BULLETIN *of the* MALAYSIAN MATHEMATICAL SCIENCES SOCIETY

Bull. Malaysian Math. Sc. Soc. (Second Series) **24** (2001) 29-44

The Application of Minimax Fit in Linear Model and its Extension in Generalised Linear Models

NOR AISHAH HAMZAH AND DAUD YAHAYA

Institute of Mathematical Sciences, Universiti Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia.
e-mail: nah@mnt.math.um.edu.my

Abstract. The minimax fitting procedure is employed in some robust parameter estimation in the normal linear model. In this paper, we shall reformulate the procedure by using the deviance residual measures instead of the usual residual measures for the robust fitting procedure. This extends the range of application from the normal distribution to a more general exponential family of distributions. Special cases of the exact solution for the single parameter case and its algorithm will be discussed.

1. Introduction

We consider the standard linear regression model

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

where $\boldsymbol{y}$ is an n-vector of responses, $\boldsymbol{X}$ is an $n \times p$ matrix representing explanatory variables with rank $p\,(< n)$, whose *i*-th row is $\boldsymbol{x}_i^T$, $\boldsymbol{\beta}$ is a *p*-vector of unknown parameters to be estimated, and $\boldsymbol{e}$ is an *n*-vector of unknown errors with mean, $E(e) = 0$, and variance, $\mathrm{var}\,(e) = \sigma^2 I$, where $\boldsymbol{I}$ is the identity matrix of size *n*.

The most commonly used method in estimating $\boldsymbol{\beta}$ is the method of Least Squares (LS) in which $\hat{\boldsymbol{\beta}}$ minimises over $\boldsymbol{\beta}$ the sum of squared residuals,

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 \equiv \min_{\beta} \sum_{i=1}^{n} r_i^2 \tag{2}$$

The method of LS has long dominated the literature (see Plackett (1972) and Stigler (1981) for some historical discussion) and application of regression techniques. However, LS is not a robust procedure in that the estimator is heavily influenced by outliers. Because data sets often contain outliers, for example gross errors, which in

general will ruin the LS fit, it is important for the data analyst to be able to identify such influential observations before proceeding with the LS method. There are several proposed alternatives to the LS method of parameter estimation which are less sensitive to outlying observations. Further discussion on alternatives to the LS method of parameter estimation in the presence of outlying observation(s) can be found in Krasker and Welsch (1982).

In this paper, we will briefly discuss the relationship between a robust parameter estimation procedure, namely the Least Median of Squares (LMS), and that of the minimax solution (section 3). In order to understand the above mentioned relationship, we will provide an overview of the minimax residual fit in the linear model in section 2. Extension of the use of minimax fit to a more general exponential family of distribution will be discussed in section 4.

2. An overview of the minimax residual fit in the normal linear model

Consider a system of linear equations (1)

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \quad i = 1, \cdots, n$$

The classical linear minimax residual problem is to find the solution of β_j $(j = 1, \cdots, p)$ which minimises

$$\Delta(\beta) = \max_i \left| r_i(\beta) \right|, \quad i = 1, \cdots, n\,(n > \mathrm{p}) \tag{3}$$

where $r_i(\beta)$ is the i-th residual as defined in (2) and $\boldsymbol{\beta}$ is a vector of p (unknown) parameters.

Cheney (1966) showed that if

(i) $n = p + 1$ and the system of $(p+1)$ linear equations has rank p, then the minimax solution is the solution of the system when all residuals are equal in absolute value,

$$\left| r_1(\beta) \right| = \left| r_2(\beta) \right| = \cdots = \left| r_{p+1}(\beta) \right|$$

(ii) $n > p+1$, and every subset of size $p+1$ has rank p, then every minimax solution of (3) is a minimax solution of an appropriate subsystem comprising $(p+1)$ equations. That is, if z is a minimum point of the function,

$$\Delta(\beta) = \max_i \left| r_i(\beta) \right|, \quad i = 1, \cdots, n \left(n > p \right),$$

then z is the minimum point of $\max_{i \in J} \left| r_i(\beta) \right|$ when J is a subset of size $p+1$ chosen from $\left\{ 1, 2, \cdots, n \right\}$.

