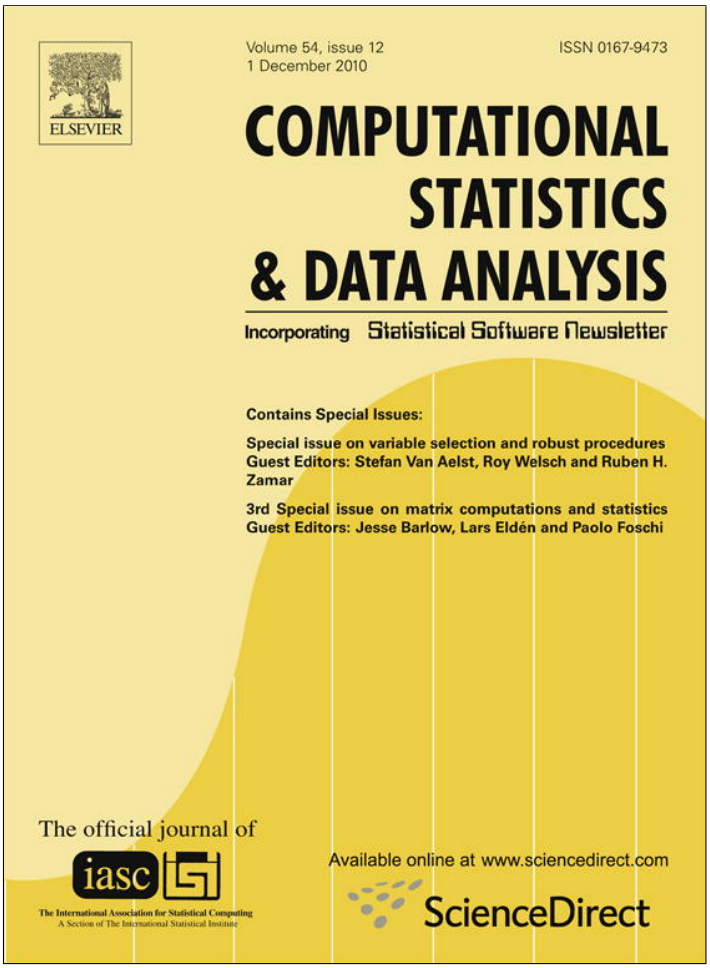

Computational Statistics and Data Analysis 54 (2010) 3168–3173

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Correcting MM estimates for "fat" data sets

Ricardo A. Maronna [a,*], Victor J. Yohai [b]

[a] Department of Mathematics, School of Exact Sciences, Universidad Nacional de La Plata and C.I.C.B.A., Argentina
[b] Department of Mathematics, School of Exact and Natural Sciences, Universidad de Buenos Aires and CONICET, Argentina

ARTICLE INFO

*Article history:*
Received 23 December 2008
Received in revised form 9 September 2009
Accepted 10 September 2009
Available online 12 September 2009

ABSTRACT

Regression MM estimates require the estimation of the error scale, and the determination of a constant that controls the efficiency. These two steps are based on the asymptotic results that are derived assuming that the number of predictors $p$ remains fixed while the number of observations $n$ tends to infinity, which means assuming that the ratio $p/n$ is "small". However, many high-dimensional data sets have a "large" value of $p/n$ (say, $\geq$0.2). It is shown that the standard asymptotic results do not hold if $p/n$ is large; namely that (a) the estimated scale underestimates the true error scale, and (b) that even if the scale is correctly estimated, the actual efficiency can be much lower than the nominal one. To overcome these drawbacks simple corrections for the scale and for the efficiency controlling constant are proposed, and it is demonstrated that these corrections improve on the estimate's performance under both normal and contaminated data.



