

2-1-1993

ROBUST REGRESSION AND OUTLIER SET ESTIMATION USING LIKELIHOOD REASONING

R. L. Kashyap'

Purdue University School of Electrical Engineering

S. Maiyuran

Purdue University School of Electrical Engineering

Follow this and additional works at: <http://docs.lib.purdue.edu/ecetr>

Kashyap', R. L. and Maiyuran, S., "ROBUST REGRESSION AND OUTLIER SET ESTIMATION USING LIKELIHOOD REASONING" (1993). *ECE Technical Reports*. Paper 33.

<http://docs.lib.purdue.edu/ecetr/33>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

ROBUST REGRESSION AND OUTLIER
SET ESTIMATION USING
LIKELIHOOD REASONING

R. L. KASHYAP
S. MAIYURAN

TR-EE 93-8
FEBRUARY 1993



SCHOOL OF ELECTRICAL ENGINEERING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1285

**ROBUST REGRESSION AND OUTLIER SET
ESTIMATION USING LIKELIHOOD REASONING***

R.L. Kashyap[†] and S. Maiyuran

Abstract

We consider the **simultaneous** estimation of the outlier set and the regression parameters using the **contaminated** data set, A , which has many members obeying the linear model, but some which **do** not. A **precise** definition of outliers is given using set theory. Fixing the **number of outliers** to be L , the optimum set of **inliers** is chosen as the set having the highest **log likelihood among** all subsets of A of size $N-L$, N being the size of A . We define the concept of **a valid partition of** the data set A into **inliers** and outliers, and show that the local maxima of the likelihood **function** yields a regression estimate which yields a valid **partition** of the data A . We show that **the** global maximum set S_L^* , the estimate of the inlier **set**, has an interesting game theoretic **interpretation**. The outlier set estimate given here is based on evaluating different partitions of data and it does not involve arbitrary thresholds characteristic of the papers in the literature.

We develop a new formula for computing the sum of minimal residual squares of any **subset** of A of size $N-L$ as a quadratic form in the **residuals** quoted from the LS coefficients obtained **from** all the **data** A . Only a particular case of this formula when $L = 1$ has been **known**. Using this method one can compute the S_L^* , the optimal set of inliers of size $N-L$. We **apply the theory** developed here to seven well known "difficult" multivariate data sets like the

[†]School of **Electrical Engineering**, Purdue University, 1285 **Electrical Engineering Building**, West Lafayette, IN 47907-1285; e-mail: lca&yap@ecn.purdue.edu

⁺⁺This work was **partially supported** by the Office of Innovative Science and Technology (IST) of the SDIO and **monitored** by the Office of Naval Research under contract No. N00014-91J-4126; and by the **Purdue Engineering Research Center for Intelligent Manufacturing Systems** funded by the **National Science Foundation** under contract No. CDR-88-03017.

stackloss data, the simulated data set of Rousseeuw, the water salinity data, the engine knock data set, the **Hawkins-Bradu-Kass** data, the star data set, and show our method extracts the correct outliers from the simulated data sets and extracts the outliers from data sets like engine **knock** where conventional methods like the least median squares fail to do.

I. Introduction

We **consider** the multivariate regression problem with linear model, $y_i = a^T x_i + \beta + e_i$, with $\{e_i\}$ being a zero mean **i.i.d.** sequence, $x_i \in \mathbb{R}^P$ and the associated contaminated data set $A = \{(x_i, y_i), i = 1, \dots, N\}$. All the members of A excepting some, the so-called outliers, obey the linear model and the identity of outliers is not known. **Our** aim is to simultaneously estimate the outlier subset in A and the unknown **parameters** a and β in the model. We will not attempt here to give a review of current robust estimation methods since excellent reviews [Rousseeuw and Leroy, 1987] are available.

The outliers in A could have been caused by the **errors** in the sensors. Also, since the linear model is chosen often on an **ad hoc** basis, there is no reason for all the members of the data set to obey the linear model. Identification of these outliers is a first step in getting a better nonlinear **regression** model valid for all the data. It is being **increasingly** recognized that data sets encountered in many engineering and science applications are contaminated with outliers which **rarely** get noticed except through serendipity as in the recent study [Hettmansperger and Sheather, 1992].

The deleterious effects of the outliers **are** well known. The **error** of the forecasts given by the estimated **model** obtained from standard methods like least **squares** is increased by the presence of outliers. **The** use of outlier resistant estimation methods partially mitigates this problem. Secondly, the presence of strong outliers in the data seems to greatly increase the sensitivity of the estimated model coefficients to relatively small perturbations of the data caused by a data transcription error. The use of a so-called outlier resistant estimation method such as

least median squares (LMS) does not necessarily decrease this sensitivity as discussed in [Hettmansperger and Sheather, 1992]. The only way to handle this sensitivity problem is to detect the outliers and delete them from the data set.

To estimate the set of outliers, we need a precise way of distinguishing the set of inliers, which is given below.

Definition 1 (outlier set): Let S be a data set which is made up of only **i.i.d.** members with a member (\mathbf{x}, y) obeying the density $p(\mathbf{y} | \mathbf{x}; \phi)$ where ϕ is an unknown vector parameter, $\mathbf{x} \in \mathbb{R}^p$, and y a scalar. Let $\hat{\phi}_S$ be the ML estimate of ϕ computed solely from S . Consider a set O whose typical member is (x', y') where $x' \in \mathbb{R}^p$ and y' is scalar. A set O is said to be an outlier set with respect to S if the following inequality is satisfied:

$$\left\{ \min_{(\mathbf{x}, y) \in S} \ln p(\mathbf{y} | \mathbf{x}; \hat{\phi}_S) - \max_{(\mathbf{x}', y') \in O} \ln p(\mathbf{y}' | \mathbf{x}'; \hat{\phi}_S) \right\} \gg 0. \quad (1.1)$$

Consider the usual multivariate Gaussian regression **problem** with $\mathbf{x}^T = (x_1, \dots, x_p, 1)$ and $\phi = (\boldsymbol{\alpha}, \sigma)$, $\mathbf{a}^T = (\alpha_1, \dots, \alpha_{p+1})$. \mathbf{x} and \mathbf{a} as defined here will be used throughout the paper.

$$p(\mathbf{y} | \mathbf{x}; \phi) = \text{Gauss}(y - \boldsymbol{\alpha}^T \mathbf{x}, \sigma^2), \quad (1.2)$$

Then the equation (1.1) simplifies to:

$$\left\{ \min_{(\mathbf{x}', y') \in O} (y' - \hat{\boldsymbol{\alpha}}_S^T \mathbf{x}')^2 - \max_{(\mathbf{x}, y) \in S} (y - \hat{\boldsymbol{\alpha}}_S^T \mathbf{x})^2 \right\} \gg 0. \quad (1.3)$$

We want to draw attention to two aspects of the above definition. Firstly, nothing has been said about the density of \mathbf{x} **variables** of the inliers, which could be **unimodal** or a mixture. Secondly, the **estimate** $\hat{\boldsymbol{\alpha}}_S$ is computed using **only** the inliers. Suppose we had an estimate, say $\hat{\boldsymbol{\alpha}}_{SO}$, obtained from both S and O data, then (1.3) does **not** imply (1.4).

$$\min_{(x',y') \in O} (y' - (\hat{\alpha}_{SO})^T x')^2 - \max_{(x,y) \in S} (y - (\hat{\alpha}_{SO})^T x)^2 > 0, \quad (1.4)$$

As a matter of fact, some values of $(y' - (\hat{\alpha}_{SO})^T x')^2$ could be very small. These outliers are the so-called masked outliers, to be discussed later.

Definition2 (valid partition): A partition $\{S, \bar{S}\}$ of the set A, $\bar{S} = A - S$, is said to be valid if $\#S > \#(A-S)$ and \bar{S} is an outlier set with respect to the inlier set S according to the definition 1.

Any data set, whether contaminated or not, has numerous valid partitions. A valid partition is interesting or significant if the left hand side of eq. (1.1) or (1.3) is relatively large. Our first task is to determine several significant valid partitions of the given data set. We initially assume that the L, the number of outliers in A, is known.

Let S_L be a subset of A of size (N-L). Let $J(S_L, \alpha)$ be the sum of the squares of the residuals of the members of S_L using the parameter α , the regression parameter. It is also the negative log **likelihood** function of S_L stripped of some unessential terms. If $(\hat{S}_L, \hat{\alpha})$ is a local **minimum** of J, then it is a valid partition and vice versa.

Let (S_L^*, α_L^*) be the global minimum of J. We will show that (S_L^*, α_L^*) is a solution to an interesting two person nonzero sum game.

To compute the global minimum S_L^* , we first develop a new expression for $J_1(S_L)$ for any S_L . It is a quadratic form of the L dimensional residual vector formed from $r_k = y_k - \alpha_A^T x_k$ associated with the data points (x_k, y_k) in the candidate outlier set $\{A - S_L\}$, and α_A being the least squares estimate of α using all the data set A. By computing the expression over all possible subsets of the residual vector of size L, we determine the optimal set S_L^* . The expression for $J_1(S_L)$ is one of the significant results of the paper. Only a special case of this method with L = 1 is **available** in the literature [Cook, 1977].

The **estimate** \mathbf{S}_L^* has all the relevant **equivariance** properties and it has a breakdown point of nearly fifty percent by appropriately choosing L . In **addition**, we discuss the consistency of the set estimate \mathbf{S}_L^* , i.e., is \mathbf{S}_L^* equal to the true inlier set from which A was constructed after adding the L outliers. We show the consistency of the estimate for the case of $L = 1$. The consistency implies that the estimate $\boldsymbol{\alpha}^*$ has zero mean and minimum variance.

Next, we offer several criteria for comparing the optimal sets \mathbf{S}_1^* , \mathbf{S}_2^* , etc., obtained with different values of L . We apply the theory developed to various data sets mentioned in the literature such as the stack-loss data, water salinity data, steam data, engine knock data, the simulated data set of Rousseeuw, etc. Our methods can identify outliers where other methods fail as in the engine knock data. In all the simulated data sets, it offers the correct outlier set solution. In others, our solutions yield significantly lower measures of forecast errors.

The above statements may seem strange to some persons since the common least squares estimate **is** the common example given for a nonrobust estimate. The commonly used least squares **estimator** has two distinct facets, namely the algorithm or procedure for computing the estimate and the input data. All the papers which are critical of the least squares approach used the **entire** set of contaminated data. The interesting point of least squares method is its extreme sensitivity to the outliers. This sensitivity has been exploited in this paper to detect the outliers. This paper shows the likelihood principle used in its completeness is a very powerful technique.

II. Global ML Estimate of Inlier Set and Its Computation

Consider the contaminated data set A

$$A = \{(\mathbf{x}_i, y_i), i = 1, \dots, N\}, \mathbf{x}_i^T = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip}, 1),$$

The **inliers** among A obey the Gaussian density characterized by parameter $(\boldsymbol{\alpha}, \sigma)$, $\mathbf{a}^T = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_{p+1})$, mentioned in eq. (1.2). Let \mathbf{S}_L be a subset of A of size $N-L$, i.e., it is

obtained from A by deleting L candidate outliers:

$$\mathbf{S}_L = A - \{(x_{i_k}, y_{i_k}), k = 1, \dots, L\}. \quad (2.1)$$

The negative likelihood of the members of \mathbf{S}_L omitting inconsequential terms is $J(\mathbf{S}, \alpha)$.

$$J(\mathbf{S}_L, \alpha) = \sum_{(x_i, y_i) \in \mathbf{S}_L} (y_i - \alpha^T x_i)^2$$

$$(\mathbf{S}_L^*, \alpha_L^*) = \text{Argument min.}_{\mathbf{S}, \alpha} J(\mathbf{S}, \alpha), \quad (2.2)$$

i.e., $J(\mathbf{S}_L^*, \alpha^*) \leq J(\mathbf{S}, \alpha)$, $\forall \mathbf{S} \subset A$, $\#\mathbf{S} = N-L$, and $\alpha \in \mathbb{R}^{p+1}$.