We now summarise the necessary condition for a minimax solution.

We shall refer to the subset, J of size $p+1$ chosen at random from the set $\left\{ 1, 2, \cdots, n \right\}$ as a reference set and denote the residual for this set as

$$r_i = y_i - x_i^T \beta, \quad i \in J, \tag{4}$$

where $x_i^T = (x_{i1}, \cdots x_{ip})$ is the i-th row of $(p+1) \times p$ submatrix $\boldsymbol{X}^J$.

Without loss of generality, let the first $p+1$ observations be in the set J. If $\boldsymbol{X}^J$ has rank p, then there exist a vector $\lambda^J = (\lambda_1, \lambda_2, \cdots, \lambda_{p+1})$ such that

$$\sum_{k=1}^{p+1} \lambda_k x_{kj} = 0, \quad j = 1, \cdots, p, \tag{5}$$

where x_{kj} is the (k, j)-th element of $\boldsymbol{X}^J$. If the reference set J solves the minimax residual problem, then (4) can be written as

$$\alpha_i h^J = y_i - x_i^T \beta$$

where h^J is the maximum deviation from the minimax fit for observations in set J and $\alpha_i = \pm 1, \ i \in J$.

Using (4) and (5), h^J is given by

$$h^J = \frac{\sum_{i \in J} \lambda_i y_i}{\sum_{i \in J} |\lambda_i|}$$

where λ_i is the i-th element of vector λ^J.

3. The relationship between LMS and minimax fit

The LMS estimate (Stromberg, 1993) minimises the q-th smallest ordered squared residual, $r^2_{(q)}$, where

$$q = \left[\frac{n-(p+1)}{2}\right] + (p+1) \tag{6}$$

Since such an estimate minimises the q-th smallest squared residual for a given set, it must minimise the maximum squared residual for some q element subset of the data. Thus the data can be divided into two smaller sets - one being the inner set which contains q elements with the smallest absolute residuals while the outer set contains $n-q$ elements with the largest absolute residuals. Hence, the LMS solution is the minimax fit to the inner set of the data.

For $n >> p$, we note that if the minimax solutions to each element subset of size $p+1$ are distinct, then the exact LMS solution will have: $p+1$ observations with squared residuals equal to $r^2_{(q)}$, $q-p-1$ observations with squared residuals less than $r^2_{(q)}$, and $n-q$ observations with squared residuals greater than $r^2_{(q)}$. This generalizes the results of Steele and Stigler (1986) for an LMS fit with a single explanatory variable.

By computing the minimax solution to all possible subsets of size $p+1$, Stromberg's (1993) LMS algorithm considers the solution $\tilde{\boldsymbol{\beta}}$ only when q observations have squared residuals less than or equal to $r^2_{(q)}$. The LMS solution, $\hat{\boldsymbol{\beta}}_{LMS}$ is that value of $\tilde{\boldsymbol{\beta}}$ with minimum $r^2_{(q)}$.

To set the idea of the LMS estimation, let us first consider the one dimensional, location, case. Let θ be the LMS estimate of location and let

$$r_i(\theta) = |y_i - \theta|, \quad i = 1, 2, \cdots, n,$$

and

$$m^2_\theta = \text{median}_i (y_i - \theta)^2 = r^2_{(q)}(\theta)$$

where $q = [n/2] + 1$, from equation (6).

Note that $r_i^2(\theta) \geq 0$ and attains its minimum at zero when $\theta = y_i$.

Let $\hat{\theta}$ denotes the value of θ which minimises m_θ over θ. It is clear that $\hat{\theta}$ must satisfy the condition that $r_{(q-1)}(\hat{\theta}) = r_{(q)}(\hat{\theta})$, for, if $r_{(q-1)}(\hat{\theta})$ and $r_{(q)}(\theta)$ are the residuals for the *i*-th and *j*-th observations respectively, then the value of θ which minimises the larger of $|y_i - \theta|$ and $|y_j - \theta|$ is $(y_i + y_j)/2$, the midpoint of y_i and y_j. Consequently, the only values which are candidates for $\hat{\theta}$ are the midpoints of pairs of observations.

