A Robust Method for Multiple Linear Regression

Mainack Paul (MD2111)

May 13, 2023

1 Introduction

Linear regression models are preferred in most of statistical computing due to its simplicity and easily interpretable nature. If analysis of variance is included as a special case of linear regression, this fraction is increased. Currently regression models are being applied widely in Linguistics, Sociology and History. Almost every discipline is making use of regression analysis. Least-squares is an optimal procedure when the errors in a regression model have a Gaussian distribution but not optimal in many non-Gaussian situations with longer tails. Hence, some alternative to least squares is required. If the form of the model is not known exactly, then a least squares fit to a hypothesized, invalid model may obscure the inappropriateness of that model. This inappropriateness may be revealed in certain plots of residuals. A robust fit may leave several residuals much larger, more clearly indicating that something is wrong. Procedures have been developed and will be described below which are resistant to gross deviations of a small number of points and relatively efficient over a broad range of distributions. If the data is Gaussian they will yield, with high probability, results very similar to those of a least squares analysis.

2 Results on Estimates of Location

New estimates of location were studied which had high efficiency for the Gaussian distribution and strong robustness under extreme departures from normality. These estimates may be usefully extended to regression situations. An estimate $\hat{\mu}$ of location may be defined for a set of numbers $x_1, ..., x_n$, as a solution to the equation

$$\sum \varphi\left(\frac{x_i - \hat{\mu}}{s}\right) = 0 \qquad (1)$$

where $s(x)$ is an estimate of spread. Such an estimate is called an M-estimate

The form of the function φ and the definition of the scale parameter s determine the properties of $\hat{\mu}$. Andrcws proposed solving for $\hat{\mu}$ using φ defined by

$$\varphi = \begin{cases} sin\left(\frac{z}{c}\right) & |z| < c\pi \\ 0 & |z| \geq c\pi \end{cases} \qquad (2)$$

where $s = median\{|x_i - median\{x_i\}|\}$.

3 Extension to the Regression Problem

The M-estimates for location are defined to be solutions of the equation (1). This is equivalent to finding a local maximum of the function $\sum \psi\left(\frac{x_i-\mu}{s}\right)$where $\varphi(z) = -\frac{d}{dz}\psi(z)$. In this second form, it can be extended to regression models since $x_i - \mu$ may be considered as a residual, r_i , and s as a scale statistic. The estimate is defined as the values of parameters for which

$$\sum \psi\left(\frac{r_i}{s}\right), \qquad (3)$$

a function of the corresponding residuals, attains a local maximum.

Consider the model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k + \sigma e_i = x_i^{'}\beta + \sigma e_i$$

where β is a vector of unknown parameters, a row vector of independent variables, σ is an unknown scale parameter and e_i is a residual.

Given any k-vector b, the residuals

$$r_i(b) = y_i - x_i^{'}b,$$

can be formed.

A robust scale estimate can be defined by

$$s(b) = median\{|r_i(b)|\}.$$

The parameters β may be estimated by the location of a local maximum of the function $\sum \psi\left(\frac{r_i(b)}{s(b)}\right)$ where $-\psi$ is the integral of (2) given by

$$\psi(z) = \begin{cases} 1 + cos\left(\frac{z}{c}\right) c & |z| < c\pi \\ 0 & |z| \geq c\pi \end{cases} \qquad (4)$$

The particular local maximum found by an iterative optimization program will depend on the starting value b_0, and on the numerical maximization procedure used.

4 Robust Procedure

First we select a starting value (b_0). Let us take the least square estimate as the starting value. This starting value is further iterated to improve the efficiency of the procedure which will be described below. This can be done by maximizing the function

$$\sum_j \psi\left(\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right) \qquad (5)$$

with respect to b_i where $s(b_i) = median\{|r_j(b_i)|\}$starting with $b_0$and $c = 1.5$ is used.

The function to be optimized (5) is a sum of cosines. Since most of the arguments of the cosines are small, the sum is nearly quadratic.

This follows from noting that the maximum of (5) satisfies the system of k equations $(l = 1, ...k)$

$$\sum_{j=1}^{n} x_{jl}\psi'\left(\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right) = 0$$

which may be rewritten as

$$\sum x_{jl} w_j r_j(b^{(i)}) = 0$$

where $w_i = \frac{\psi'\left(\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right)}{r_j(b^{(i)})}$.

Hence,

$$w_i = \begin{cases} \frac{sin\left(\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right)}{r_j(b^{(i)})} & \left|\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right| \leq c\pi \\ 0 & \left|\frac{r_j(b^{(i)})}{s(b^{(i-1)})}\right| > c\pi \end{cases}$$

Thus, the estimate can be easily calculated by

i). selecting an initial estimate $b^{(0)}$,

ii). using the estimate to find the residuals $r(b^{(0)})$, scale estimate $s^{(0)}$and weights $w^{(0)}$,

iii). solving the least-squares equations associated with the model

$$y_i = x_i^{'}\beta + noise$$

with weights w_j for the next estimate in the iteration. If a weighted least-squares program is not available, the system

$$w_j y_i = w_j x_j b_i + noise$$

can be solved using ordinary least-squares.