Let $r_i = y_i - (\alpha_L^*)^T x_i$. Rank the r_i^2 in decreasing order.

$$r_{j_1}^2 \geq r_{j_2}^2 \cdots \geq r_{j_N}^2 \quad (2.3)$$

Theorem 1:

A. α_L^* is the least squares estimate of α using only the members of \mathbf{S}_L^* ,

$$\text{i.e., } \mathbf{a} = (\sum x_i x_i^T)^{-1} (\sum x_i y_i)$$

where the summation is over all members of \mathbf{S}_L^* .

B. $\mathbf{S}_L^* = A - \{(x_{j_k}, y_{j_k}), \dots, k = 1, \dots, L\}$.

C. $\mathbf{S}_L^* = \text{Arg. Min.}_{\mathbf{S} \subset A, \#\mathbf{S}=L} J_1(\mathbf{S})$,

where

$$J_1(S) = J(S, \alpha_S), \quad (2.4)$$

and α_S is the LS estimate of \mathbf{a} using members of S only.

D. The: **partition** of A into $\{S_L^*, A - S_L^*\}$ is **valid**, i.e., $(A - S_L^*)$ is an outlier set with respect to the inlier set S_L^* according to definition 1.

....

Proof:

Part A follows from the inequality $J(S_L^*, \alpha_L^*) \leq J(S_L^*, \mathbf{a}) \forall \mathbf{a} \in \mathbf{R}^{p+1}$.

Part B follows from the inequality $J(S_L^*, \alpha_L^*) \leq J(S, \alpha_L^*)$ and the definition of the indices j_1, j_2 , etc., in (2.3).

Part C.

$$\begin{aligned} \min_{S, \alpha} J(S, \alpha) &= \min_S \{ \min_a J(S, \alpha) \} \\ &= \min_S \{ J(S, \alpha_S) \} = \min_S J_1(S) \end{aligned}$$

Part D: It follows from Part (B) and (2.3),

To compute the global minimum S_L^* , we need to simplify the expression for $J_1(S)$, the square **sum** of residuals associated with S .

$$J_1(S) = \min_{\alpha} \sum_{(x_i, y_i) \in S} (y_i - \alpha^T x_i)^2$$

$$= \sum_{(x_i, y_i) \in S} y_i^2 - \alpha_S^T P_S^{-1} \alpha_S, \text{ where} \quad (2.5)$$

$$P_S = \left(\sum_{(x_i, y_i) \in S} x_i x_i^T \right)^{-1}, \quad \alpha_S = P_S \left(\sum_{(x_i, y_i) \in S} x_i y_i \right) \quad (2.6)$$

$$\text{Let: } P_A = \left(\sum_{(x_i, y_i) \in A} x_i x_i^T \right)^{-1}, \quad \alpha_A = P_A \left(\sum_{(x_i, y_i) \in A} x_i y_i \right). \quad (2.7)$$

Recall the definition of the set S_L in (2.1) with L outliers $((x_{i_k}, y_{i_k}), k = 1, \dots, L)$.

$$\text{Let: } B = [x_{i_1} \mid x_{i_2} \mid \dots \mid x_{i_L}], \quad y_O = \text{col.}(y_{i_1}, y_{i_2}, \dots, y_{i_L}), \quad (2.8)$$

$$H_{A,L} = B^T P_A B, \quad (H_{A,L})_{jk} = x_{i_j}^T P_A x_{i_k}, \quad j, k = 1, \dots, L. \quad (2.9)$$

Theorem 2: α_A and P_A are related to α_S and P_S in the following ways:

$$(i) \quad P_S = P_A + P_A B [I - H_{A,L}]^{-1} B^T P_A$$

$$(ii) \quad P_A = P_S - P_S B [I + B^T P_S B]^{-1} B^T P_S$$

$$(iii) \quad \alpha_S = \alpha_A - P_A B [I - H_{A,L}]^{-1} (y_O - B^T \alpha_A)$$

$$(iv) \quad \alpha_A = \alpha_S + P_S B [I + B^T P_S B]^{-1} (y_O - B^T \alpha_S).$$

The proof is in the Appendix. Results (ii) and (iv) are well known in the **Kalman filtering** literature. Results (i) and (iii) have not appeared in the literature before. The existence of the inverse matrix follows from the theorem below.

Theorem .3:

$$[\mathbf{I} - \mathbf{B}^T \mathbf{P}_A \mathbf{B}]^{-1} = \mathbf{I} + \mathbf{B}^T \mathbf{P}_S \mathbf{B}, \text{ where} \quad (2.10)$$

$$\mathbf{P}_S = [\mathbf{P}_A^{-1} - \mathbf{B} \mathbf{B}^T]^{-1}. \quad ***$$

The proof is in the appendix.

Theorem 4: Consider the set \mathbf{S}_L defined in (2.1), \mathbf{P}_S in (2.6), \mathbf{P}_A in (2.7), \mathbf{B} and y_O in (2.8).

$$\text{A) } J_1(\mathbf{S}_L) = J_1(\mathbf{A}) - \mathbf{r}_{AO}^T [\mathbf{I} - \mathbf{B}^T \mathbf{P}_A \mathbf{B}]^{-1} \mathbf{r}_{AO}, \text{ where} \quad (2.11)$$

$$\mathbf{r}_{AO} = y_O - \mathbf{B}^T \alpha_A$$

$$= \text{col.}(r_{A,i_k}, k = 1, \dots, L), \quad r_{A,k} = y_k - \alpha_A^T x_k$$

$$\text{and } J_1(\mathbf{A}) = \sum_{(x_i, y_i) \in \mathbf{A}} y_i^2 - \alpha_A^T \mathbf{P}_A^{-1} \alpha_A.$$

$$\text{B) } J_1(\mathbf{A}) - J_1(\mathbf{S}_L) = \mathbf{r}_{AO}^T [\mathbf{I} + \mathbf{B}^T \mathbf{P}_S \mathbf{B}] \mathbf{r}_{AO}, \quad (2.13)$$

$$\text{C) } \mathbf{S}_L^* = \mathbf{A} - \{(x_{j_k}, y_{j_k}), k = 1, \dots, L\}, \text{ where}$$

$$\{j_1, j_2, \dots, j_L\} = \text{Arg.} \max_{i_1, i_2, \dots, i_L, i_k \in [1, N] \forall k} [J_1(\mathbf{A}) - J_1(\mathbf{S}_L)]$$

The proof is in the appendix.

Comment 1: The above theorem is one of the key results of this paper. The above result has been known, in different notation, only for the case of $L = 1$, i.e., \mathbf{S} is the result of deleting only one member from \mathbf{A} [Cook, 1977]. To compute \mathbf{S}_L^* , we need only compute α_A with all the

data, then $r_{A,k}$ for all k , compute r_{AO} and form the quadratic form in (2.11) or (2.13) for all combinations of indices (i_1, i_2, \dots, i_L) .

Comment 2: The expression in (2.11) involves the inversion of a $L \times L$ matrix whereas the expression in (2.13) involves the inversion of a $(p+1) \times (p+1)$ matrix P_S . We use one or the other depending on the size of L and P .

Comment 3: Some simplifications:

When $L = 1$, and $A - S_1 = \{(x_i, y_i)\}$

$$J_1(S_1) = J_1(A) - \frac{r_{Ai}^2}{1 - x_i^T P_A x_i}$$

when $L = 2$ and $S_2 = A - \{(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2})\}$,

$$J_1(S_2) = J_1(A) - \{(r'_{i_1})^2 + (r'_{i_2})^2 + 2r'_{i_1} r'_{i_2} h'_{i_1, i_2}\} / [1 - (h'_{i_1, i_2})^2], \text{ where}$$

$$r'_{i_k} = \frac{r_{Ai_k}}{\sqrt{1 - x_{i_k}^T P_A x_{i_k}}}, \quad k = 1, 2, \quad h'_{i_1, i_2} = \frac{(x_{i_1}^T P_A x_{i_2})}{\sqrt{(1 - x_{i_1}^T P_A x_{i_1})(1 - x_{i_2}^T P_A x_{i_2})}}$$

$|r'_{i_1}|$ and $|r'_{i_2}|$ may be relatively small even when **outliers are** present so that $J_1(S_1)$ may not indicate the presence of outlier may any significant test as discussed in [Cook, 1977]. But $\{J_1(S_2) - J_1(A_1)\}$ becomes significant when outliers are present because of the denominator and the cross terms like $r'_{i_1} r'_{i_2} h'_{i_1, i_2}$ in the numerator. This effect will become more dominant as L increases.

Comment 4: 1 Even without any attempt at optimizing the program, the search time for finding S_L^* with $N = 20$, $p = 5$ and $L = 4$ and using eq. (2.11) was about five minutes on a Sun Spark

machine. With $N = 16$, $p = 4$, $L = 7$, the time is of the same order.

Comment 5: Let us express the statistic in Theorem 4 in **terms** of the empirical mean \bar{v} and covariance matrix R associated with P_S .

$$\text{Let: } \mathbf{x}_i = \text{col.}(\mathbf{v}_i^T, 1), \quad \bar{v} = \sum_{i \in S} \mathbf{v}_i / (N-L), \quad (2.14)$$

$$R = \left[\frac{1}{N-L} \sum_{i \in S} (\mathbf{v}_i - \bar{v})(\mathbf{v}_i - \bar{v})^T \right]^{-1}, \quad (2.15)$$

$$\text{Then } P_S = \left[\begin{array}{cc} R, & -R\bar{v} \\ -\bar{v}^T R, & \bar{v}^T R\bar{v} + 1 \end{array} \right] / (N-L). \quad (2.16)$$

The formula (2.13) can be simplified as follows:

$$(\mathbf{B}\mathbf{r})^T = \left(\sum r_i \mathbf{v}_i^T, \sum r_i \right)$$

$$\begin{aligned} J_1(\mathbf{A}) - J_1(\mathbf{S}) &= \mathbf{r}^T (\mathbf{I} + \mathbf{B}^T P_S \mathbf{B}) \mathbf{r} \\ &= \mathbf{r}^T \mathbf{r} + \left(\sum r_i \right)^2 / (N-L) + \left(\sum r_i (\mathbf{v}_i - \bar{v}) \right)^T R \left(\sum r_i (\mathbf{v}_i - \bar{v}) \right) / (N-L), \end{aligned} \quad (2.17)$$

where i is summed over the set $I_L = \{i_1, i_2, \dots, i_L\}$. The traditional Mahalanobis **matrix**^C involves quadratic form $(\mathbf{v}_i - \bar{v})^T R (\mathbf{v}_i - \bar{v})$. Note the summation in the third **term** of (2.17).

When there is only one independent variable ($p = 1$), the formula (2.17) can be used without a need for matrix inversion. This is stated as Theorem 5.