It is now easy to see that, if $\hat{\theta} = (y_i + y_j)/2$, then exactly $q-2$ other observations must lie between y_i and y_j and the range $|y_i - y_j|$ is smallest amongst all pairs (y_i, y_j) with this property. Thus $\hat{\theta}$ is the midpoint of the range of the shortest half sample.

Geometrically, (assuming that the points are in general position) the LMS estimate $\hat{\boldsymbol{\beta}}_{LMS}$ for the simple regression with intercept, corresponds to the middle line of the narrowest strip containing half of the observations. In higher dimensions $(p > 2)$, $\hat{\boldsymbol{\beta}}_{LMS}$ corresponds to the middle hyperplane of the thinnest hyperstrip which is the region between two parallel hyperplanes that contain half of the observations. Finally, we should note that the LMS algorithm proposed earlier assumes that not more than *q* of the observations have zero residuals.

Linear models of the form (1) have found widespread applications across a variety of disciplines, but this formulation is restrictive because there are many situations in which the additive model (1) is not appropriate for a given set of data. An example of this may occur when the response $\boldsymbol{y}$ only take strictly positive values, thus constraining $\boldsymbol{X\beta}$ to take strictly positive values only.

We can extend the range of application from the normal distribution to a more general exponential family of distributions. Given $\boldsymbol{X}$ an explanatory variable, the response vector, $\boldsymbol{Y}$ follows a generalized linear model (GLMs) with likelihood function

$$f(y_i|x_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi) \right\}$$

where b and c are known functions and the linear function, $\boldsymbol{X\beta}$, is related to θ_i via the relationship,

$$E(y_i) = \mu_i = b'(\theta_i),$$

$$g(\mu_i) = x_i^T \beta = \eta_i,$$

and ϕ is a scale parameter while *g* is a link function.

The usual method of parameter estimation in GLMs is the maximum likelihood estimation (MLE), in which the parameter $\hat{\boldsymbol{\beta}}_{MLE}$ minimises over $\boldsymbol{\beta}$ the sum of log likelihood function,

$$\max_{\beta} \sum_{i=1}^{n} ln\ f\left(y_i \middle| x_i^T\right) \equiv \max_{\beta} \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi) \right\}.$$

As in the case of the linear model, the maximum likelihood estimation is not robust in that it is subject to influence by outliers. We can derive a robust approach by reformulating the criterion of LMS, which now uses the deviance residual, d_i, given by

$$d_i(\theta_i) = 2\left\{ l(y_i, \theta_i) - l(y_i, \hat{\theta}_i) \right\}, \tag{7}$$

where

$$l(y_i;\theta) = ln\ f(y_i | x_i)$$

and $\hat{\theta}_i$ is the maximum likelihood estimation based on the *i*-th observation alone. The robust method of estimation, which uses the deviance residual (7), will be discussed in Section 4. We will look at the special case of the exact solution for the single parameter case and its algorithm in section 4.1.

4. On the theory of minimax estimation for GLMs

The minimax estimation for GLMs problem is to find $\beta_j, j = 1, \cdots, p$ to minimise the maximum of $d_{(q)}(\theta)$ where $d_i(\cdot)$ denotes the *i*-th deviance residual of observation y_i from the fitted model.

Suppose $n = p+1$. The following theorem shows that the minimax solution is the exact solution when $d_i(\theta)$, $i = 1, \cdots, p+1$ are all equal.

