A Review of Robust Regression

Chun Yu\textsuperscript{1}, Weixin Yao\textsuperscript{1}, and Xue Bai\textsuperscript{1}

\textsuperscript{1}Department of Statistics,
Kansas State University, Manhattan, Kansas, USA 66506-0802.

Abstract

Ordinary Least-squares (OLS) estimators for a linear model are very sensitive to unusual values in the design space or outliers among y values. Even one single atypical value may have a large effect on the estimate. For the purposes of obtaining the estimates that are not influenced by outliers and have high efficiencies relative to OLS estimates under the assumption of normally distributed errors, several robust techniques are introduced in this paper. Some of them are resistant to outliers, that is, their breakdown points are near 0.5, but have low relative efficiencies. Some have high efficiencies but have low breakdown points. Finally, the robust estimators which achieve both high breakdown points and high efficiencies are recommended in this paper.

Key words: Robust; Regression.
1 Introduction

Suppose the observations are the i.i.d. (independent and identical) (p+1) dimensional random vectors \((x_i, y_i), i = 1, \ldots, n\). In order to understand how the \(y_i\) values are related to the \(x_i\) values, we consider the linear model

\[ Y_i = x'_i \beta + \varepsilon_i, \quad (1.1) \]

where \(\beta\) is an unknown \(p \times 1\) vector, and the \(\varepsilon'_i\)s are i.i.d. and independent of \(x_i\) for each \(i\). \(E(\varepsilon_i) = 0\). The Ordinary Least Square (OLS) Estimates are achieved by minimizing the sum of squared residuals

\[ \sum_{i=1}^{n} r_i (\hat{\beta})^2 = \text{min}. \quad (1.2) \]

where the residuals \(r_i(\cdot)\) equal \(y_i - x'_i \hat{\beta}\). Under the assumption that the errors in the regression model have normal distribution with mean 0 and variance \(\sigma^2\), the OLS Estimates are the most efficient unbiased estimates of \(\beta\). However, if the errors are not normally distributed, i.e., there are unusual observations, a single outlier can spoil the result of OLS estimates completely. If the assumption of normal errors is violated, we should consider the alternatives to answer the question for the truth. Robust methods for regression can be used for this purpose. The criteria for our estimators should be: (1) robust estimators work quite well on the non-normal distributed errors, i.e., the robust estimators are not influenced by the outliers; (2) the robust estimators are not that much worse than the OLS estimator when the errors are normally distributed.

Several robust approaches for regression were introduced and discussed in this paper. M-estimates (Huber 1981) are solutions of the normal equation with appropriate weight functions. they are resistant to unusual y observations, but sensitive to high leverage points on x, hence the breakdown point of an MM-estimate is 0 when the num-
ber of observations becomes large. Least Median of Squares (LMS) estimates (Siegel 1982) minimize the median of squared residuals. Least Trimmed Squares (LTS) estimates (Rousseeuw 1983) minimize the trimmed sum of squared residuals. S-estimates (Rousseeuw and Yohai 1984) minimize the variance of the residuals; they have high breakdown point but have low efficiencies. Generalized S-estimates (GS-estimates) (Croux et al. 1994) maintain the high breakdown point as S-estimates and improve the efficiencies a little. MM-estimates (Yohai 1987) are obtained by a iteration procedure with a robust estimator as initial estimate and a robust estimate of scale, they simultaneously attain high breakdown point and efficiencies. Mallows Generalized M-estimates (Mallows 1975) and Schweppe Generalized M-estimates (Handschin et al. 1975) down-weight the high leverage points on x but cannot distinguish "good" and "bad" leverage points, thus resulting in a loss of efficiencies. These two estimators have low breakdown points. SIS Generalized M-estimates (Coakley and Hettmansperger 1993) overcome the problems of Schweppe Generalized M-estimates and are calculated in one step. They have both high breakdown points and high efficiencies. R-estimates (Jaeckel 1972) minimize the sum of score of the ranked residuals. They have relatively high efficiencies but their breakdown points are as low as those of OLS estimates. Robust and Efficient Weighted Least Squares Estimator (REWLSE) (Gervini and Yohai 2002) are weighted Least-squares estimates with a pair of initial robust estimate of regression and scale. they can achieve both high breakdown points and full efficiencies. Robust regression based on regularization of case-specific parameters (Lee, MacEachern, and Jung 2007) are obtained by minimizing a robust loss function associated with regularization coefficients constraining the coefficients of the explanatory variables and the indicator variable. We described the advantages and disadvantages of these robust approaches by comparing the robustness measuring properties and the relative efficiencies to the OLS estimator when the the errors are normal distributed.
The rest of the paper is organized as follows. Measuring robustness are presented in Section 2. Section 3 introduces various robust techniques. We compare the advantages and disadvantages of the various robust techniques in Section 4.

2 Measuring Robustness

2.1 Breakdown Point

In this paper, finite sample breakdown point is used and defined as follow: Let $Z_i = (x_i, y_i)$. Given any sample $Z = (z_i, \ldots, z_n)$, denote $T(Z)$ the estimate of the parameter $\beta$. Let $Z'$ be the corrupted sample where any $m$ of the original points of $Z$ are replaced by arbitrary bad data. Then the finite sample breakdown point $\delta^*$ is defined as

$$
\delta^* (Z, T) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Z'} \| T(Z') - T(Z) \| = \infty \right\}.
$$

(2.1)

where $\| \cdot \|$ is Euclidean norm.