Theorem 5: Let $p = 1$. $\mathbf{x}_i = (\mathbf{v}_i, 1)$. **Define** the scalars \bar{v} and \bar{R} as in (2.14) and (2.15).

$$J_1(\mathbf{A}) - J_1(\mathbf{S}_L) = \sum \mathbf{r}_i^2 + (\sum \mathbf{r}_i)^2 / (N-L) + (\sum \mathbf{r}_i (\mathbf{v}_i - \bar{\mathbf{v}}))^2 / R(N-L), \quad (2.18)$$

where the \mathbf{i} is summed over the set $\{\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_L\}$.

Comment: The third term in (2.18) suggests a simple heuristic for locating the outliers. Let $\bar{\mathbf{v}}_A = (1/N) \sum_{i=1}^N \mathbf{v}_i$. Alternatively, $\bar{\mathbf{v}}_A$ could be the robust location estimate of the sequences $\{\mathbf{v}_i, i = 1, \dots, N\}$. The statistics $(\mathbf{r}_i (\mathbf{v}_i - \bar{\mathbf{v}}_A))^2$ are ranked. The indices with **relatively** large value of this statistic are the candidates for outliers. We have used this statistic successfully in solving all the one dimensional regression examples, one of them being the example 7 in the Section VI. This statistic may be generalized for multivariate examples also.

Since $\alpha_{\mathbf{L}}^*$ is a least squares estimate, we have the following result.

Theorem 6: The estimate $\alpha_{\mathbf{L}}^*$ possesses the three equivariance properties defined in [Rousseeuw and Leroy, 1987; p. 116], namely regression equivariance, equivariance to scale and **equivariance** to affine **transformations**.

Theorem 7: The estimate $\alpha_{\mathbf{L}_0}^*$ has a breakdown point equal to $[(N - \mathbf{L}_0)/N]100$ percent.

$$\mathbf{L}_0 = N/2, \text{ if } N \text{ is even}$$

$$= (N+1)/2, \text{ if } N \text{ is odd}$$

Proof: **Suppose** the data set A has $(N - \mathbf{L}_0)$ **data** points $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ one or more components of which take arbitrarily large values so that the corresponding value of $(\mathbf{y}_i - \mathbf{a}^T \mathbf{x}_i)^2$ is very large regardless of \mathbf{a} . The set $\mathbf{S}_{\mathbf{L}_0}^*$ by definition excludes these points and consequently $\alpha_{\mathbf{L}_0}^*$ is finite.

There exist numerous estimates of α for any given problem which have **nearly fifty** percent breakdown point, as indicated below, clearly showing that the possession of a high breakdown point by an estimate does not imply that the estimate possesses any other "good." properties.

Theorem 8: Let $v_i = \text{col.}(x_{i1}, x_{i2}, \dots, x_{ip})$, $x_i^T = (v_i^T, 1)$,

$$\gamma = \text{col.}(\alpha_1, \dots, \alpha_p), \alpha_{p+1} = \beta, \alpha^T = (\gamma^T, \beta)$$

Fix the value of γ arbitrarily.

$$(\mathbf{S}^*, \beta^*) = \min_{\mathbf{S}, \beta} \sum_{(x_i, y_i) \in \mathbf{S}} (y_i - \gamma^T v_i - \beta)^2$$

where \mathbf{S} is minimized over all subsets of \mathbf{A} of size $(N - L_0)$. Then $\alpha^T = (\gamma^T, \beta^*)$ has a breakdown point equal to $100(N - L_0)/N$ percent for any arbitrary finite γ vector.

The proof is similar to that of Theorem 7.

Theorem 9: The set estimate \mathbf{S}_L^* is consistent for the case $L = 1$, with the interclass separation (the LHS of (1.3)) sufficiently large.

Proof is in the appendix. **The** theorem can **be** generalized for $L > 1$ also.

Comment: Consistency implies that all the members of \mathbf{S}_L^* obey the Gaussian assumption and consequently the estimate α_L^* has zero mean and minimum variance.

Masked outliers: This idea, repeatedly mentioned in several places [see, for instance, Rousseeuw and Leroy, 1987], can be clarified using the framework developed here. Moreover, we provide here a procedure for simulating masked outliers. We need the following theorem.

Theorem 10: Recall the definition of S in (2.1), B in (6), y_O in (2.7). r_{AO} , the residuals vector in (2.12).

Let: $r_{SO} = y_O - B^T \alpha_S = \text{col.}(y_{i_k} - \alpha_S^T x_{i_k}, k = 1, \dots, L)$.

Then: $r_{AO} = (I + B^T P_S B)^{-1} r_{SO}$.

$$= (I - H_{A,L}) r_{SO}$$

The proof follows directly from part (iv) of Theorem 2.

Suppose S is outlier free. Then by our definition, members of the set $(A - S)$ are outliers. Suppose they are dominant outliers. Then each component of r_{SO} is relatively large in magnitude. Note that elements of B can be chosen independently of P_S , so that $(I + B^T P_S B)^{-1}$ can be made relatively small in magnitude rendering several or all elements of r_{AO} relatively small; *i.e.*, even though the outliers residuals r_{SO} are dominant in magnitude, their counterparts r_{AO} in the contaminated data set may not be dominant. This phenomena is called the masking of outliers. An illustration of this phenomena of masked outliers with the multivariate data set of Rousseeuw is given in Section VI.

Game theoretic interpretation of (S^*, α^*)

Consider a two person nonzero **sum** with two players, N and P , with their respective payoff functions J_N given below and $J(S, \alpha)$ given earlier.

$$J_N(S, \alpha) = \sum_{(x_i, y_i) \in A-S} (y_i - \alpha^T x_i)^2$$

S is under the control of player N who wants to *maximize* J_N . The range of S is all subsets of A of size $N-L$. The vector α , $\alpha \in \mathbb{R}^{p+1}$ is under the control of player P who wants to *minimize* J . Both the players know both the functions J_N , J , the data set A and the integer L . The two

players are assumed to be rational, i.e., they are interested in optimizing their respective functions.

Theorem 11: $(\mathbf{S}_L^*, \alpha_L^*)$ defined in Theorem 1 is a stable equilibrium point for the game, i.e.,

$$J_N(\mathbf{S}_L^*, \alpha_L^*) \geq J_N(\mathbf{S}, \alpha_L^*), \forall \mathbf{S} \subset A, \#\mathbf{S} = N-L, \quad (2.19)$$

$$J(\mathbf{S}_L^*, \alpha_L^*) \leq J(\mathbf{S}_L^*, \alpha), \forall \alpha \in \mathbb{R}^{p+1}. \quad (2.20)$$

Proof: (2.20) follows from the least squares property of α_L^* . (2.19) follows from the characterization of set \mathbf{S}_L^* in part (B) of Theorem 1.

Comment: When the two players N and P announce their strategies \mathbf{S}_L^* and α_L^* , each player cannot do any better by deviating from its strategy, if the other player takes the declared action. If N chooses $\mathbf{S} = \mathbf{S}_L^*$, there is nothing to gain for P by deviating from P. Similarly for N.

$(\mathbf{S}_L^*, \alpha_L^*)$ would still be the solution to the game if J_N were defined as the sum of the absolute deviations $|y_i - \alpha^T x_i|$ or any of its other monotonic transformations.

III. Estimate with Univariate Gaussian Data

Let $A = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be made up of $(N-L)$ **inliers** obeying Gauss density $\mathbf{N}(\boldsymbol{\theta}, \boldsymbol{\rho})$ and L outliers. Let S be a subset of A , $\#\mathbf{S} = N-L$. The negative log **likelihood** expression for members of S , after omitting some nonessential terms is:

$$J(\mathbf{S}, \boldsymbol{\theta}) = \sum_{\mathbf{x}_k \in \mathbf{S}} (\mathbf{x}_k - \boldsymbol{\theta})^2$$

The value, of $\boldsymbol{\theta}$ which minimizes $J(\mathbf{S}, \boldsymbol{\theta})$ is $(1/(N-L)) \sum_{\mathbf{x}_k \in \mathbf{S}} \mathbf{x}_k$.

$$\begin{aligned} J_1(S) &= \min_{\theta} J(S, \theta) \\ &= \sum_{x_k \in S} x_k^2 - \frac{1}{N-L} \left[\sum_{x_k \in S} x_k \right]^2 \end{aligned}$$

We need to minimize $J_1(S)$ with respect to S over all subsets of A . The next lemma indicates that we **need** to compare only contiguous subsets of A of length $(N-L)$, which are only $(N-L+1)$ in number.

Definition: A subset S_1 of the data set $A = \{x_1, \dots, x_N\}$ is said to be contiguous if for every pair of members x_i and x_j of S_1 where $x_i < x_j$, the existence of a member $x_k \in A$ and $x_i < x_k < x_j$ implies ($x_k \in S_1$).

Theorem 12: Given a subset S_2 of size $(N-L)$, $S_2 \subseteq A$ which is not contiguous, there exists a contiguous subset S_c of size $(N-L)$, $S_c \subseteq A$ so that:

$$J_1(S_c) \leq J_1(S_2).$$

The proof is in the appendix. The above theorem and definition of $J_1(S)$ yield the following Theorem.

Theorem 13: The global minimum S_L^* of $J_1(S)$ has the following structure: Rank the data $\{x_i\}$

$$x_{i_1} < x_{i_2} < \dots < x_{i_N} ; y_k = x_{i_k}, \quad (3.1)$$

Then $S_L = \{y_k, M_L \leq k \leq M_L + N-L - 1\}$, where

$$M_L = \text{Arg. min.}_{1 \leq M \leq L+1} (\sum y_k^2 - (\sum y_k)^2 / (N-L)), \quad (3.2)$$

and the index k runs from M_L to $M_L + N - L - 1$.

$$\theta_L = \frac{1}{N-L} \sum_{k=M_L}^{M_L+N-L-1} y_k .$$

Comment 1: S_L is usually unique. Non-uniqueness occurs while dealing with integer data, some of whose values may be repeated.

Comment 2: The partition $\{S_L, A - S_L\}$ is valid, i.e., the residual square of *every* outlier is greater than that of all the inliers.

$$\min_{x_k \in A - S_L} (x_k - \theta_L)^2 > \max_{x_i \in S_L} (x_i - \theta_L)^2 .$$

For computing S_L we need to compute the statistic in (3.2) for all $M = 1, 2, \dots, L+1$ and pick the minimum. We can avoid the brute force search by developing a gradient like search procedure. This will be treated elsewhere.

IV. Local Minima and Valid Partitions

A global minimum (S^*, α^*) of $J(S, \alpha)$ is difficult to compute if N is large. It is easier to obtain a local minimum which has interesting properties.

Definition 3 (local minimum): $(\hat{S}, \hat{\alpha})$ is a local minimum of $J(S, \alpha)$ if:

$$J(\hat{S}_L, \hat{\alpha}) \leq J(S, \alpha) \forall \hat{S}, S_L \subset A, \#S = L, \#\hat{S}_L = L, \quad (4.1)$$

$$J(\hat{S}_L, \hat{\alpha}_L) \leq J(\hat{S}_L, \alpha), \forall \alpha \in R^{p+1} . \quad (4.2)$$

A problem may have several local minima. A global minimum is a local minimum, but not vice-versa.

Theorem 14: If $(\hat{\mathbf{S}}, \hat{\boldsymbol{\alpha}})$ is a local minima, then the

- 1) partition $[\hat{\mathbf{S}}, \mathbf{A} - \hat{\mathbf{S}}]$ is valid.
- 2) $\hat{\boldsymbol{\alpha}}$ is the LS estimate of \mathbf{a} w.r.t. $\hat{\mathbf{S}}$.