5 Data Study

Table 1: Data from Operation of A Plant for the Oxidation of Ammonia to Nitric Acid

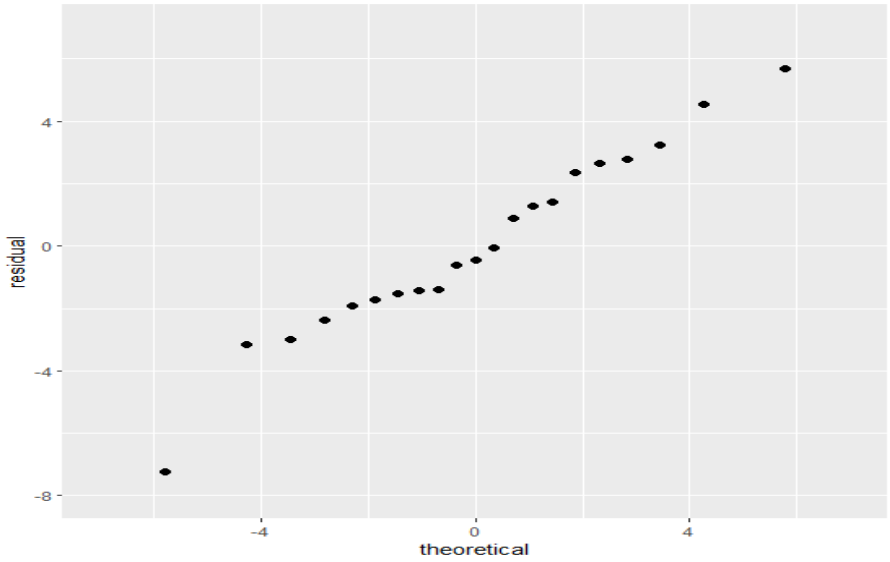
Observation Number	Stack Loss y	Air Flow x_1	Cooling Water Inlet Temperature x_2	Acid Concentration x_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

At first, we fit a standard least-squares linear regression model and also have plotted a normal probability plot for the residuals. The fit came out to be

$$E(y) = -39.9 + 0.72x_1 + 1.30x_2 - 0.15x_3$$

and the normal probability plot is as follows:

Figure 1: Probability plot of residuals from Least Square Fit of x_1, x_2, x_3

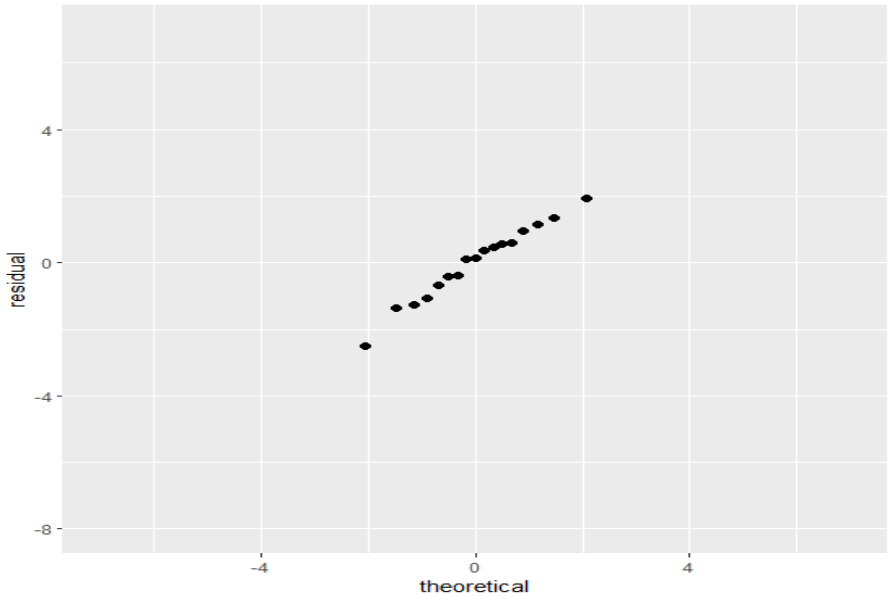


It is apparent that obseravation 21(bottom left corner) has an abnormally large residual. This observation has altered the coefficients of the fitted model considerably. Apart from this, observations 1,3 and 4 also has a large residual compared to the other observations. So, we set this 4 points aside and fit again a standard least-squares linear regression model along with a probability plot. The fit came out to be

$$E(y) = -37.6 + 0.80x_1 + 0.58x_2 - 0.07x_3$$

and the normal probability plot is as follows:

Figure 2: Probability plot of residuals from Least Square Fit of x_1, x_2, x_3 after removing 4 points

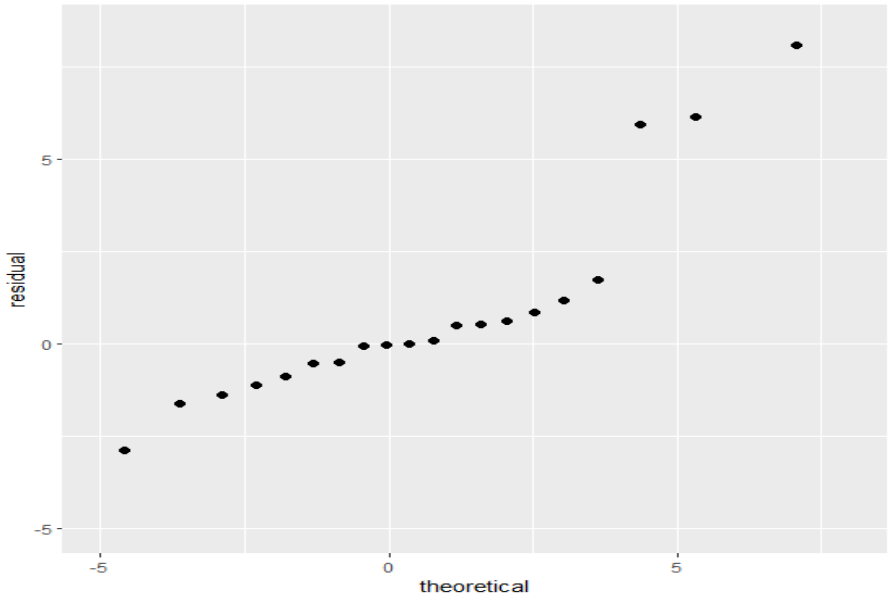


Fit (3) is a robust fit with c = 1.5. The fit came out to be

$$E(y) = -37.2 + 0.82x_1 + 0.52x_2 - 0.07x_3$$

The probability plot of residuals from this fit, Figure 3, identifies the 4 points.

Figure 3: Probability plot of residuals from Robust Fit of x_1, x_2, x_3

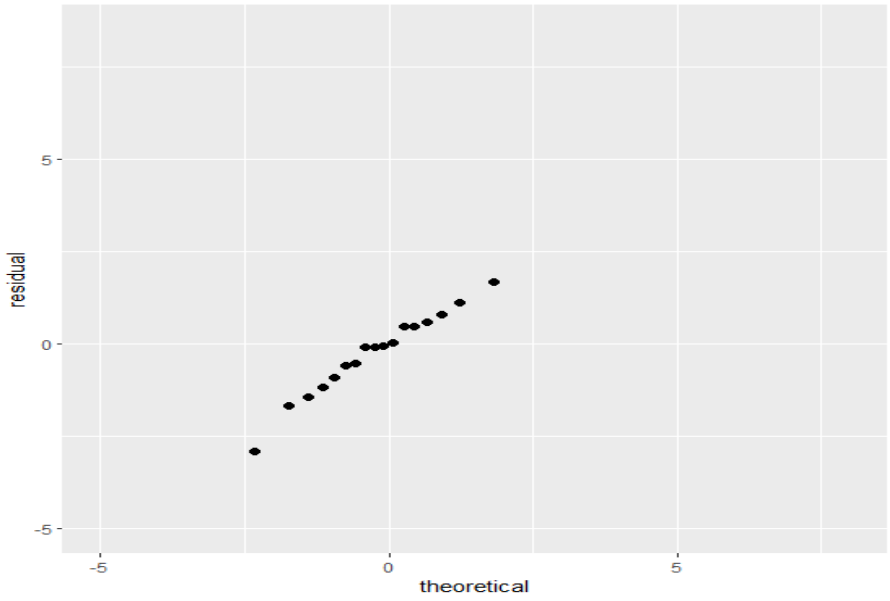


Fit (4) is the same fitting procedure applied to the data with the 4 points removed. The fit came out to be

$$E(y) = -37.2 + 0.82x_1 + 0.52x_2 - 0.07x_3$$

Note that the fit is unaffected by the 4 points. The robust fitting procedure (3) has immediately and routinely led to the identification of 4 questionable points. The fit is independent of these points. The probability plot for fit 4 is as follows:

Figure 4: Probability plot of residuals from Robust Fit of x_1, x_2, x_3 after removing 4 points



6 Conclusion

In this paper, a method for estimation in robust regression has been developed but it requires a safe initial fit. This procedure is iterative and is insensitive to moderate numbers of extreme observations with the result that these may be readily detected by examining residuals. However, the principal advantage lies in the detections of observations to be studied further.