}$$

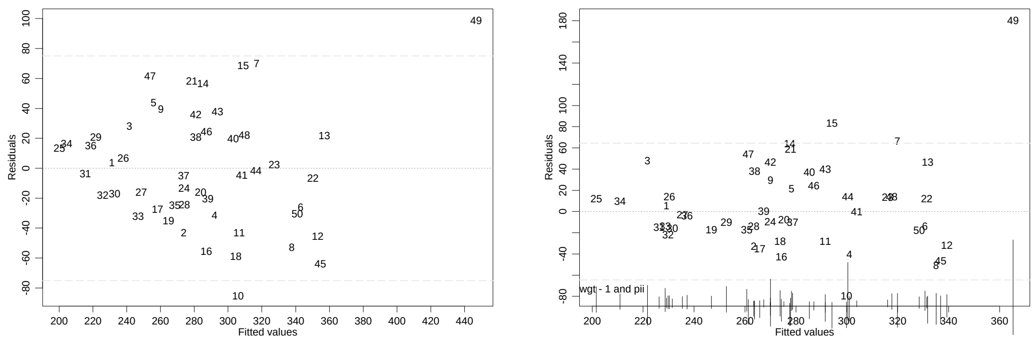
and should be compared with other published values.

The standard errors given for the coefficients are the weighted least squares (WLS) values, for consistency multiplied by the factor 1.38373 as computed in the appendix.

Figure 1b also shows the weights given to each point in the final fit. Points with very low weights should now be examined for their "deficiency", be it due to departure from the model or its anomalous nature with respect to the dependent variables. If it is warranted, the model may be modified so that a point which was an outlier is no longer an outlier.

b. Education Expenditure Data

The data deals with education expenditure for 50 U.S. states and was given in Chatterjee and Price (1991, p. 131) and was used by Rousseeuw and Leroy (1987) to illustrate robust fitting. The aim was to explain the per capita expenditure on public education projected in 1975 (y), by three explanatory variables x_1 (per capita income in 1973), x_2 (number of residents per thousand under 18 years of age in 1974), and x_3 (the number of residents per thousand living in urban areas in 1970). The TA plot for the least squares fit is given in Figure 2a. The plot



(a) Least Squares (b) New IRLS approach.

Figure 2: Tukey-Anscombe [TA] plots of education expenditure data; LS and our IRLS.

shows clearly that there is an outlier (Alaska, observation 49) and the residuals have an expanding V shaped structure showing the presence of heteroscedasticity. A satisfactory analysis of the data can be arrived at after deleting the outlier, and doing a weighted least squares analysis after grouping the states by geographical region. The reader is referred to Chatterjee and Price (1991) for further details.

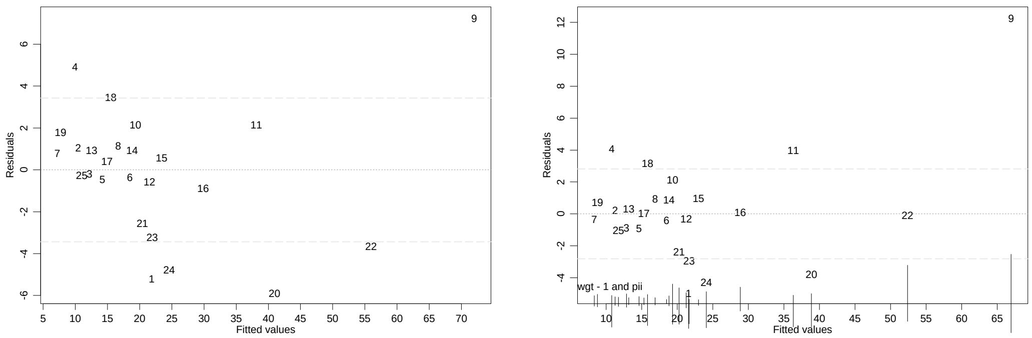
Figure 2b gives the TA plot for the model fitted by our robust procedure. Alaska (observation 49) is clearly an outlier. The plot from the robust fit does not exhibit heteroscedasticity and consequently can be taken as a satisfactory fit. The fitted model is given by

$$y = -292.63 + 0.050\, x_1 + 0.88\, x_2 + 0.072\, x_3.$$

Our values and plot for the data should be compared with Rousseeuw and Leroy (1987, p. 111). The coefficients and the plots are very similar, although our estimates are obtained by standard statistical software.

c. Soft Drink Delivery Data

The data set was originally given by Montgomery and Peck (1992, p. 116) and has been used by Coakley and Hettmansperger (1993) to illustrate several robust procedures. The data consists of 25 observations and attempts to explain the soft drink delivery time in minutes by number of cases and distance in feet. The TA plots for the least squares and our robust fit are given in Figures 3a and



(a) Least Squares (b) New IRLS approach.