## 1. Introduction

An important issue in robust estimation is the balance between robustness (as measured by the contamination bias) and efficiency at a central model. Regression MM estimates require the estimation of the error scale, and the determination of a constant that controls the efficiency. The pertinent methodology is based on the asymptotic results that are derived assuming that the number of predictors $p$ remains fixed while the number of observations $n$ tends to infinity, which in practical terms means that they will hold approximately if $p/n$ is sufficiently small. However, many modern data sets are "fat"in the sense that $p/n$ is "large"(say, $\geq$0.2). In these cases the standard asymptotic theory does not adequately take into account the loss of residual degrees of freedom due to parameter estimation. We shall demonstrate that the pertinent asymptotic results do not hold if $p/n$ is large; namely that the estimated scale underestimates the true error scale, and even if the scale is correctly estimated, the actual efficiency can be much lower than the nominal one. To overcome these drawbacks we shall propose simple corrections for the scale and for the efficiency controlling constant, and we shall demonstrate that these corrections do not affect the estimate's robustness.

An asymptotic analysis of these situations would require the study of the asymptotic distribution of the *residuals* from robust estimates when both $p$ and $n$ tend to infinity with $p/n$ remaining constant. Chen and Lockhart (2001) study the least squares residuals when $p^3 \log^2 p/n \to 0$, a situation too restricted for our purposes. Mammen (1996) and Portnoy (1986) derive important expansions for the asymptotic distribution of residuals when $p$ tends to infinity with $n$, assuming $p^2/n \to \infty$ and $p^2/n \to c$ respectively. Their results imply that the residuals' distribution may be very different from the error distribution. Despite the importance of these results, they cannot be used to describe the situations we face. For this reason our approach is based on heuristics and simulations. Rousseeuw and Leroy (1987, p. 44) propose an empirically derived correction factor for the residual MAD of the Least Median of Squares estimate, of the form $1 + 5/(n - p)$. Another possible approach, not explored here, to correct the asymptotic standard deviations of the robust regression estimates is the use of techniques of "finite sample asymptotics". A review of this approach can be found in Ronchetti (1990).

* Corresponding address: Departamento de Matemáticas, Facultad de Ciencias Exactas, C.C. 172, La Plata 1900, Argentina.
*E-mail address:* rmaronna@retina.ar (R.A. Maronna).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.09.015

Section 2 gives the main definitions, Section 3 describes the correction for the scale, Section 4 deals with the efficiency, and contains the conclusions. The Appendix contain the proofs.

## 2. S and MM estimates

We consider robust estimation in the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n \tag{1}$$

with $\mathbf{x}_i$, $\boldsymbol{\beta} \in R^p$ and $\{e_i\}$ i.i.d. and $e_i$ independent of $\mathbf{x}_i$. An M estimate of scale (an *M-scale* for short) of the data vector $\mathbf{r} = (r_1, \ldots, r_n)$ is defined as the solution $S = S(\mathbf{r})$ of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{S}\right) = \delta, \tag{2}$$

where $\rho$ is a bounded $\rho$-function in the sense of Maronna et al. (2006), i.e., $\rho(r)$ is a nondecreasing function of $|r|$ which is increasing for $\rho(r) < \sup \rho = 1$; here $\delta \in (0, 1)$ controls the breakdown point (BP) of the estimate.

A regression S estimate (Rousseeuw and Yohai, 1984) is defined as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\mathbf{r}(\boldsymbol{\beta})), \tag{3}$$

where $\mathbf{r}(\boldsymbol{\beta})$ is the residual vector with elements $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ and $S$ is an M-scale. It is known that for data in general position the maximum BP is attained by putting in (2)

$$\delta = \delta_{\text{opt}} =: 0.5\left(1 - \frac{p}{n}\right), \tag{4}$$

which yields BP=$\delta_{\text{opt}}$; see Maronna et al. (2006) for details. A popular choice for $\rho$ is the bisquare function $\rho_{\text{bis}}(r) = \min\{1, 1 - (1 - r^2)^3\}$.

S estimates are known to have a low efficiency under normal errors. MM estimates (Yohai, 1987) allow to control the efficiency while conserving a high BP. Let $\widehat{\boldsymbol{\beta}}_0$ be an initial estimate with high BP but possibly low efficiency (typically an S estimate) with residual vector $\mathbf{r}_0$. Let $\rho$ be a bounded $\rho$-function. Define the scale $\widehat{\sigma}$ by

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_{0i}}{h_0(\delta)\,\widehat{\sigma}}\right) = \delta, \tag{5}$$

where $h_0$, defined by

$$\mathrm{E}\rho\left(\frac{z}{h_0(\delta)}\right) = \delta; \quad z \sim N(0, 1) \tag{6}$$

makes $\widehat{\sigma}$ consistent at the normal model.