Theorem. Suppose that g and b' are both strictly monotone function. Further suppose that $\boldsymbol{\beta} \in \Re^p$ and that rank $(X) = p$.
Define for $i = 1, \cdots, p+1$,

$$d_i(\theta_i) = 2\left\{ b(\theta_i) - y_i\theta_i - b(\tilde{\theta}_i) + y_i\tilde{\theta}_i \right\}$$

where

$$\theta_i = \theta_i(\beta) = (b')^{-1}\left(g^{-1}(x_i\beta)\right)$$

and $\tilde{\theta}$ is the MLE based on the *i*-th observation alone. Then

(i) there exists $\boldsymbol{\beta} \in \Re^p$ such that $d_1(\theta_1), d_2(\theta_2), \cdots, d_{p+1}(\theta_{p+1})$ are all equal,

(ii) the value of $\boldsymbol{\beta} \in \Re^p$ which minimises $\max_{1 \le j \le p+1} d_j(\theta_j)$ is such that

$$d_1(\theta_1) = d_2(\theta_2) = \cdots = d_{p+1}(\theta_{p+1}).$$

The proof of this theorem will be published elsewhere and interested readers may refer to the Appendix.

Since the above estimate minimises the *q*-th smallest deviance residual for a given set, it must minimise the maximum deviance residual for some *q*-subset of the data. Hence we call this a Least Median of Deviance (LMD) estimate.

4.1. An exact algorithm for a single parameter case

In this special case of a single parameter, the process discussed in section 3 can be further simplified, thus resulting in an exact algorithm to handle this problem. To find the LMD estimate when $p = 1$, we find θ which minimises the

$$\begin{aligned} \Delta(\theta) &= \text{median } d_i(\theta), \quad i = 1, \cdots, n \\ &= d_{(q)}(\theta) \end{aligned}$$

where $q = 1 + \left[\frac{n}{2}\right]$.

Suppose that g and b' are strictly monotone functions. From the definition of $d_i(\theta)$ in the above theorem, we note the following facts:

(1) $d'(\theta) = b'(\theta) - y_i = E_\theta(Y_i) - y_i$

(2) $d''(\theta) = b''(\theta) = \text{var}_\theta(Y_i) \ge 0$

(3) the deviance, $d_i(\theta) \ge 0$ and attains its minimum at 0 when $\theta = \tilde{\theta} = (b')^{-1}(y_i)$

(4) θ_i is an increasing function of y_i because $b'(\theta)$ is monotone.

Without loss of generality, assume that

$$y_1 \le y_2 \le \cdots \le y_n.$$

Then we have,

$$\theta_1 \le \theta_2 \le \cdots \le \theta_n.$$

For any pair (i, j), $i < j$ where $j = i + q - 1$, $d_i(\theta)$ and $d_j(\theta)$ intersect at the point

$$\theta_{ij} = \frac{\{f(y_j) - f(y_i)\}}{(y_j - y_i)}$$

where

$$f(y_i) = y_i\theta_i - b(\theta_i) \quad \text{and} \quad \theta_i \le \theta_{ij} \le \theta_j .$$

The LMD estimate, $\hat{\theta}_{LMD}$, is given by $\hat{\theta}_{LMD} = \theta_{h,h+q-1}$ which corresponds to

$$\Delta(\theta_{h,h+q-1}) = \min_{1 \le i \le n-q+1} \{\Delta(\theta_{i,i+q-1})\} .$$

As an alternative, θ_{ij} may be calculated from

$$\theta_{ij} = \theta + \frac{\{d_j(\theta) - d_i(\theta)\}}{2(y_j - y_i)}$$

where θ is the MLE estimate obtained by minimizing $\sum_{i=1}^{n} d_i(\theta)$.

The computation of $\hat{\theta}_{LMD}$ can be described as follows.

(1) Order y_i, $i = 1, \cdots, n$ such that

$$y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$$

and let $d_i(\theta)$ denote the deviance for observation y_i from θ.

(2) For each $i = 1, \cdots, n - q + 1$ where $q = 1 + \left[\frac{n}{2}\right]$, let θ_i^* be the solution of

$$d_i(\theta) = d_{(i+q-1)}(\theta) .$$

(3) Calculate $d_i(\theta_i^*)$ for each $i = 1, \cdots, n - q + 1$.