Proof: $\mathbf{r}_i = \mathbf{y}_i - (\hat{\boldsymbol{\alpha}})^T \mathbf{x}_i$. Rank \mathbf{r}_i^2 in increasing order.

$$\mathbf{r}_{i_1}^2 \geq \mathbf{r}_{i_2}^2 \geq \dots \geq \mathbf{r}_{i_N}^2, \quad (4.3)$$

Then the **inliers** set $\hat{\mathbf{S}}$ has the L points with the smallest residuals squares. The outliers will have the remaining points with larger values of residual squares. Hence the partition is valid.

Theorem 15: Suppose the partition $\{\hat{\mathbf{S}}, \mathbf{A} - \hat{\mathbf{S}}\}$ is valid and $\hat{\boldsymbol{\alpha}}$ be a LS estimate of \mathbf{a} using $\hat{\mathbf{S}}$. Then $\{\hat{\mathbf{S}}, \hat{\boldsymbol{\alpha}}\}$ is a local minimum of $\mathbf{J}(\mathbf{S}, \boldsymbol{\alpha})$.

Proof: Let $\mathbf{r}_i = (\mathbf{y}_i - (\hat{\mathbf{a}})^T \mathbf{x}_i)^2$. Rank \mathbf{r}_i^2 in increasing order as in (4.3). Then by definition of valid partition

$$\tilde{\mathbf{S}} = \{(\mathbf{x}_{i_k}, \mathbf{y}_{i_k}), k = L+1, \dots, N\} \quad (4.4)$$

By definition of $\hat{\mathbf{S}}$, $\mathbf{J}(\hat{\mathbf{S}}, \hat{\boldsymbol{\alpha}}) < \mathbf{J}(\mathbf{S}, \hat{\boldsymbol{\alpha}})$ VS, $|\mathbf{S}| = \hat{\mathbf{S}}, \mathbf{S} \subset \mathbf{A}$.

Since $\hat{\boldsymbol{\alpha}}$ is a LS estimate using \mathbf{S} , $\mathbf{J}(\hat{\mathbf{S}}, \hat{\boldsymbol{\alpha}}) \leq \mathbf{J}(\hat{\mathbf{S}}, \boldsymbol{\alpha}) \forall \boldsymbol{\alpha} \in \mathbf{R}^{p+1}$.

V. Miscellaneous Topics

A) Comparison of the inlier candidate sets S_L^* , $L = 1, 2, \dots$

By using the methods of Section II, we obtain a sequence of data subsets S_1^* , S_2^* , etc., with # outliers = 1, 2, etc. We will not develop **here** a multiple hypothesis test to compare them. **Pairwise** comparison does not seem to be useful. Instead, we will evaluate them by several different criteria which do not involve thresholds.

The key idea of the paper has been that we do not test each data point as an outlier or inlier, but we compare different partitions of the data into inlier and outlier sets. To evaluate the different partitions we need a measure of the separation between the inlier and outlier classes of the partition. One such measure suggested by the definition of outlier in (1.1) or (1.3) is the interclass distance (ICD).

$$\text{ICD} = [\min_{(x_i, y_i) \in \bar{S}_L} |r_i| - \max_{(x_i, y_i) \in S_L} |r_i|] / \sigma_L$$

where

$$\sigma_L^2 = \frac{1}{N-L} \sum_{(x_i, y_i) \in S_L} (y_i - \alpha_S^T x_i)^2.$$

The larger the value of ICD, the better the partition. This will be the main criterion.

We can also compare the models using σ_L . σ_L decreases as L increases. If the change from σ_L to σ_{L+1} is not significant, then the model with smaller L is **preferred**.

Thirdly, we can use the usual MAD statistic, the median of all the N absolute residuals.

Finally, we can develop a likelihood criterion involving all the observations by modelling the y values of outliers by a Gaussian density $\mathbf{p}(y_i; \boldsymbol{\beta}, \rho_o)$. This can be done **only** if the number of outliers is not very small. We compute the overall negative log likelihood of all the N observations, say $\mathbf{J}(\mathbf{S}_L)$, using the partition $\{\mathbf{S}_L^*, \mathbf{A} - \mathbf{S}_L^*\}$ and choose the one with the least value.

$$\mathbf{J}_4(\mathbf{S}_L^*) = (N-L) \ln \sigma_L^2 + L \ln \hat{\rho}_{O,L}$$

where $\hat{\rho}_{O,L}$ is the mean of the squares of the outlier residuals.

$$\hat{\rho}_{O,L} = \sum_{(x_i, y_i) \in \mathbf{A} - \mathbf{S}_L} (y_i - \bar{y})^2 / L, \quad \bar{y} = \sum y_i / L$$

B) Local minima using univariate optimum estimators

If the outliers come from sensor errors, it is likely that an inspection of the individual sequences $\{\mathbf{y}_k\}$, $\{\mathbf{x}_{ki}, k = 1, \dots, N\}$, $\mathbf{i} = 1, \dots, \mathbf{p}$ can help us to determine the set of the outliers in these sequences and thus provide a candidate for the outlier array $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ in the entire data. We use the univariate outlier estimation methods mentioned in Section III with much success since they give us precisely a set of outliers of specified length L . This is especially useful when the data size N is large and L is not very small. If the sequence $\{\mathbf{x}_{ki}, k = 1, \dots, N\}$ of the i th independent variable \mathbf{x}_i , yields the outliers $\{\mathbf{x}_{i_1, i}, \mathbf{x}_{i_2, i}, \mathbf{x}_{i_3, i}, \dots\}$ etc. then the $(\mathbf{x}_{i_1}, \mathbf{y}_{i_1})$, $(\mathbf{x}_{i_2}, \mathbf{y}_{i_2})$, etc. are candidates for the outliers in set \mathbf{A} . We can test the corresponding partition for local minima very easily. In Hawkins-Bradu-Kass data set with $N = 75$ or the Rousseeuw data with $N = 20$, the outlier set given by this method has also been the global minimum. The reason for the success of this method is related to the statistic $\mathbf{r}_i(\mathbf{v}_i - \bar{\mathbf{v}})$ mentioned in comment 5 of Theorem 4.

VI. Examples

We consider seven different data sets, two of which are simulated and the remaining five represent some physical process. The number of observations N ranges from 16 in the engine **knock** data set to 75 in the Hawkins-Bradu-Kass set. The number of independent variables p range from 1 to 5.

For each data set, we list the sets \mathbf{S}_L^* for $L = 1, 2, 3, \dots$, along with the values of four statistics for comparison mentioned earlier in Section V, namely the Interclass distance (ICD), \mathbf{a} , the standard deviation (root mean square) of the inlier residuals, the median of the absolute values of all residuals (MAD) and the **likelihood** statistic J_L mentioned in Section V. We also give in each case the regression coefficients with their standard deviation $\sqrt{(\mathbf{P}_S)_{ii}} \cdot \sigma$, $i = 1, \dots, p+1$.

The model with the highest value of interclass distance is often the best model.

Example 1 (stackloss data): This real life data set introduced by **Brownlee** (1965) has $N = 21$ points and has $p = 3$ independent variables. It has been studied by numerous investigators [**Andrews**, 1974; Rupert and **Carrol**, 1980; Rousseeuw and Leroy, 1987 (which lists several other references)]. The conclusion is that the data set has 4 outlier points **1,3,4,21** and possibly the point 2.

TABLE 1 (Stackloss Data)

| L | \bar{S}_L^* | ICD | σ | MAD | J_L | α_1 | α_2 | α_3 | α_4 |
|---|------------------|------|----------|--------|---------|----------------|-----------------|------------------|------------------|
| 0 | {} | - | 2.918 | 1.9175 | | 0.72 (0.12) | 1.29 (0.33) | -0.152 (0.14) | -39.92 (10.7) |
| 4 | {1,3,4,21} | 3.56 | 1.095 | 1.0579 | 21.7225 | 0.80 (0.06) | 0.577 (0.15) | -0.067 (0.05) | -37.65 (4.14) |
| 5 | {1,3,4,13,21} | 1.60 | 0.887 | 0.8496 | 21.0690 | 0.85 (0.05) | 0.445 (0.12) | -0.092 (0.04) | -35.41 (3.43) |
| 6 | {1,3,4,13,20,21} | 0.42 | 0.794 | 0.8598 | 22.7080 | 0.84 (0.04) | 0.45 (0.11) | -0.078 (0.04) | -36.72 (3.12) |

We give in Table 1 the results of the optimal sets of outliers \bar{S}_4^* , \bar{S}_5^* , \bar{S}_6^* with 4,5,6 outliers, respectively, along with the coefficients and interclass distance. Among them, [1,3,4,21] yields the maximum value of **normalized** class separation distance, namely 3.56. Hence S_4 is preferred. The residual plot of model S_4^* is in Figure 1.

With $S_4 = \{1,3,4,21\}$, the normalized residual of (x_2, y_2) (r/σ) is only 1.05 and hence (x_2, y_2) cannot be considered as a potential outlier. The coefficients associated with S_5 are within one standard deviation of the corresponding values of S_4 . They are all consistent.

Our next task is to explain the outliers. We analyze the individual data sets $\{y_k\}$, $\{x_{k1}\}, \dots, \{x_{k3}\}$ separately using the method of Section III. We fix the number of outliers as 4. The four outliers with y (stackloss) data is {1,2,3,4} which is a local minimum of the entire data. The four outliers with x_1 (rate) data is {1,2,3,21}. The x_2 (temp) has several points with the same value. The outlier set with $L = 4$ arbitrarily breaks up the subset of variables having the same value into some outliers and others as inliers. We can avoid the problem with $L = 6$ in which case the outlier subset is [1,2,3,4,7,8].

We see that points 1,3 are in all of them. Although the point 2 is in all of them, it is not an outlier because its relatively large values for all the variables are consistent.. If we plot the

residuals of the least square method with data excluding {1,2,3,21}, then the residual plot indicates that point 4 has a large residual even though it is used as an **inlier** here. Adding {4} to the outlier set gives us the set {1,2,3,4,21}. We can remove {2} since it is not a dominant outlier.

Example 2 (Rousseeuw **data**): This data set with $N = 20$ and $p = 5$ was introduced by Rousseeuw [1984]. It is obtained by adding outliers to a data set in [Draper and Smith, 1966] dealing with wood specific gravity. The outliers are 4,6,8 and 19.

TABLE 2 (Rousseeuw Data)

| L | \bar{S}_L^* | ICD | σ | MAD | J_L | α_1 | α_2 | α_3 | α_4 | α_5 | α_6 |
|---|----------------|-------|----------|--------|---------|---------------|----------------|----------------|----------------|---------------|---------------|
| 0 | {} | - | 0.020 | 0.0129 | - | 0.44 (.10) | -1.48 (.41) | -0.26 (.09) | 0.021 (.13) | 0.17 (.17) | 0.42 (.14) |
| 4 | {4,6,8,19} | 30.11 | 0.006 | 0.0065 | -196.63 | 0.22 (.03) | -0.09 (.16) | -0.56 (.03) | -0.40 (.05) | 0.61 (.06) | 0.37 (.04) |
| 5 | {4,5,6,8,19} | 2.223 | 0.005 | 0.0052 | -190.87 | 0.20 (.03) | 0.03 (.13) | -0.57 (.03) | -0.34 (.04) | 0.53 (.05) | 0.41 (.03) |
| 6 | {4,5,6,7,8,19} | 1.570 | 0.004 | 0.0045 | -190.57 | 0.24 (.03) | -0.09 (.12) | -0.57 (.02) | -0.34 (.04) | 0.59 (.06) | 0.35 (.04) |

In this example, $\bar{S}_1 = \{11\}$, $\bar{S}_2 = (3,111)$. Only \bar{S}_4 yields {4,6,8,19}, \bar{S}_5 and \bar{S}_6 are supersets of \bar{S}_4 . The corresponding statistics are in Table 2. Note the large value of interclass distance, namely 30.12, with the set {1,3,4,21}. Clearly it is the **preferred** model. The points 5 and 7 occurring in S_5^* and S_6^* are clearly very weak outlier candidates since the ICD falls from 30.12 to 2.22 in S_5^* and to 1.57 in S_6^* . The residual plot of S_4^* is given in Figure 2.