Then the MM estimate $\widehat{\boldsymbol{\beta}}_{\text{MM}}$ is defined as a local minimum of

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{h_1\widehat{\sigma}}\right),$$

where the minimum is computed iteratively starting from $\widehat{\boldsymbol{\beta}}_0$, and $h_1 > h_0$ is chosen in order to attain a given asymptotic efficiency at the normal model. It is shown that $\widehat{\boldsymbol{\beta}}_{\text{MM}}$ has the BP of $\widehat{\boldsymbol{\beta}}_0$ and the given normal efficiency. We compute the MM-estimator using the iterative reweighted least squares algorithm starting from an S-estimator; this procedure yields a local minimum close to the starting estimator. As demonstrated in Maronna et al. (2006), the resulting estimator has the desired efficiency, and its maximum bias is lower than that of the estimator which yields the *global* minimum of the loss function.

## 3. Correcting the scale

It would be ideal (but impossible) to base the MM estimate on the actual scale of the errors $S(\mathbf{e})$, while we actually have the (possibly biased) residual scale $\widehat{\sigma}_r = S\left(\mathbf{r}\left(\widehat{\boldsymbol{\beta}}_0\right)\right)$. To compare both we performed a simulation of model (1) with both $x_{ij}$ and $e_i$ i.i.d. standard normal, $n = 50$ and different values of $p$. Given the equivariance of the estimates, we can always take $\boldsymbol{\beta} = 0$ without loss of generality. For each case we generated 1000 samples and computed the bisquare S estimate and

$$q = \text{median}\left(\frac{S(\mathbf{e})}{\widehat{\sigma}_r}\right), \qquad M = \text{MAD}\left(\frac{S(\mathbf{e})}{\widehat{\sigma}_r}\right). \tag{7}$$

Table 1 gives the values of $q$ and the "coefficient of variation" $M/q$. It is seen that the underestimation can be serious. Unfortunately, the expansions given in the references cited in the Introduction do not allow us a theoretical understanding of this phenomenon. Since the ratios $M/q$ are low, we may consider $q$ to be a representative value. If we knew $q$ we could correct $\widehat{\sigma}_r$ by multiplying it by $q$. We propose two approaches to estimate $q$: one based on a Taylor expansion of $S$ and the other on an empirical fit of $q$.

**Table 1**
Median $q$ and "coefficient of variation" of scale ratios.

| $p$ | 5 | 10 | 15 | 25 |
|---|---|---|---|---|
| $q$ | 1.18 | 1.41 | 1.78 | 2.31 |
| $M/q$ | 0.06 | 0.08 | 0.09 | 0.14 |