The LMD estimate $\hat{\theta}_{LMD}$ is that value of θ_i^* which corresponds to

$$\Delta(\theta^*) = \min_{1 \le i \le n-q+1} \{\Delta(\theta_i^*)\} .$$

This algorithm was initially proposed by Seheult (1986).

5. Numerical example: Multiple linear regression

The data set presented by Brownlee (1965) describes the operation of a plant for oxidation of ammonia to nitric acid and consist of 21 observations as listed in Table 1. The stackloss (Y) can be explained by the rate of operation (X_1), the cooling water inlet temperature (X_2) and the acid concentration (X_3).

The results found in most cited literature (see Daniel and Wood (1971), Andrews and Pregibon (1978), Atkinson (1982), Carroll and Ruppert (1985) and Rousseeuw (1986)) mostly concluded that observations 1, 3, 4 and 21 were outliers. According to some of these results, observation 2 is reported as an outlier too (see Rousseeuw (1986)). In this study, we shall refer to an observation with absolute standardised residual greater than 2.5 as an outlier.

Table 1. Stackloss Data

Index	**Rate**	**Temperature**	**Acid Concentration**	**Stackloss**
(i)	X_1	X_2	X_3	(y)
1.	80	27	89	42
2.	80	27	88	37
3.	75	25	90	37
4.	62	24	87	28
5.	62	22	87	18
6.	62	23	87	18
7.	62	24	93	19
8.	62	24	93	20
9.	58	23	87	15
10.	58	18	80	14
11.	58	18	89	14
12.	58	17	88	13
13.	58	18	82	11
14.	58	19	93	12
15.	50	18	89	8
16.	50	18	86	7
17.	50	19	72	8
18.	50	19	79	8
19.	50	20	80	9
20.	56	20	82	15
21.	70	20	91	15

Table 2. Summary of fitting procedures Stackloss Data

Variable	LS	exact LMS
Intercept	–38.1242898	–37.2197214
X_1	0.7757662	0.7350743
X_2	1.1050111	0.4108615
X_3	–0.1707237	0.0107184

Predicted values ($\hat{y}$)

Index (i)	*y*	LS	Exact LMS
1	42	40.9347	33.8653
2	37	37.5596	33.6755
3	37	35.6582	29.3454
4	28	26.0845	19.5464
5	18	18.5574	18.5702
6	18	18.5754	18.4259
7	19	19.6983	19.6763
8	20	20.3567	19.3237
9	15	15.8217	15.6763
10	14	13.5219	13.3237
11	14	13.3867	13.7285
12	13	12.4566	13.6361
13	11	11.6674	13.9139
14	12	12.0526	14.2380
15	8	7.2266	7.7870
16	7	6.6225	7.6763
17	8	8.5059	8.3583
18	8	8.1549	8.2679
19	9	9.0467	8.6062
20	15	14.7238	13.4771
21	15	17.3881	23.4420