The MAD statistics with S_6^* , (0.0045) is less than the MAD value given by the LMS solution which is 0.0048. LMS ideally should give the minimum achievable value of MAD. The program to compute the LMS solution is a Monte Carlo procedure, which, by its very nature, cannot guarantee a global minima. The LMS criteria has numerous local minima.

This example also illustrates the power of the one dimensional robust scheme of this paper. If we analyze the **univariate** data $\{y_k\}$, $\{x_{k4}\}$, $\{x_{k5}\}$, individually and look for optimum four outliers using the method of Section III, they are exactly the points $\{4,6,8,19\}$ in all the three cases. Thus the outliers would have been detected by a relatively simple procedure.

Masked Outliers

This data set gives us a good demonstration of the concept of masked outliers. According to the outlier free model, with α_S as in S_4^* model, the residuals of the four outliers $\{4,6,8,19\}$ are

$$\begin{aligned} r_{SO} &= \text{col.}(y_i - \alpha_S^T x_i, i=4,6,8,19) \\ &= \text{col.}(-0.19, -0.22, -0.21, -0.25) \end{aligned}$$

But the residuals of the same four points computed from the coefficient α_A **obtained** for all the data is

$$\begin{aligned} r_{AO} &= \{(y_i - \alpha_A^T x_i, i=4,6,8,19)\} \\ &= \text{col.}\{0.0085, -0.0109, -0.0008, -0.0261\} \end{aligned}$$

i.e., the outliers are masked in the residuals computed from the complete data set. As mentioned, the difference between them can be explained as in Section II.

$$r_{SO} = (I - H)^{-1} r_{AO}$$

$$I - H = \begin{matrix} & 0.74 & -0.25 & -0.18 & -0.21 \\ -0.25 & 0.74 & -0.20 & -0.23 \\ -0.18 & -0.20 & 0.71 & -0.28 \\ -0.21 & -0.23 & -0.28 & 0.70 \end{matrix}$$

$$(I - H)^{-1} = \begin{matrix} & 7.02 & 6.29 & 6.35 & 6.75 \\ 6.29 & 7.60 & 6.68 & 7.09 \\ 6.35 & 6.68 & 7.91 & 7.30 \\ 6.75 & 7.09 & 7.30 & 8.72 \end{matrix}$$

Example 3 (water salinity data): This data set introduced by Rupert and Carrol (1980) has $N = 28$ points with $p = 3$ independent variables. This data set has also been investigated by several authors. The x_3 component of data point 16 is very large, indicating a possible outlier. LMS method [Rousseeuw and Leroy, 1987] yields {16,5,23,24} as outliers whereas Rupert and Carrol's method yields {1,11,13,15,16,17} as outliers.

Here $\bar{S}_1^* = \{16\}$ and \bar{S}_i^* is superset of \bar{S}_{i-1}^* . The results are given in Table 3.

TABLE 3 (Water Salinity Data)

| L | \bar{S}_L^* | ICD | σ | MAD | J_L | α_1 | α_2 | α_3 | α_4 |
|---|----------------|-------|----------|--------|---------|-----------------|------------------|------------------|-----------------|
| 0 | } | - | 1.232 | 0.7178 | - | 0.777 (0.08) | -0.026 (0.15) | -0.30 (0.10) | 9.590 (2.89) |
| 2 | {15,16} | 0.470 | 0.878 | 0.4610 | -18.754 | 0.728 (0.06) | -0.242 (0.11) | -0.64 (0.10) | 18.60 (2.83) |
| 3 | {15,16,17} | 1.285 | 0.763 | 0.498 | -11.995 | 0.741 (0.05) | -0.241 (0.10) | -0.596 (0.90) | 17.65 (2.48) |
| 4 | {5,15,16,17} | 1.107 | 0.686 | 0.5321 | -11.531 | 0.741 (0.05) | -0.321 (0.09) | -0.775 (0.11) | 21.93 (2.77) |
| 5 | {5,8,15,16,17} | 0.145 | 0.635 | 0.569 | -13.857 | 0.724 (0.04) | -0.279 (0.09) | -0.786 (0.10) | 22.30 (2.56) |

Model with outliers {15,16} is not acceptable because its interclass distance is very small (0.71). Model \mathbf{S}_3^* with outliers {15,16,17} is preferable with its ICD of 1.2851. The model coefficients of \mathbf{S}_4^* and \mathbf{S}_5^* are within one standard deviation of the coefficients of \mathbf{S}_3^* . The residuals of the model \mathbf{S}_3^* are in Figure 3.

Here **univariate** methods do not throw any light. With $L = 4$, \mathbf{x}_1 yields outliers {3,4,5,6}; \mathbf{x}_2 , {3,9,15,19,23}; \mathbf{x}_3 , {3,5,16,24}; \mathbf{y} , {3,4,5,6}.

Example 4 (steam data): This data set, originally introduced by Draper and **Smith** [1966], has $N = 25$ points and $p = 2$ independent variables. **Hampel** et al. [1986] used hypotheses testing methods to explore the significance of the coefficients of the fitted model. They noticed the two extreme values of the \mathbf{x}_2 ($= \mathbf{x}_6$) variable in the points 7 and 19. Our analysis shows that there are other outlier points 11,23, which yield large positive residuals with respect to the correct model and whereas points 7 and 19 yield large negative residuals. Each group of outliers by itself yields a model which masks the effect of the other.

The results of our study are in Table 4. $\bar{\mathbf{S}}_1^*$, the optimal outlier set with one variable was (11) and $\bar{\mathbf{S}}_2^*$ was (11,23). Thus (7,19) was not the optimal set for $L = 2$. But the model with (7,19) outliers is a local minima.

TABLE 4 (Steam Data)

| L | \bar{S}_L^* | ICD | σ | MAD | J_L | α_1 | α_2 | α_3 |
|---|---------------------|-------|----------|--------|---------|---------------------------|-------------------|-------------------|
| 0 | {} | - | 0.621 | 0.4478 | - | -0.0724 (0.008) | 0.2028 (0.043) | 9.1269 (1.03) |
| 2 | {11,23} | 0.884 | 0.491 | 0.4512 | -37.421 | -0.0755 (0.006) | 0.1958 (0.034) | 9.5525 (0.825) |
| 2 | {7,19} [†] | 5.182 | 0.509 | 0.360 | -36.849 | -0.0822 (0.007) | 0.5246 (0.093) | 2.8135 (1.882) |
| 3 | {7,11,19} | 0.936 | 0.440 | 0.3051 | -37.449 | -0.0823 (0.006) | 0.4820 (0.081) | 3.7732 (1.660) |
| 4 | {7,11,19,25} | 0.650 | 0.388 | 0.3919 | -34.866 | -0.0875 (0.006) | 0.5361 (0.075) | 2.9525 (1.495) |
| 5 | {7,11,19,23,25} | 0.873 | 0.336 | 0.3056 | -38.502 | -0.0880 (0.005) | 0.5048 (0.065) | 3.6849 (1.32) |

†: not optimal

This is an excellent example with two very strong negative outliers (i.e., outliers with negative residual), namely 7 and 19, and three weak positive outliers, 11, 23 and 25. The optimal outlier set with $L = 2$ is {11,23} whose O value is only slightly less than the O value of {7,19} model. But, their corresponding coefficients are very different. The {7,19} outlier model has a very large value of interclass distance (5.18) indicating its superiority. Its only drawback is the high standard deviation associated with the coefficient α_3 (intercept).

The optimal set with $L = 3$ yields the outliers {7,11,19}. Its ICD value (0.936) is much less than before. But the standard deviation of α_3 is 1.66 which is much less than the value of a , namely 3.77. Its MAD value of 0.305 is less than the earlier case.

The optimal set with $L = 5$ and outliers {7,11,19,23,25} is similar to the model with {7,11,19}. Its ICD value (0.873) is less than that of {7,11,19}. Note the coefficients (α_1, α_2) of models S_3^* , S_4^* , S_5^* and that with (7,19) are similar. They differ only in the α_3 coefficient. Summing up, the model with {7,11,19} or {7,11,19,23,25} as outliers can be preferred. The residuals of the model S_3^* and S_5^* are graphed in Figures 4A and 4B.

Example 5 (engine knock data): This data set [Mason et al., 1982] was analyzed in [Hettmansperger and Sheather, 1992] in great detail by using both least squares and least median squares methods, without much success. The data has 16 observations, with $p = 4$, the dependent variable being the engine knock. Accidentally one of the investigators in [Hettmansperger and Sheather] **mistyped** air-component value of 2nd observation as 15.1 instead of 14.1, the correct value. Both 14.1 and 15.1 are not the extreme values of the corresponding component data set. Both the LS and LMS estimates with the two data sets 14.1 and 15.1 are given below, LMS values being repeated from the paper quoted above.

| | α_1 | α_2 | α_3 | α_4 | α_5 |
|---------------|------------|------------|------------|------------|------------|
| LS with 14.1 | 1.1 | 2.19 | 0.93 | -0.002 | 12.01 |
| | (1.03) | (0.92) | (0.33) | (0.019) | (29.03) |
| LS with 15.1 | 1.06 | 1.69 | 1.17 | -0.008 | 15.97 |
| | (1.06) | (0.79) | (0.28) | (0.02) | (29.5) |
| LMS with 14.1 | 0.21 | 2.9 | 0.56 | -0.009 | 30.1 |
| LMS with 15.1 | 4.6 | 1.2 | 1.5 | 0.069 | -86.5 |

The variability in LS coefficients are consistent with the corresponding standard deviation, but the standard deviation of the coefficients are very high. But the variability in the LMS coefficients are drastic. The authors speculated that the drastic differences between the two sets may be caused by several outliers in the data set, but could not determine the outliers.

Analysis of the data set 14.1

The results of our analyses are in Table 5. The optimum model S_3^* with outlier set {5,7,9,13,15} is superior in all the four categories. Note the relatively high interclass distance of 7.14. The a has dropped from 1.64 (for all data) to 0.157. The MAD value of 0.206 is lower than the MAD value of the LMS result, namely 0.226.