Figure 3: Tukey-Anscombe [TA] plots of soft drink delivery data; LS and our IRLS.

b respectively. The estimated coefficients and the standard errors for four procedures are given in Table 5. Two of the procedures, Mallows One-step estimator

TABLE I

Estimators and Standard Errors for Delivery Data

Term	LS	M1S	S1S	IRLS
Intercept	2.34 (1.10)	4.30 (.94)	4.27 (1.08)	3.94 (.86)
Cases	1.62 (0.17)	1.51 (.16)	1.45 (.17)	1.44 (.14)
Distance	0.014(.004)	0.011(.003)	0.012(.005)	0.013(.003)

(M1S), and Schweppe one-step estimator (S1S) have been proposed by Coakley and Hettmansperger (1993). Also shown are the least squares estimator (LS) and the proposed iterative weighted least squares (IRLS). Numerically, there is not much difference between the M1S, S1S, and IRLS. The simplicity and the ease of computation makes IRLS highly attractive.

Further Examples
We have applied our method to several other real and artificial data sets. These data sets are reputed to be "difficult". Our method deals effectively with the data set presented by Hadi and Simonoff (1993). Our method identifies the three points which are outliers and high leverage points. It is then able to overcome the effect of these points by downweighting them in the fitting.

We applied our method to Atkinson's Scottish Hill Race data presented in Chatterjee and Hadi (1988). In this data set there is a masking effect. Our weighting scheme is able to detect the 3 offending points and downweight them. Several problematic data sets are presented in Rousseeuw and Leroy (1987). Our method provided a satisfactory fit to the Belgian telephone data. It did not perform well on the Hawkins–Bradu–Kass and the modified wood gravity data. These are two artificial data sets containing high leverage points some of which are outliers and others which are not. The problem with the data sets are transparent from the pairwise scatter plots of the variables (the "*scatter plot matrix*"). Our method does not sufficiently downweight the offending points. In general, our method is not very effective when there is extensive masking as it is in these two data sets. It must be pointed out however that these are artificial data sets, deliberately altered, containing a high percentage of masking points. This situation is rarely encountered in practice.

We have applied our method to the data set provided by Antille and El May (1992). In this example complex methods such as Least Median of Squares and Least Trimmed Squares fail, as well as MM–estimators based on them. Our method however works perfectly. This leads us to believe that the method proposed in this paper will be useful to all practicing statisticians and lead to the adoption of robust regression methods in everyday application.

6 SUMMARY

Models fitted to data should not be sensitive to a few extreme observations. Robust fits are highly desirable. Currently available robust regression methods are highly complex and very computer intensive. They are not accessible (both conceptually and operationally) to everyday statistical users. We have proposed an iteratively weighted least squares algorithm which produces robust fits. The performance of the new procedure is illustrated by its ability to match the performance of previously published procedures. We hope that the simplicity and the intuitive appeal of our procedure will lead to its adoption as a practical everyday tool in statistical analysis — in places and circumstances where robust regression has been unavailable before.

The IRLS algorithm is implemented by a code written in S-Plus. We will be glad to make it available to any interested researcher. The implementation is very simple. The procedure can be carried out on any statistical package which has a routine for doing weighted least squares and the ability to calculate leverage values p_{ii} and least squares residuals. Code for this procedure is now also available in a SAS IML version as well as a Minitab macro.

APPENDIX
VARIANCE ESTIMATION

Several years back (e.g., Holland and Welsch, 1977), it was noted that WLS (weighted least squares) can be used iteratively to compute robust regression M-estimates. As Street, Carroll and Ruppert (1988) further pointed out, the WLS variance estimate, $\mathrm{var}_{\mathrm{WLS}}\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, is not correct but rather

$$\begin{aligned} \mathrm{var}\widehat{\boldsymbol{\beta}} &= \sigma^2 V_0 \cdot (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{d} \cdot \mathrm{var}_{\mathrm{WLS}}\widehat{\boldsymbol{\beta}}, \quad \text{where} && (11) \\ \frac{1}{d} &= V_0 \cdot \frac{\mathbf{E}\left[\psi(\tilde{\varepsilon})/\tilde{\varepsilon}\right]}{\mathbf{E}\left[\tilde{\varepsilon}\psi(\tilde{\varepsilon})\right]}, && (12) \\ V_0 &= \frac{\mathbf{E}\left[\psi^2(\tilde{\varepsilon})\right]}{(\mathbf{E}\left[\psi'(\tilde{\varepsilon})\right])^2} = \frac{A}{B^2}, && (13) \end{aligned}$$