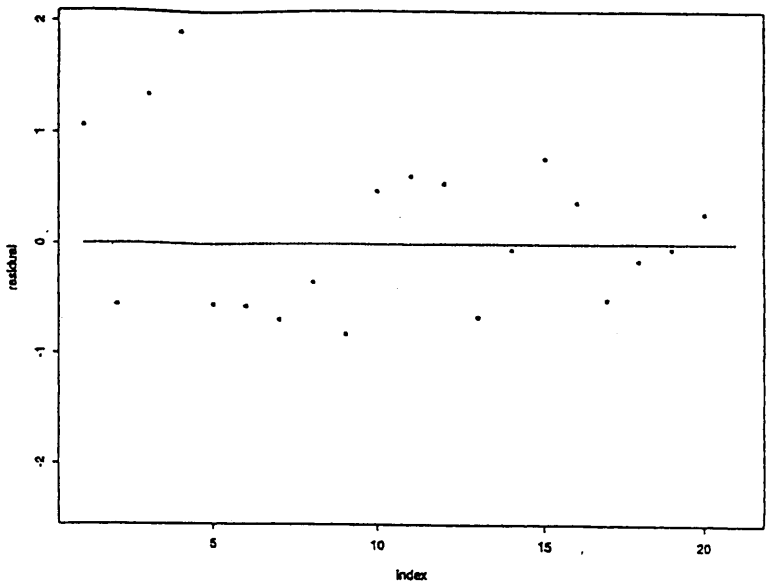

Figure 1(a). Stackloss data: Residual plot from LS fit

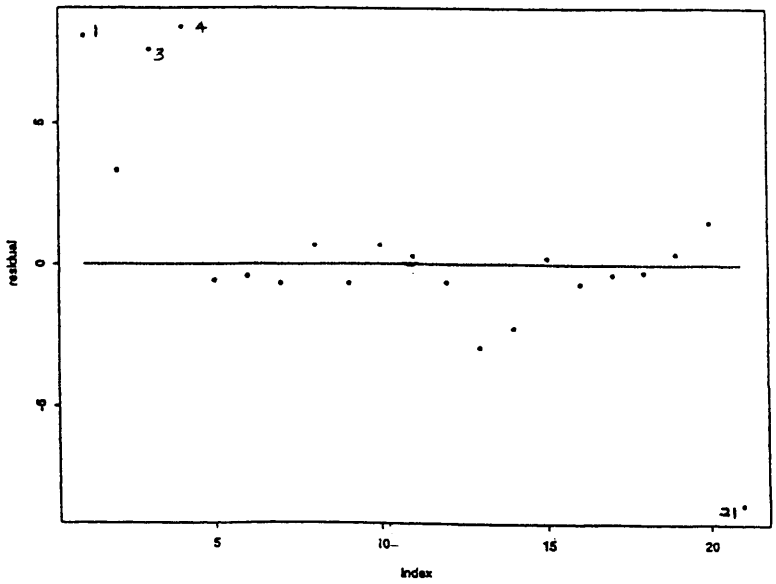

Figure 1(b). Stackloss data: Residual plot from LMS fit.

Table 2 displays the estimates of the parameters and the predicted responses $\hat{y}$ obtained via the LS and LMS procedure. Fitting the data using the LS criterion (the equivalent of MLE in the normal distribution case) yields the equation

$$\hat{y} = -38.124 + 0.776X_1 + 1.105X_2 - 0.171X_3$$

Figure 1(a) displays the residual plot from the LS fit. The standardised residuals were calculated by dividing the raw residuals by the scale estimate of the corresponding fit. Since all standardised residuals lie within ± 2.5 standard error, we conclude that none of the observations could be classified as deviant.

The exact LMS fit to the stackloss data is given by

$$\hat{y} = -37.2197 + 0.7351X_1 + 0.4109X_2 + 0.0107X_3$$

Figure 1(b) displays the residual plot associated with the exact LMS fit. The standardised residuals were obtained by dividing the raw residuals by a robust scale estimate. This plot does reveal the presence of anomalous points mentioned by the previous authors. From this residual plot, it becomes apparent that observations 1, 3, 4 and 21 are the most outlying while observation 2 is intermediate because it is on the edge of the region containing the outliers.

6. Concluding remarks

In the normal linear regression, the LS method, as theory predicted, is clearly the best method of parameter estimation when no contamination (outliers) in the data is suspected. However, in the presence of some contamination, the LS is subjected to influence by this contamination. The minimax estimation, in particular LMS, offer alternative methods of estimation which aim to provide methods of fitting reliable models in the presence of some anomalous data points. Similar phenomenon can be observed with the MLE method of estimation in GLMs.

The difficulty of LMS and hence minimax LMD is that, it is computationally expensive. The added complexity is due to the number of iterations required for convergence for each subset of observation of size $p+1$. For this, the adaptation of the feasible subset algorithm (Hawkins,1993) certainly offers great help when exhaustive enumeration is intractable. With further development on the aspects of computational programming, this difficulty may be overcome in the near future.

In this paper, we are assuming that the model is correctly specified. However, it is possible that the model is incorrectly specified in which the data appears to be contaminated. This should be borne in mind and more work will be needed to investigate this possibility.