TABLE 5 (Engine **Knock-14.1** Data)

| L | \bar{S}_L^* | ICD | a | MAD | J_L | α_1 | α_2 | α_3 | α_4 | α_5 |
|---|------------------|-------|-------|--------|---------|------------------|-----------------|-----------------|------------------|-----------------|
| 0 | {} | - | 1.645 | 0.8494 | - | 1.095 (1.03) | 2.189 (0.92) | 0.935 (0.33) | -0.002 (0.02) | 12.01 (29.0) |
| 2 | {5,15} | 0.481 | 0.855 | 0.4756 | -2.1600 | 0.181 (0.57) | 3.025 (0.49) | 0.550 (0.18) | 0.002 (0.01) | 21.67 (15.8) |
| 3 | {5,13,15} | 4.036 | 0.491 | 0.3234 | -15.010 | -0.59 (0.35) | 2.860 (0.29) | 0.440 (0.11) | -0.011 (0.01) | 45.90 (10.1) |
| 4 | {5,9,13,15} | 2.538 | 0.331 | 0.3104 | -22.143 | -0.493 (0.24) | 3.136 (0.20) | 0.347 (0.07) | -0.011 (.004) | 43.65 (6.86) |
| 5 | {5,7,9,13,15} | 7.146 | 0.157 | 0.2062 | -27.923 | -0.028 (0.13) | 2.949 (0.10) | 0.477 (0.04) | -0.009 (.002) | 35.11 (3.50) |
| 6 | {5,7,9,12,13,15} | 3.319 | 0.104 | 0.1383 | -30.907 | 0.003 (0.09) | 2.852 (0.07) | 0.525 (0.03) | -0.008 (.001) | 33.75 (2.35) |

Note that the coefficient α_2 is very significant, namely 2.85. So when we alter the value of the x_2 -component of the data point 2 from 14.1 to 15.1, the corresponding change in $(\alpha_2 x_2)$ is 2.85 which is about **18 times** the standard deviation of a which is 0.1569. **Obviously** it becomes an outlier even though it may not appear as an outlier by casual inspection of the data. This fact is **confirmed** by the analysis of data set 15.1. The residual plots are in Figure 5.

Analysis of the 15.1 data set

The results of this data set are in Table 6. Again we analyzed this data set from the beginning. Note that the data point 2 starts appearing in the outlier sets from \bar{S}_4^* onward. The optimum set is S_6^* . It is superior in all the four categories, namely highest ICD, least **a**, least MAD and least J_L . The ICD value of 7.68 is fairly high showing the clear separation of the outlier and inlier classes.

Note \bar{S}_5^* is a singularity. It is not a **superset** of \bar{S}_4^* , nor a subset of \bar{S}_6^* . Its coefficients are quite distinct from the coefficients of others. If this set were analyzed in isolation, it looks good

with relatively high ICD, low MAD, low J_L , etc. Its singularity appears only when we compare it with others. The example indicates the dangers of accepting a solution looking only at the residual plots, without considering other solutions.

TABLE 6 (Engine Knock-15.1 Data)

| L | \bar{S}_L^* | ICD | a | MAD | J_L | α_1 | α_2 | α_3 | α_4 | α_5 |
|---|--------------------|-------|-------|--------|---------|------------------|------------------------|------------------------|--------------------------|-----------------|
| 0 | {} | - | 1.690 | 1.0132 | - | 1.063 (1.06) | 1.695 (0.80) | 1.172 (0.28) | -0.008 (0.02) | 15.97 (29.5) |
| 2 | {5,7} | 1.660 | 0.977 | 1.0501 | 4.3823 | 1.447 (0.71) | 1.944 (0.49) | 1.191 (0.18) | -0.0002 (0.01) | 1.032 (18.2) |
| 3 | {5,13,15} | 2.317 | 0.776 | 0.8062 | -3.1508 | -0.616 (0.56) | 2.147 (0.38) | 0.77 (0.14) | -0.019 (0.01) | 51.70 (15.6) |
| 4 | {2,5,13,15} | 3.907 | 0.504 | 0.4657 | -5.6013 | -0.598 (0.36) | 2.898 (0.30) | 0.40 (0.12) | -0.009 (.007) | 45.94 (10.4) |
| 5 | {4,5,7,12,14} | 6.075 | 0.257 | 0.2014 | -16.975 | 4.719 (0.34) | 1.058 (0.18) | 1.569 (.054) | 0.068 (.006) | -88.73 (8.9) |
| 6 | {2,5,7,9,13,15} | 7.678 | 0.143 | 0.1677 | -22.259 | -0.05 (0.12) | 2.990 (.095) | 0.441 (0.04) | -0.008 (.002) | 35.36 (3.19) |
| 7 | {2,5,7,9,12,13,15} | 4.747 | 0.079 | 0.1140 | -27.233 | -0.02 (0.07) | 2.893 (.050) | 0.491 (0.03) | -0.007 (.001) | 34.02 (1.79) |

Note again that the MAD value of 0.168 in \bar{S}_6^* is less than the MAD value of the LMS result. The residual plot is given in Figure 6. It clearly indicates that point 2 is a dominant outlier.

Note that the outliers associated with the optimum model S_5^* with 14.1 set, namely {5,7,9,13,15} are the same as the outliers of the optimum model S_6^* with 15.1 set, namely {2,5,7,9,13,15} but for the point 2 which has been explained already. The regression coefficients of the outlier free models S_5^* in 14.1 set and S_6^* in 15.1 set are fairly close to one another. Thus our method offers protection against accidental typing errors also.

Example 6 (Hawkins-Bradu-Kass (1984) data): This is a simulated data set with $N = 75$ and $p = 3$. The data has ten extreme outliers and four other points which obey the regression model,

but are located away from other inliers.

We performed only the analysis of the four univariate data sets. In addition, we chose the J_L criteria described in Section V to fix L , i.e., choose L which yields the least value of J_L .

The data set $\{y_k\}$ yielded $L = 10$ and indicated the outliers as points 1 through 10. The data sets $\{x_{k1}\}$, $\{x_{k2}\}$ and $\{x_{k3}\}$ yielded $L = 14$ and indicated as outliers the points 1 through 14. The residuals are graphed in Figure 7. The distinction between the **inliers** and outliers is very clear. All of them detect the ten (bad) outliers. In addition, $\{x_{1k}\}$, $\{x_{2k}\}$ and $\{x_{3k}\}$ detect the four other points which are located away from others, even though they obey the linear model.

Example 7 (Herksprung-Russel star data): This data set was introduced by Rousseeuw and Leroy [1987]. Here $N = 47$, $p = 1$ with $x = \log$ temperature of the star and $y = \log$ intensity of the star. It has four very strong outliers, the so-called giant stars, points 11,20,30,34 (not errors). The LS fit to the entire data is almost perpendicular to the correct fit. The standard methods of diagnostics as discussed in [Rousseeuw and Leroy] do not indicate any outliers.

Since $p = 1$, the computation of statistics is very simple without involving any matrix **inversion** as indicated in Theorem 5. S_4^* indicates the 4 giant stars (points 11,20,30,34) as the outlier and the interclass distance is very large.

The univariate method applied to the x data with $L = 4$ picks $\{11,20,30,34\}$ as the outliers. This heuristic $r_i(v_i - \bar{v}_A)$ mentioned in the comment after Theorem 5 gave excellent results. Note \bar{v}_A is the average of all x -values. The statistic is graphed versus i in Figure 8. The 6 outliers **are** located in Figure 9. This includes the 4 joint stars **and** some others. The least squares fit given by the corresponding inlier set is closer to the "correct" linear fit to the data.

The same heuristic is successful in detecting the outliers in the one dimensional simulated data set of Rousseeuw [1984] which has defied all the attempts at solution except LMS. The heuristic gives a set of inliers whose least squares fit is close to the correct linear model.

Discussion of Examples

The precise choice of L is not crucial for the choice of the regression coefficients. In all the examples, once the dominant outliers have been removed, *i.e.*, $L > L_0$, some integer, then the corresponding regression coefficients stabilize and increasing L further **results** only in a minor perturbation of these coefficients. The perturbation is within one or two **of** their standard deviations.

Notice also the σ value, the standard deviation of the inlier residuals of the final model is substantially less than the original σ value with all data. In many examples, the decrease is an order of magnitude. Clearly in these examples, the outliers are not errors, but **observations** which do not obey the linear model.

Next the role played by the interclass distance (ICD) statistic is interesting. Except in special cases like the steam data of example 3, the chosen model has the highest ICD. The overall likelihood statistic J_L is also useful.

In some of the examples, the model with high ICD has also the least value of the MAD. We have shown at least two instances in which the MAD values are less than the value given by the LMS method, clearly indicating that the computational procedure for **computing** LMS is giving a local minimum.

The examples clearly illustrate that when the model is free of dominant outliers, then any slight perturbation of data points will cause only minor perturbation in the regression coefficients. We can develop precise upper bounds on their variability.

The simple method of searching for outliers in the univariate data sets of the components has been useful in detecting the overall systems outliers in several, though not **all**, problems.

Finally, the use of the suggested model for forecasting is done as follows. If the given x value is closer to the x value of inliers according to some standard statistic, the model is used for forecasting. If the x value is closer to the outliers, then the forecast of y is just the empirical mean of the y values of the outliers.

VII. Discussion and Comparison

We will make a brief comparison of our methods with others in the **literature**. There have been several attempts at getting an estimate of \mathbf{a} by trimming the residual squares [Rousseeuw and Leroy, 1987], [Rupert and **Carrol**, 1980]. We have exploited this idea in a systematic manner.

Huber's and related theories of robust estimation [Huber, 1981; **Hampel**, et al., 1986] assume that all the members of the given data set have a common distribution. There is no empirical support for this assumption. All successful attempts at generating contaminated data in regression utilize two distributions, one for inliers and the other for outliers. There has been no analysis of the reasons for the failure of the M estimates in the presence of even a single outlier [**Hampel** et al., 1987; Rousseeuw, 1984].

The least median squares (LMS) approach is the most popular method in robust regression. It offers an interesting criterion for minimization. LMS is clearly a useful data explanatory tool. Any algorithm will yield a set of outliers. There is no discussion why the set of outliers given by the LMS method is superior to other sets. In examples with one independent variable ($p = 1$), it gives good results. However, when $p > 1$, the global minimum needed in the LMS is computed using Monte **Carlo** type procedures and there are doubts whether this numerical procedure indeed gives the global minima. A principal support of the LMS approach is its nearly fifty

percent breakdown point. We have shown elsewhere that high breakdown estimates for the regression problem can be constructed, but they may be irrelevant as solutions to the regression problem.

The key idea of this paper is the characterization of robust estimation as a problem of classification of sets. Outliers are those which do not obey the linear model characterizing the inliers and consequently their residuals are large in magnitude with those of the inliers. As a consequence, the two problems of outlier detection and parameter estimation must be handled simultaneously, not sequentially as in the earlier studies. The concept of valid partition of a contaminated data set \mathbf{A} was introduced and showed that $(\hat{\mathbf{S}}, \hat{\boldsymbol{\alpha}})$ is every local minimum of the function $\mathbf{J}(\mathbf{S}, \boldsymbol{\alpha})$ if and only if the partition $\{\hat{\mathbf{S}}, \mathbf{A} - \hat{\mathbf{S}}\}$ is valid. A valid partition clearly picks out some of the outliers. Only the global minimum of $\mathbf{J}(\mathbf{S}, \boldsymbol{\alpha})$ picks out all the outliers provided that L , the number of outliers, is known.

There have been a large number of papers, as discussed in the book by Rousseeuw and Leroy [1987], which deal with the topic of diagnostics, of the least squares result for possible presence of outliers. Many papers, beginning with Cook [1977], deal with the effect of deleting one observation at a time. Some others [Cook and Weisberg, 1982, and others] considered the effect of deleting several observations at one stroke and determining the change in the regression coefficients. But the key problem is the choice of the subset to be deleted.

A result of great instrumental importance is the Theorem 4 which allows us to evaluate the $\mathbf{J}_1(\mathbf{S})$, the log likelihood of a set of $(N-L)$ observations and thus determine the optimal set yielding the global optimum. This result nicely connects the concepts of robust regression and robust diagnostics. This method allows us to compare the effect of deleting all possible subsets of data as long as the data size N and subset size L are not large. Development of effective heuristics to limit the computation when N and/or L are large needs attention.