and $\tilde{\varepsilon} = \varepsilon/\sigma$ is the rescaled error of model (1). We have seen above that ψ is Huber's function $\psi^H_{c_*}$ with $c_* = \Phi^{-1}(\frac{3}{4}) \approx .6745$. Note that the weight factor $v_i^2 = (1 - p_{ii})^2$ in (7) is treated as non-random observation weight, being part of the design matrix $\mathbf{X}$ in equation (11).

In order to be consistent under the "null model" of Gaussian errors, we compute the expected values in (12) and (13) under $\tilde{\varepsilon} \sim \mathcal{N}(0,1)$. This computation gives $\mathbf{E}\left[\psi'(\tilde{\varepsilon})\right] = 2\Phi(c_*) - 1 =: B$, $\mathbf{E}\left[\psi^2(\tilde{\varepsilon})\right] =: A = B - 2c_*(\phi(c_*) - c_*(1 - \Phi(c_*)))$, $\mathbf{E}\left[\tilde{\varepsilon}\psi(\tilde{\varepsilon})\right] = B$, and $\mathbf{E}\left[\psi(\tilde{\varepsilon})/\tilde{\varepsilon}\right] =: C = B + \frac{c_*}{\sqrt{2\pi}} E_1(c_*^2/2)$, where $E_1(x)$ is the exponential integral function $E_1(x) := \int_x^\infty e^{-t}/t\, dt$ (Abramowitz and Stegun, 1972, Ch. 5, p. 228f).

The numerical values in our case are $B = \frac{1}{2}$, $A = .298794$ (hence, $V_0 = 1.195$, i.e., efficiency $1/V_0 = 83.7\%$), $E_1(c_*^2/2) = E_1(.2275) = 1.11877$, $C = 0.80102$ and hence correction factor $1/d = CA/B^3 = 1.91472$ for the variance or factor $\sqrt{1/d} = 1.38373$ for standard errors.

ACKNOWLEDGMENTS

We would like to thank Frank Hampel for his many thoughtful comments and encouragement during the course of this work. For one of the authors (S.C.), he provided a stimulating working environment at ETH Zurich. We would also like to thank Andreas Ruckstuhl and Werner Stahel for their comments and computational help. We would like to thank the associate editor and the referees for their very careful reading and enlightening comments from which we benefited greatly.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*, Dover, New York.

Antille, G. and El May, H. (1992). The use of slices in the LMS and the method of density slices: Foundation and comparison, *in* Y. Dodge and J. Whittaker (eds), *Computat. Statist: Proc. 10th Symp. Computat. Statist., Neuchatel*, Vol. 1, Physica-Verlag, pp. 441–445.

Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Oxford Statistical Science Series, Clarendon Press, Oxford.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, Wiley, N. Y.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd edn, Wiley, N. Y.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, N. Y.

Chatterjee, S. and Price, B. (1991). *Regression Analysis By Example*, 2nd edn, Wiley, N. Y.

Coakley, C. W. and Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator, *J. Am. Statist. Assoc.* **88**(423): 872–880.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, N. Y.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N. Y.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edn, Wiley, N. Y.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *J. Am. Statist. Assoc.* **88**: 1264–1272.

Hampel, F. R. (1978). Modern trends in the theory of robustness, *Math.Operationsforschung Statist.,Ser.Statist.* **9**: 425–442.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N. Y.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares, *Comm. Statist., Theory Meth.* **6**: 813–888.

Huber, P. J. (1981). *Robust Statistics*, Wiley, N. Y.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation, *J. Am. Statist. Assoc.* **77**: 595–604.

Mallows, C. L. (1975). On some topics in robustness, *Technical report*.

Merrill, H. M. and Schweppe, F. C. (1971). Bad data suppression in power system static state estimation, *IEEE Trans. Power App. Syst.* **PAS-90**: 2718–2725.

Montgomery, D. C. and Peck, E. A. (1992). *Introduction to Linear Regression Analysis*, Wiley, N. Y.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N. Y.

Street, J. O., Carroll, R. J. and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares, *Amer. Statist.* **42**(2): 152–154. *(Corr: **43** p.69)*

Weisberg, S. (1980). *Applied Linear Regression*, Wiley, N. Y.