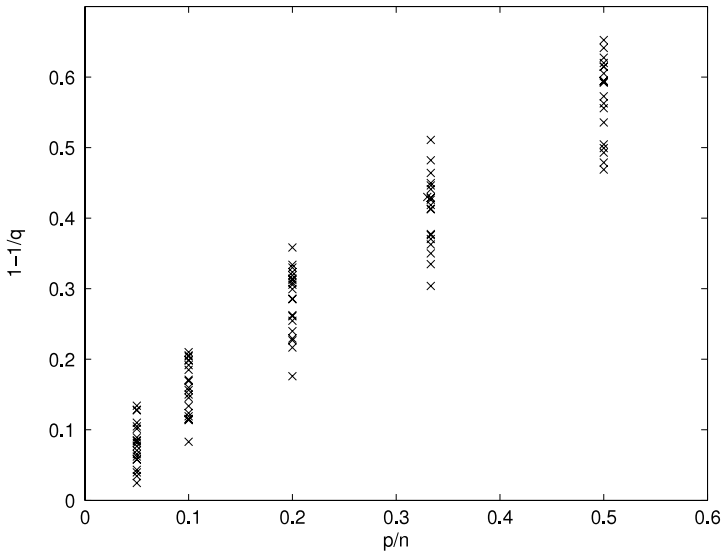


**Fig. 1.** $1 - 1/q(n, p, F)$ vs. $p/n$ for all configurations.

For the Taylor approach, the results in the Appendix suggest the estimate of $q$

$$\widehat{q}_{\mathrm{T}} = 1 + \frac{p}{2n}\frac{\widehat{a}}{\widehat{b}\widehat{c}} \tag{8}$$

with

$$\widehat{a} = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{r_i}{\widehat{\sigma}_r}\right)^2, \qquad \widehat{b} = \frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{r_i}{\widehat{\sigma}_r}\right), \qquad \widehat{c} = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{r_i}{\widehat{\sigma}_r}\right)\frac{r_i}{\widehat{\sigma}_r}, \tag{9}$$

where $\psi = \rho'$. Note that $\widehat{a}$, $\widehat{b}$ and $\widehat{c}$ depend on $h_0$.

The empirical approach is based on the fact that for the classical estimate $S(\mathbf{r}) = \sqrt{n^{-1}\sum_{i=1}^{n} r_i^2}$ we have

$$\sqrt{\frac{\mathrm{ES}(\mathbf{e})^2}{\mathrm{ES}(\mathbf{r})^2}} = \frac{1}{\sqrt{1 - p/n}} \approx \frac{1}{1 - 0.5p/n}.$$

This fact suggests to estimate $q$ by an expression of the form $1/\sqrt{1 - Kp/n}$ or the form $1/(1 - Kp/n)$ where $K$ is some constant. We deal first with the second form. For this purpose we undertook a simulation study. For each configuration we generated 1000 samples from model (1) with $x_{ij}$ and $e_i$ i.i.d. with the same distribution $F$. Five distributions were employed representing different degrees of heavy-tailedness: Normal, Student with 1, 3 and 5 degrees of freedom, and the standard normal distribution truncated at $\pm 1.65$. The values of $n$ were 25, 50, 100 and 400, and for each $n$ the values of $p$ were $[an]$ with $a$ =1/10, 1/5, 1/3 and 1/2. For each configuration the median ratio $q = q(n, p, F)$ was computed corresponding to the S estimate with bisquare $\rho$ and $\delta = \delta_{\mathrm{opt}}$ in (4). For given $n$ and $p$, the results were remarkably stable on $F$.

The relationship $q \approx 1/(1 - Kp/n)$ (resp. $1/\sqrt{1 - Kp/n}$) implies an approximately proportionality between $p/n$ and $1 - 1/q$ (resp. $1 - 1/q^2$). Fig. 1 supports the first relationship, while the corresponding plot for the second one does not support it. For this reason we fitted our "data set"with an expression of the form $q(n, p, F) \approx 1/(1 - Kp/n)$. More precisely, $K$ was determined by

$$\sum\left(1 - \frac{1}{q(n, p, F)} - K\frac{p}{n}\right)^2 = \min,$$

where the sum runs over all configurations of $(n, p, F)$.

The results were encouraging, but the quality of the fit worsened for small $n$. This fact suggested adding a "second order" term, and therefore we fit an expression of the form

$$\widehat{q}_{\mathrm{E}} = \frac{1}{1 - (k_1 + k_2/n)p/n}, \tag{10}$$

**Table 2**
Relative errors with maximum absolute values.

| $p/n$ | 0.05 | 0.10 | 0.20 | 0.33 | 0.5 |
|---|---|---|---|---|---|
| $\widehat{q}_T$ | 0.05 | 0.06 | −0.17 | −0.31 | −0.46 |
| $\widehat{q}_E$ | 0.04 | 0.08 | 0.14 | 0.15 | −0.35 |

**Table 3**
Efficiencies of MM estimate with nominal efficiency 0.85 for corrected and uncorrected scales, with $n = 50$.