References

1. D.F. Andrews and D. Pregibon, Finding outliers that matter, *Journal of the Royal Statistical Society B* **40** (1978), 85-93.
2. A.C. Atkinson, Regression diagnostics, transformations and constructed variables, *J. Royal Stat. Soc. B* **44** (1982), 1-36.
3. K.A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, John Wiley and Sons, New York, second edition, 1965.
4. R.J. Carroll and D. Ruppert, Transformation in regression: A robust analysis, *Technometrics* **27** (1985), 1-11.
5. E.W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill Inc., 1966.
6. C. Daniel and F.S. Wood, *Fitting Equations to Data*, John Wiley and Sons, New York, 1971.
7. D.M. Hawkins, The feasible set algorithm for the least median of squares regression, *Comput. Stat. and Data Anal*. **16** (1993), 81-101.
8. W.S. Krasker and R.E. Welsch, Efficient bounded-influence regression estimation, *J. Amer. Stat. Assoc.* **77** (1982), 595-604.
9. R.J. Plackett, Studies in the history of probability and statistics XX1X: The discovery of the method of least squares, *Biometrika* **81** (1972), 977-990.
10. P.J. Rousseeuw and A.M. Leroy, *Robust regression and outlier detection*, New York, John Wiley, 1986.
11. A.H. Seheult and P.J. Green, An exact LMD algorithm for a single parameter case, personal communication (notes), 1986.
12. J.M. Steele, and W.A. Stigler, Algoritms and complexity for least median of squares regression, *Discrete Applied Math*. **14** (1986), 93-100.
13. S.M. Stigler, Gauss and the invention of Least Squares, *The Annals of Statist*. **9** (1981), 465-474.
14. A.J. Stromberg, Computing the exact least median of squares estimate and stability diagnosis in multiple linear regression, *SIAM J. Sci. Comput*. **14** (1993), 1289-1299.

Appendix

Proof of Theorem 1

(i) The function $d_i(\theta_i) \geq 0$ attains its minimum $d_i(\theta_i) = 0$ at $\theta_i = \tilde{\theta}_i$.
Now for any $\boldsymbol{\beta} \in \Re^p$, let

$$\mu_i = b'(\theta_i) = g^{-1}(x_i^T \beta), \qquad i = 1, \cdots, p+1$$

For fixed $\theta_1, \cdots, \theta_p$ suppose $\boldsymbol{\beta}$ be the solution of

$$\theta_i = (b')^{-1}\left(g^{-1}(x_i^T \beta)\right) \quad \text{and} \quad c_i = g\left(b'(\theta_i)\right).$$

Let $\boldsymbol{X}_p$ be the $p \times p$ matrix whose i-th row is x_i^T, $i = 1, \cdots, p$.
Then,

$$\boldsymbol{X}_p \boldsymbol{\beta} = c \tag{8}$$

and since rank $(X_p) = p$, equation (8) has a unique solution for $\boldsymbol{\beta}$.

Now choose $k > 0$ and put $d_i(\theta_i^k) = k, \quad i = 1, \cdots, p.$
Since for each i, $d_i(\theta_i) \to \infty$ as $\theta_i \to \pm\infty$, such a θ_i^k always exists.
Put $c_i^k = g(b'(\theta_i^k))$, $i = 1, \cdots, p$ and define

$$x_i^T \beta_k = c_i^k, \quad i = 1, \cdots, p.$$

Then $\boldsymbol{\beta}_k$ is the unique solution of $X_p \beta = c^k$.
Now let

$$c_{p+1}^k = x_{p+1}^T \beta_k \quad \text{and} \quad \theta_{p+1}^k = (b')^{-1}\left(g^{-1}(c_{p+1}^k)\right)$$

Then either

$$d_{p+1}(\theta_{p+1}^k) < k \quad \text{or} \quad d_{p+1}(\theta_{p+1}^k) > k \quad \text{or} \quad d_{p+1}(\theta_{p+1}^k) = k.$$

We consider each of these cases separately.