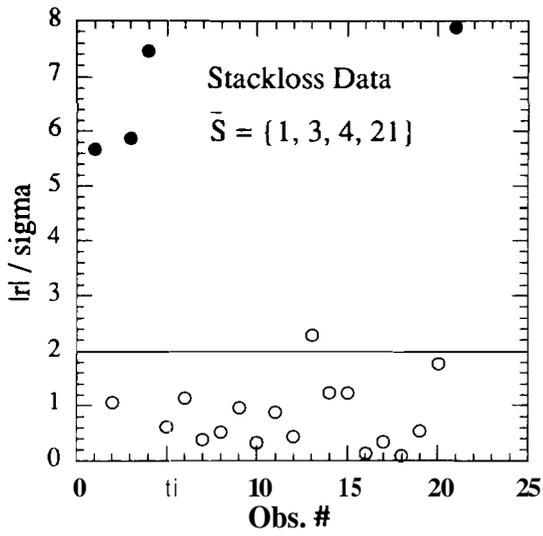


Figure 1 : Normalized residuals

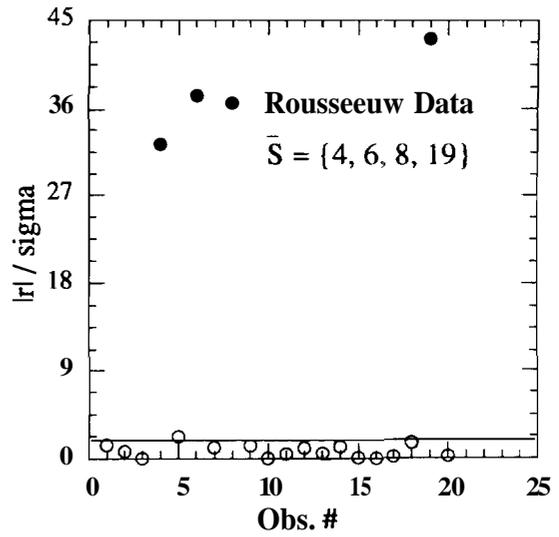


Figure 2 : Normalized residuals

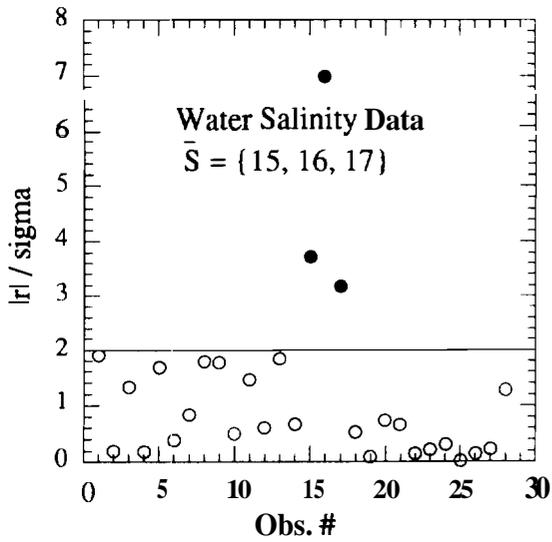


Figure 3 : Normalized residuals

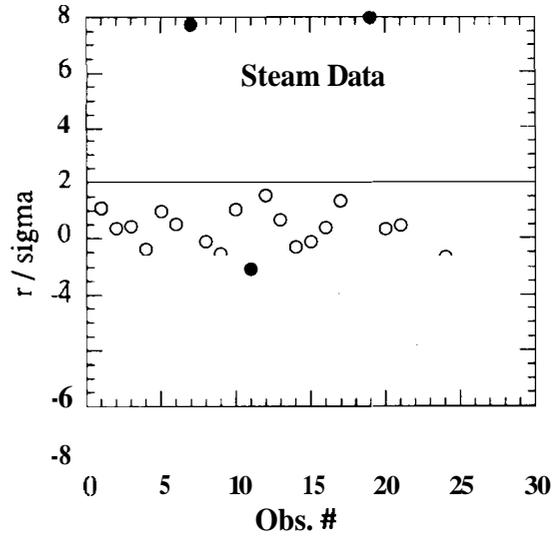


Figure 4a : Normalized residuals

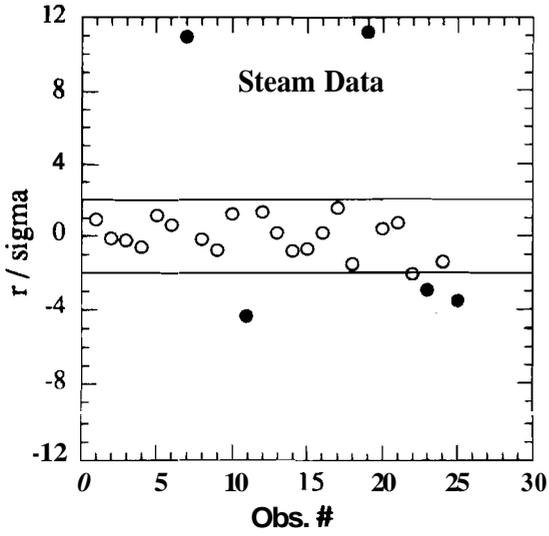


Figure 4b : Normalized residuals

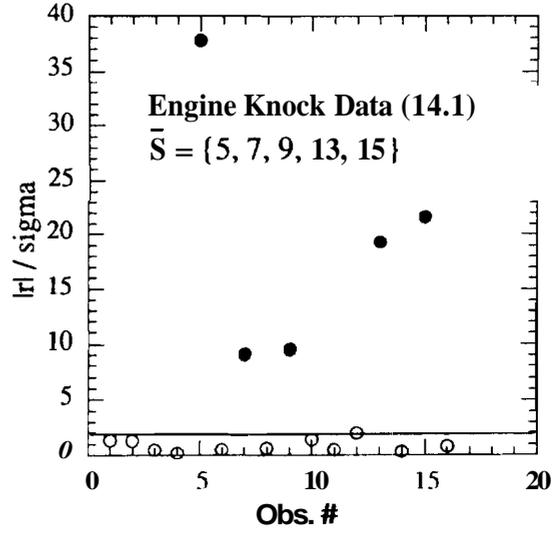


Figure 5 : Normalized residuals

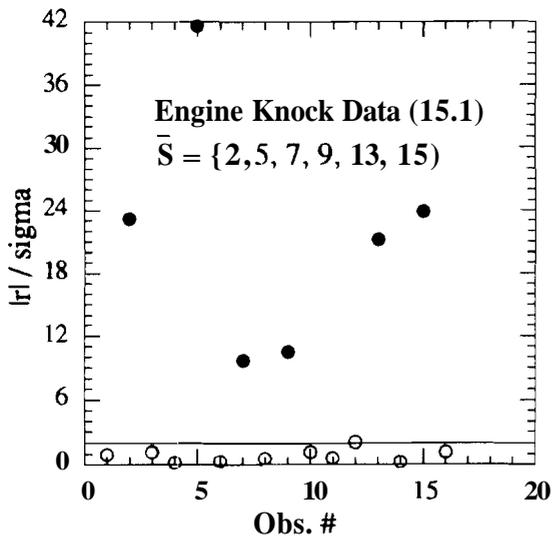


Figure 6 : Normalized residuals

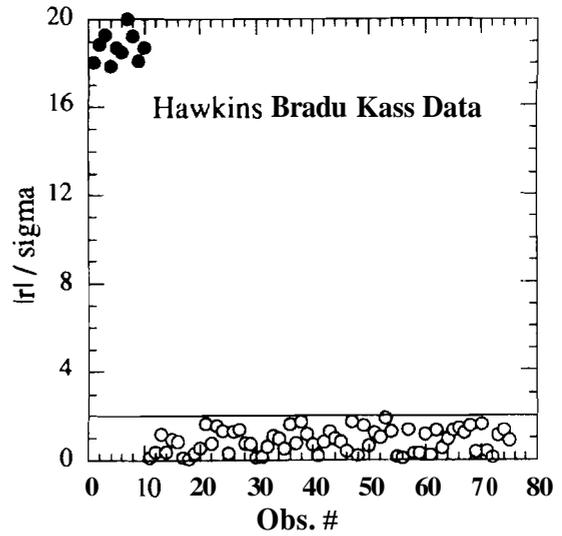


Figure 7 : Normalized residuals

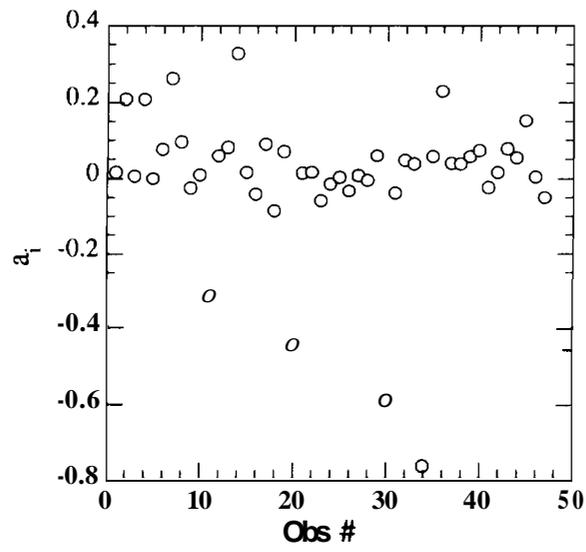


Figure 8: Star Data

$$a_i = r_i(x_i - \bar{x})$$

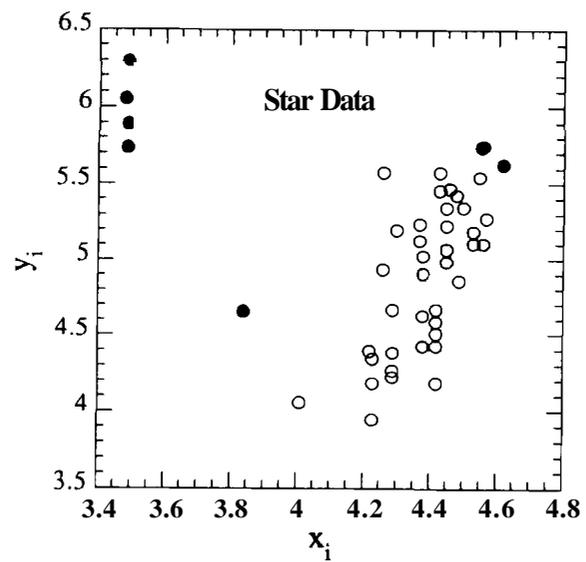


Figure 9: Selected Outliers

References

- [1] **Andrews, D.F. (1974)**, "A Robust Method for Multiple Linear Regression," *Technometrics*, Vol. 16, pp. 523-531.
 - [2] **Brownlee, K.A. (1965)**, *Statistical Theory and Methodology in Science and Engineering*, New York: Wiley.
 - [3] **Cook, R.D. (1977)**, "Detection of Influential Observation in Linear Regression," *Technometrics*, Vol. 19, pp. 15-18.
 - [4] **Cook, R.D. and Weisberg (1982)**, *Residuals and Influence in Regression*, London: Chapman & Hall.
 - [5] **Donoho, D.L., and Huber, P.J. (1983)**, "The Notion of Breakdown Point," in *A Festschrift for E.L. Lehmann*, Wadsworth.
 - [6] **Draper, N.R., and Smith, H. (1966)**, *Applied Regression Analysis*, New York: Wiley.
 - [7] **Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986)**, *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.
 - [8] **Hawkins, D.M., Bradu, D., and Kass, G.V. (1984)**, "Location of Several Outliers in Multiple Regression Using Elemental Sets," *Technometrics*, Vol. 26, pp. 197-208.
 - [9] **Hettmansperger, T.P., and Sheather, S.J. (1992)**, "A Cautionary Note on the Method of Least Median Squares," *American Statistician*, Vol. 46, No. 2, pp. 79-83.
 - [10] **Huber, P.J. (1981)**, *Robust Statistics*, New York: Wiley.
 - [11] **Mason, R.L., Gunst, R.F., and Hess, J.L. (1989)**, *Statistical Design and Analysis of Experiments*, New York: John Wiley.
 - [12] **Rousseeuw, P.J. (1984)**, "Least Median Squares Regression," *J. Am. Stat. Assoc.*, Vol. 79, pp. 871-880.
 - [13] **Rousseeuw, P.J., and Leroy, A.M. (1987)**, *Robust Regression and Outlier Detection*, John Wiley.
-

- [14] Rupert, D. and Carrol, R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *J. of Amer. Stat. Assoc.*, Vol. 75, pp. 828-838.