| Scale | $p$ | | | |
|---|---|---|---|---|
| | 5 | 10 | 15 | 25 |
| Uncorrected | 0.75 | 0.55 | 0.49 | 0.52 |
| Corrected with $\widehat{q}_T$ | 0.86 | 0.76 | 0.69 | 0.59 |
| Corrected with $\widehat{q}_E$ | 0.82 | 0.76 | 0.74 | 0.73 |

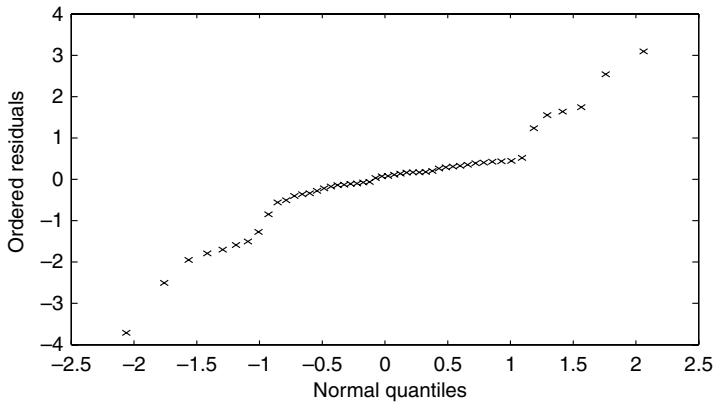


**Fig. 2.** QQ plot of residuals from S estimate of normal sample with $n = 50$ and $p = 10$.

which yielded $k_1 = 1.29$ and $k_2 = -6.02$. As to the other fit based on the form $1/\sqrt{1 - Kp/n}$, it turned out to be very unsatisfactory, as was to be expected, and was therefore dropped.

For both $\widehat{q}_T$ and $\widehat{q}_E$ we computed the relative errors $(\widehat{q} - q)/q$ with $q$ defined in (7). Table 2 gives the relative errors with maximum (over $n$, $p$ and $F$) absolute values for each value of $p/n$. It is seen that $\widehat{q}_T$ and $\widehat{q}_E$ have similar behaviors for $p/n \leq 0.2$, with the latter being better for larger values.

The correction (9) can be applied to any estimate using smooth bounded $\rho$-functions; for the correction (10) the required constants $k_1$ and $k_2$ must be recomputed for each $\rho$.

## 4. Correcting the efficiency

We then proceeded to assess the normal efficiency of the MM estimate using standard values of $h_1$. For this purpose we performed another simulation study with the normal model with $\boldsymbol{\beta} = 0$, which entails no loss of generality due to the estimates' equivariance. The MM estimate was computed using as initial $\widehat{\boldsymbol{\beta}}_0$ the bisquare S estimate with maximal breakdown point, and $h_1 = 3.44$ which under the standard theory ensures 85% normal asymptotic efficiency; and employing both the uncorrected and corrected scales. The values of $n$ and $p$ were like in the former study. The criterion was the mean squares error (MSE) defined as the Monte Carlo average of $\left\|\widehat{\boldsymbol{\beta}}\right\|^2$. Since robust estimates may have heavy-tailed distributions even under normal data, we also used a 10% upper trimmed mean. Both trimmed and non-trimmed means yielded qualitatively similar results, but we consider the latter more representative since the former was in some cases influenced by a few atypical values. The efficiency was therefore computed as the ratio of the (trimmed) MSEs of $\widehat{\boldsymbol{\beta}}_{MM}$ and of the least squares estimate. Table 3 reports the efficiencies for $n = 50$, the other cases being similar.

It is seen that using the uncorrected scale entails a serious loss of efficiency. Using the corrected ones brings a clear improvement, with $\widehat{q}_E$ better than $\widehat{q}_T$ for larger $p/n$. However, the efficiency remains lower than the nominal one. A clue to the reasons of this inefficiency may be given by Fig. 2, which is the normal Q–Q plot of the residuals from the S estimate applied to a standard normal sample with $n = 50$ and $p = 10$.