Case I. Suppose that $d_{p+1}(\theta_{p+1}^k) < k$.

Then reduce k until $d_{p+1}(\theta_{p+1}^k) = k$. To see that this is possible, note that $d_{p+1}(\theta_{p+1}^k)$ is continuous as a function of k and that choosing $k = 0$ gives

$$c_i = g(y_i), \quad i = 1, \cdots, p \quad \text{and} \quad d_i = 0, \quad i = 1, \cdots, p.$$

However,

$$d_{p+1}(\theta_{p+1}^0) = d_{p+1}\left((b')^{-1}(g^{-1}(x_{p+1}^T\beta_0))\right) > 0$$

since $d_{p+1}(\cdot)$ has a unique minimum at $\tilde{\theta}_{p+1} = (b')^{-1}(y_{p+1})$ and the minimum is 0.

Case II. Suppose $d_{p+1}(\theta_{p+1}^k) > k$.

This is a special case of the general situation where $d_1, \cdots, d_{p+1}$ are not all equal.
Since b' and g are strictly monotone, $d_{p+1}(\cdot)$ does not attain its minimum unless $x_{p+1}^T\beta_k = g(y_{p+1})$.
Since $d_{p+1}(\theta_{p+1}^k) > 0$, this is not the case for $\boldsymbol{\beta}_k$.
$\boldsymbol{\beta}$ defines a hyperplane in $p+1$ dimensional space. The hyperplane is horizontal with probability 0. Hence there is a direction in which $\boldsymbol{\beta}$ can move which will be in the direction of reducing d_{p+1}. Choose the steepest such direction and move $\boldsymbol{\beta}$ until $d_{p+1} = d_j$, for some $j \le p$.

We can now change the value of $\boldsymbol{\beta}$ in such a way that $d_{p+1}(\theta_{p+1}(\beta))$ is reduced; this is certainly possible, since for any $k^* > 0$, the equation

$$d_{p+1}\left((b')^{-1}(g^{-1}(x_{p+1}^T\beta))\right) = k^* \tag{9}$$

has a $p-1$ dimensional solution set.

Choosing progressively smaller k^* and arbitrary corresponding $\boldsymbol{\beta}^*$ values in the solution set of equation (9) will eventually lead to a position where

$$\max_{i=1,\cdots,p} d_i\left((b')^{-1}(g^{-1}(x_i^T\beta^*))\right)=k^*$$

Suppose, without loss of generality, that

$$\max_{i=1,\cdots,p} d_i\left((b')^{-1}(g^{-1}(x_i^T\beta^*))\right)= d_1\left((b')^{-1}(g^{-1}(x_1^T\beta^*))\right)$$

Now, impose the restriction that $d_1 = d_{p+1}$. This imposes a linear constraint on $\boldsymbol{\beta}^*$ and, subject to this constraint, the solution set of equation (9) is now $p-2$ dimensional.

Now continue to choose progressively smaller values of k^*, and values of $\boldsymbol{\beta}^*$, subject to the constraint that $d_1 = d_{p+1} = k^*$, until

$$\max_{i=2,\cdots,p} d_i\left((b')^{-1}(g^{-1}(x_i^T\beta^*))\right)= k^*.$$

Continue this process until $d_i = d_{p+1}$ for all but one of the d_i, $i=1,\cdots,p$. This case then reduces to case I.

Case III. Suppose $d_{p+1}(\theta_{p+1}^k)=k$.

Since $|d_1(\theta_1)| = \cdots = |d_{p+1}(\theta_{p+1})|$, $\boldsymbol{\beta}$ is a minimax solution and this procedure stops.

(ii) Suppose that the value of $\boldsymbol{\beta}$ which minimises $\max_{j=1,\cdots,p+1} d_j(\theta_j)$ is such that

$$d_{p+1}(\theta_{p+1}) > d_j(\theta_j),\quad j=1,\cdots,p.$$

Then, applying the procedure for Case (II) will always reduce the maximum deviance.