Appendix

Proof of Theorem 2: Let us denote $H_{A,L}$ by H .

$$\text{Part (i): } P_S^{-1} = \sum_{(x_i, y_i) \in S} x_i x_i^T = P_A^{-1} - BB^T$$

$$(P_A^{-1} - BB^T)(P_A + P_A B(I - H)^{-1} B^T P_A)$$

$$= I + B(I - H)^{-1} B^T P_A - BB^T P_A - BB^T P_A B(I - H)^{-1} B^T P_A$$

$$= I + B[(I - H)^{-1} - I - H(I - H)^{-1}] B^T P_A = I.$$

Part (ii) can be proved similarly.

$$\text{Part (iii): } \alpha_S = P_S \begin{pmatrix} x_i \\ y_i \end{pmatrix}_{(x_i, y_i) \in S}$$

$$= P_S [P_A^{-1} \alpha_A - B y_0]$$

$$= (P_A + P_A B(I - H)^{-1} B^T P_A)(P_A^{-1} \alpha_A - B y_0), \text{ by part (i)}$$

$$= \alpha_A - P_A B y_0 + P_A B(I - H)^{-1} B^T \alpha_A - P_A B(I - H)^{-1} H y_0$$

$$= \alpha_A - P_A B[I + (I - H)^{-1} H] y_0 + P_A B(I - H)^{-1} B^T \alpha_A$$

$$= \alpha_A - P_A B(I - H)^{-1}(y_O - B^T \alpha_A)$$

Part (iv) can be proved similarly.

Proof of Theorem 3:

$$I + B^T P_S B = I + B^T [P_A + P_A B(I - H)^{-1} B^T P_A] B, \text{ from part (i) of Th. 2,}$$

$$= I + H + H(I - H)^{-1} H = (I - H)^{-1}.$$

Proof of Theorem 4:

$$A) \quad J_1(A) - J_1(S) = \|y_O\|^2 + \alpha_S^T P_S^{-1} \alpha_S - \alpha_A^T P_A^{-1} \alpha_A, \text{ by (2.4)} \quad (1)$$

$$P_S^{-1} \alpha_S = \sum_{(x_i, y_i) \in S} x_i y_i = P_A^{-1} \alpha_A - B y_O \quad (2)$$

$$\text{RHS of (1)} = \|y_O\|^2 + (P_A^{-1} \alpha_A - B y_O)^T (\alpha_A - P_A B(I - H)^{-1} r_{AO}) - \alpha_A^T P_A^{-1} \alpha_A,$$

using (2) and part (iii) of Th. 1,

$$= \|y_O\|^2 - (y_O)^T B^T \alpha_A + \alpha_A^T P_A^{-1} \alpha_A - \alpha_A^T B(I - H)^{-1} r_{AO} + (y_O)^T B^T P_A B(I - H)^{-1} r_{AO} - \alpha_A^T P_A^{-1} \alpha_A, \quad (3)$$

$$\text{Sum of first two terms on RHS of (3)} = (y_O)^T (y_O - B^T \alpha_A)$$

$$= y_O^T r_{AO} \quad (4)$$

Sum of remaining terms of RHS of (3)

$$\begin{aligned}
&= -\alpha_A^T \mathbf{B}(\mathbf{I} - \mathbf{H})^{-1} \mathbf{r}_{\text{AO}} + (\mathbf{y}_O)^T \mathbf{H}(\mathbf{I} - \mathbf{H})^{-1} \mathbf{r}_{\text{AO}} \\
&= -\alpha_A^T \mathbf{B}(\mathbf{I} - \mathbf{H})^{-1} \mathbf{r}_{\text{AO}} + (\mathbf{y}_O)^T (\mathbf{I} - \mathbf{H})^{-1} \mathbf{r}_{\text{AO}} - (\mathbf{y}_O)^T \mathbf{r}_{\text{AO}}, \text{ since } \mathbf{H}(\mathbf{I} - \mathbf{H})^{-1} = (\mathbf{I} - \mathbf{H})^{-1} - \mathbf{I}. \\
&= \mathbf{r}_{\text{AO}}^T (\mathbf{I} - \mathbf{H})^{-1} \mathbf{r}_{\text{AO}} - \mathbf{y}_O^T \mathbf{r}_{\text{AO}}. \tag{5}
\end{aligned}$$

Adding (4) and (5) yields the required result.

B): It follows from \mathbf{A} and Theorem 3.

Proof of Theorem 9: Since $L = 1$, the set \mathbf{A} has a **single** outlier which will be assumed to be $(\mathbf{x}_N, \mathbf{y}_N)$ without any loss of generality. LHS of (1.4) being arbitrarily large implies:

$$|y_i - \alpha_A^T x_i| / |y_N - \alpha_A^T x_N| < \delta \ll 1, \forall i, \tag{8}$$

where δ can be chosen to be arbitrarily small. Recall that the optimal outlier given by Theorem 4 is $(\mathbf{x}_i, \mathbf{y}_i)$ which maximizes $(y_i - \alpha_A^T x_i)^2 / (1 - x_i^T P_A x_i)$.

We need to prove that:

$$|y_N - \alpha_A^T x_N| / \sqrt{(1 - x_N^T P_A x_N)} \geq |y_i - \alpha_A^T x_i| / \sqrt{(1 - x_i^T P_A x_i)}, \forall i. \tag{9}$$

Let α_N be the estimate of α obtained from the subset of \mathbf{A} omitting $(\mathbf{x}_N, \mathbf{y}_N)$. The corresponding value of P is P_N .

By part (iv) of Theorem 2:

$$\alpha_A = \alpha_N + P_N x_N (1 + h)^{-1} (y_N - \alpha_N^T x_N) \tag{10}$$

$$\begin{aligned}
\text{LHS of (9)} &= |y_i - \alpha_A^T x_i| / \sqrt{1 - x_i^T P_A x_i} \\
&= \frac{[r_i - r_N(x_i^T P_N x_N)/(1+h)]}{\sqrt{(1 - x_i^T P_N x_i)(1+h) + (x_i^T P_N x_N)^2}} \sqrt{1+h}, \text{ by (11) \& (13)} \\
&\leq \frac{(|r_i|/|r_N|)(1+h) + |x_i^T P_N x_N| |r_N|/\sqrt{1+h}}{\sqrt{(1 - x_i^T P_N x_i)(1+h) + (x_i^T P_N x_N)^2}} \tag{16}
\end{aligned}$$

LHS of (9) = $(\beta_1 + \beta_2)$ [RHS of (9)]

Note α_N does not involve (x_N, y_N) . For any x_N we can choose y_N to make $|y_N - \alpha_N^T x_N|$ arbitrarily large. β_1 can be made arbitrarily small by making δ in (8) arbitrarily small. Further, $|\beta_2| < 1$.

Hence **LHS** of (9) < **RHS** of (9) $\forall i$.

Proof of Theorem 12: Let $K = N-L$.

Let $S_2 = \{x_{j_k}, k=1, \dots, K\}$. Let the members of S_2 be ordered.

$$x_{i_1} < x_{i_3} < x_{i_4} < \dots < x_{i_{L+1}} ; i_k \in \{j_i, i=1, \dots, L\} .$$

Let $y_k = x_{i_k}$. Note the absence of x_i , in the above list. Since S_2 is non-contiguous, there must exist a member, say $x_{i_2} \in S$, $x_{i_2} \notin S_2$, and $x_{i_1} < x_{i_2} < x_{i_{L+1}}$. Assume without loss of generality that $x_{i_1} < x_{i_2} < x_{i_3}$. Construct two subsets S_1 and S_{K+1} which are "less contiguous" than S_1 , by adding the member y_2 and deleting one of the two extreme values, y_{K+1} or y_1 . By "less contiguous" we mean that if there exist m members in S_2 which are not in S , but have values within the extremes x_{i_1} and $x_{i_{K+1}}$, then S_1 or S_{K+1} have only $(m-1)$ such members.

$$S_{K+1} = \{y_k, k=1, \dots, K\}, \quad S_1 = \{y_k, k=2, \dots, K+1\}, \quad S_2 = \{y_1, y_3, y_4, \dots, y_{K+1}\},$$

The subscript i in the subset S_i means y_i is absent in that subset.

$$J_i = J_1(S_i)$$

$$= \left(\sum_{k=1}^{K+1} y_k^2 - y_i^2 \right) - (1/K) \left(\sum_{k=1}^{K+1} y_k - y_i \right)^2.$$

We will prove that either $J_{K+1} < J_2$ or $J_1 < J_2$.

$$\text{Let } \bar{y} = \sum_{k=1}^{K+1} y_k.$$

$$\begin{aligned} J_2 - J_{K+1} &= y_{K+1}^2 - y_2^2 + (1/K)[(\bar{y} - y_{K+1})^2 - (\bar{y} - y_2)^2] \\ &= y_{K+1}^2 - y_2^2 + (1/K)[(y_2 - y_{K+1})(2\bar{y} - y_2 - y_{K+1})] \\ &= (y_{K+1} - y_2)[(y_{K+1} + y_2)(1 + \frac{1}{K}) - 2\bar{y}/K] \\ &= 2(y_{K+1} - y_2) \left[\frac{y_{K+1} + y_2}{2} - \frac{\bar{y}}{K(1 + 1/K)} \right] \end{aligned} \tag{17}$$

There are only two possibilities:

$$\text{Case (i): } \frac{y_{K+1} + y_2}{2} \geq \bar{y}/K(1 + \frac{1}{K})$$

$$\text{Case (ii): } \frac{y_{K+1} + y_2}{2} \leq \bar{y}/K(1 + \frac{1}{K}) \tag{18}$$

If Case (i) is true, then

$$(J_2 - J_{K+1}) \geq 0, \text{ since } (y_{K+1} - y_2) > 0$$

Suppose Case (ii) is true. We can derive the following expression as in (17) above.

$$J_2 - J_1 = 2(y_1 - y_2) \left[\frac{y_1 + y_2}{2} - \frac{\bar{y}}{K(1 + \frac{1}{K})} \right].$$

$y_2 - y_1 > 0$, by definition.

$$\frac{\bar{y}}{K(1 + \frac{1}{K})} - \frac{y_1 + y_2}{2} \geq \frac{y_{K+1} + y_2}{2} - \frac{y_1 + y_2}{2}, \text{ by (18)}$$

Hence $J_2 - J_1 > 0$.