It is seen that the tails are far from normal. This implies that using the standard value of $h_1$ will result on a higher proportion of observations being downweighted than is convenient for the desired efficiency. This heavy-tailedness of the residuals might be hinted from the discussion in Section 3 of Mammen (1996).

The slope of the center part is about 0.58, which is near $\text{median}_i\,(|r_i|)\,/0.675 = 0.53$. This means that the underestimation of the scale is not a feature of the M-scale, but is due to the majority of residuals being "smaller"than standard normal.

**Table 4**
Corrected tuning constants for the bisquare family.

| $p/n$ | <0.1 | 0.1 | 0.2 | 0.33 |
|---|---|---|---|---|
| $h_1$ | 3.5 | 3.7 | 4.0 | 4.2 |

**Table 5**
Maximum MSEs of estimates.

| $p/n$ | $n = 50$ | | | $n = 200$ | | |
|---|---|---|---|---|---|---|
| | S-E | $h_1 = 3.44$ | $h_1 = 4.0$ | S-E | $h_1 = 3.44$ | $h_1 = 4.0$ |
| 0.1 | 0.94 | 0.66 | 0.65 | 1.14 | 0.59 | 0.59 |
| 0.2 | 1.78 | 0.98 | 0.90 | 2.10 | 1.11 | 1.06 |
| 0.3 | 2.79 | 1.44 | 1.34 | 9.78 | 5.36 | 4.57 |

It follows that attaining the desired efficiency requires a larger $h_1$ than the one given by standard asymptotic theory. We therefore performed further simulations of the normal model using different values of $h_1$. The results were rather stable for different $n$ and $p$. The suggested values to be used are given in Table 4.

Using a larger $h_1$ does not affect the BP, but it may be suspected that it will increase the contamination bias. To assess this effect further simulations were performed for contaminated normal data with $n = 50$ and 100 and $p/n = 0.1$, 0.2 and 0.3. In each configuration 10% of the observations were replaced as follows, $\mathbf{x}_i$ by $(k_{\text{lev}}, 0, \ldots, 0)$ and $y_i$ by $k_{\text{lev}}k_{\text{slo}}$, where $k_{\text{lev}}$ and $k_{\text{slo}}$ regulate the leverage and slope of the contamination. Both trimmed and non-trimmed MSEs were computed. For the reasons given above, we prefer the latter. The trimmed MSE was computed for the S estimate and for the MM estimates using the scale corrected with $\widehat{q}_{\text{E}}$, and the constants $h_1$ =3.44 and $h_1 = 4$. In all cases $h_1 = 4$ yielded a smaller MSE than $h_1 = 3.44$. As a synthesis, Table 5 shows the maximum over $k_{\text{slo}}$ of the MSEs for $n = 50$ and 200, and $k_{\text{lev}} = 5$. The results for $k_{\text{lev}} = 10$ were similar. Using the non-trimmed MSE yielded qualitatively similar results, but they were in some cases unduly influenced by a few large atypical values. We see that the two versions of MM are better than the S estimate, and that $h_1 = 4$ is better than 3.44. The likely reason is that when $p/n$ is not small, the decrease in variability more than compensates the increase in bias. Comparing Tables 3 and 5 suggests that increasing $h_1$ from 3.44 to 4 increases the efficiency without loss of robustness.

## 5. Conclusions

It has been demonstrated that unless the ratio $p/n$ is small, the efficiency of an MM regression estimate can be much lower than the nominal one for two reasons: (1) the usual estimate of the error scale is affected by an important downwards bias, and (2) the tuning constant calibrated according to asymptotic theory is not large enough. For problem (1) two methods for correcting the scale are proposed: one based on a Taylor expansion, and the other on an empirical fit. For problem (2) larger values of the tuning constants are proposed; and it is shown that their use does not decrease the robustness of the regression estimate. According to the results of the simulations in the preceding sections, it is recommended to use either the Taylor- or the Empirical-based scale correction when $p/n \leq 0.1$, and to use the Empirical one otherwise; and to increase the tuning constant of the bisquare function to 4 when $p/n > 0.1$.

## Acknowledgements

This research was partially supported by Grants X-018 from Universidad de Buenos Aires, PID 5505 from CONICET and PICTs 21407 and 00899 from ANPCYT, Argentina.

## Appendix. Justification of (8)

Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is the true parameter vector and $\widehat{\boldsymbol{\beta}}_0$ is the S estimate. A Taylor expansion yields

$$S(\mathbf{e}) = S(\mathbf{r}(\boldsymbol{\beta})) \approx S\left(\mathbf{r}\left(\widehat{\boldsymbol{\beta}}_0\right)\right) - \boldsymbol{\Delta}'\mathbf{d}\left(\widehat{\boldsymbol{\beta}}_0\right) + \frac{1}{2}\boldsymbol{\Delta}'\mathbf{D}\left(\widehat{\boldsymbol{\beta}}_0\right)\boldsymbol{\Delta}, \tag{11}$$

where $\mathbf{d}(\boldsymbol{\beta}) = \partial S(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\mathbf{D}(\boldsymbol{\beta}) = \partial^2 S(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2$. It follows from (3) that $\mathbf{d}\left(\widehat{\boldsymbol{\beta}}_0\right) = \mathbf{0}$ and

$$\mathbf{D}\left(\widehat{\boldsymbol{\beta}}_0\right) = -\frac{1}{\widehat{\sigma}_r\widehat{t}_3}\mathbf{A},$$

with $\widehat{t}_3$ defined in (9) and $\mathbf{A} = n^{-1}\sum_{i=1}^{n}\psi'(r_i/\widehat{\sigma}_r)\mathbf{x}_i\mathbf{x}_i'$. Call $\sigma_\infty$ the asymptotic value of $\widehat{\sigma}_r$, and put

$$a = \mathrm{E}\psi\left(\frac{e}{\sigma_\infty}\right)^2, \qquad b = \mathrm{E}\psi'\left(\frac{e}{\sigma_\infty}\right), \qquad c = \mathrm{E}\psi'\left(\frac{e}{\sigma_\infty}\right)\frac{e}{\sigma_\infty}.$$

When $n \to \infty$, $\mathbf{D}\left(\widehat{\boldsymbol{\beta}}_0\right)$ tends to $(b/\sigma_\infty c)\,\mathbf{X}'\mathbf{X}$, and the distribution of $\boldsymbol{\Delta}$ is approximately

$$\mathrm{N}_p\left(\mathbf{0}, \frac{\sigma_\infty^2 a}{nb}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right),$$

where $\mathbf{X}$ is the matrix having the $\mathbf{x}_i$'s as rows. Therefore $\boldsymbol{\Delta}'\mathbf{D}\left(\widehat{\boldsymbol{\beta}}_0\right)\boldsymbol{\Delta}$ is approximately distributed as $(\sigma_\infty a/nbc)$ times a $\chi^2$ variable with $p$ degrees of freedom which has mean $p$. This suggests from (11)

$$S(\mathbf{e}) \approx \widehat{\sigma}_r + \widehat{\sigma}_r \frac{pa}{2nbc},$$

which is approximated by (8).

## References


Chen, G., Lockhart, R.A., 2001. Weak convergence of the empirical process of residuals in linear models with many parameters. Ann. Statist. 29, 748–762.

Mammen, E., 1996. Empirical process of residuals for high-dimensional linear models. Ann. Statist. 24, 307–335.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, Chichister.

Portnoy, S., 1986. Asymptotic behavior of the empiric distribution of $M$-estimated residuals from a regression model with many parameters. Ann. Statist. 14, 1152–1170.

Ronchetti, E., 1990. Small sample asymptotics: A review with applications to robust statistics. Comput. Statist. Data Anal. 10, 207–223.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J, Yohai, V.J., 1984. Robust regression by means of S-estimators. In: Franke, J., Härdle, W., Martin, R.D. (Eds.), Robust and Nonlinear Time Series. In: Lectures Notes in Statistics, vol. 26. Springer Verlag, New York, pp. 256–272.

Yohai, V.J., 1987. High breakdown-point and high efficiency estimates for regression. Ann. Statist. 15, 642–665.