2.2 Influence Function

The influence function (IF) of an estimator (Maronna, Martin and Yahai 2006) is an approximation to the behavior of $\hat{\theta}_\infty$ when the sample contains a small fraction $\varepsilon$ of identical outliers, $x_0$. It is defined as

$$
IF_\theta(x_0, F) = \lim_{\varepsilon \downarrow 0} \frac{\hat{\theta}_\infty((1 - \varepsilon)F + \varepsilon \delta_{x_0}) - \hat{\theta}_\infty(F)}{\varepsilon}
$$

(2.2)

$$
= \frac{\partial}{\partial \varepsilon} \hat{\theta}_\infty((1 - \varepsilon)F + \varepsilon \delta_{x_0})|_{\varepsilon=0}
$$

(2.3)

where $\delta_{x_0}$ is the point-mass at $x_0$ and " $\downarrow$ " stands for "limit from the right". If there are $p$ unknown parameters, then $\hat{\theta}_\infty$ is a $p$-dimensional vector and so is its IF. The
quantity \( \hat{\theta}_\infty ((1-\varepsilon)F+\varepsilon\delta_{x_0}) \) is the asymptotic value of the estimate when the underlying distribution is \( F \) and a fraction \( \varepsilon \) of outliers is equal to \( x_0 \).

### 3 Robust Regression Techniques

#### 3.1 M-Estimates

In order to find a balance between the efficiency and the resistance of the unusual \( y \) observations, Huber (1964) proposed M-estimates. For a linear model, suppose the observed responses \( y_i \) are independent but not identically distributed, and have density functions

\[
f_i(Y_i) = \frac{1}{\sigma} f \left( \frac{Y_i - x_i'\beta}{\sigma} \right), \quad (3.1)
\]

The log likelihood is given by

\[
l(\beta) = -n\log\sigma + \sum_{i=1}^{n} \log f \left( \frac{Y_i - x_i'\beta}{\sigma} \right). \quad (3.2)
\]

M-estimates on regression are defined as solutions \( \hat{\beta} \) to

\[
\sum_{i=1}^{n} \rho \left( \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) = \text{min}, \quad (3.3)
\]

Where \( \hat{\sigma} \) is an error scale estimate. Differentiating (3.3) with respect to \( \beta \) yields the normal equation

\[
\sum_{i=1}^{n} \psi \left( \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) x_i = 0, \quad (3.4)
\]

Where \( \psi = \rho' \).

In particular, if \( \rho = -\log f(x) \), the solution of the normal equation is the MLE of \( \beta \).

If \( \rho = \frac{1}{2}x^2 \), then the solution of the normal equation is the OLS estimate. OLS
estimates are very sensitive to outliers. Rousseeuw and Yohai (1984) indicates that OLS estimates have a breakdown point of \( BP = \frac{1}{n} \), which tends to zero when the sample size \( n \) is getting large. Therefore, one single unusual observation can cause OLS estimates break down.

If \( \rho = |x| \), then Least Absolute Deviation (LAD) Estimates are achieved by minimizing the sum of the absolute values of the residuals

\[
\sum_{i=1}^{n} |r_i(\hat{\beta})| = \text{min}. \tag{3.5}
\]

The LAD is also called \( L_1 \) estimate due to the \( L_1 \) norm is used, which is the regression equivalent of the median. Although LAD is more resistant than OLS to unusual \( y \) values, it is sensitive to high leverage points with \( x_i's \), and thus has a breakdown point of \( BP = \frac{1}{n} \to 0 \) (Rousseeuw and Yohai 1984). Moreover, LAD estimates have a low efficiency of 0.64 when the errors are normal distributed.

For M-estimates, \( \psi \) function in the normal equation is replaced by appropriate weights, such as Huber or biweight, that decrease as the size of the residual increases

\[
\sum_{i=1}^{n} w_i \left( \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) x_i = 0. \tag{3.6}
\]

In most situations considered, \( \hat{\sigma} \) is computed previously, but it can also be computed simultaneously through a scale M-estimating equation. An iteratively reweighted least squares (IRLS) procedure is employed to find M-estimates for regression. Similar to LAD estimates, M-estimates have a \( BP = \frac{1}{n} \to 0 \) due to lack of immunity of high leverage points with \( x_i's \) (Rousseeuw and Yohai 1984). Under the Gauss-Markov assumption, however, M-estimates have a efficiency of 0.95 relative to OLS estimators.
### 3.2 LMS Estimates

The LMS estimates (Siegel 1982) are found by minimizing the median of the squared residuals

\[
\text{Med} \left( r_i \left( \hat{\beta} \right)^2 \right) = \min
\]

The advantage of LMS estimators is that it possess a high breakdown point of BP near 0.5; while the disadvantage is that it has at best a relative efficiency of 0.37 when the assumption of normal errors is met (see Rousseeuw and Croux 1993). Moreover, LMS estimates do not have a well-defined influence function because of its convergence rate of \( n^{-\frac{1}{3}} \) (Rousseeuw 1982). Despite these limitations, LMS estimates can be used as initial estimates of the residuals when calculating the much more efficient MM-estimators.

### 3.3 LTS Estimates

The LTS estimates (Rousseeuw 1983) minimize

\[
\sum_{i=1}^{q} r_{(i)} \left( \hat{\beta} \right)^2 = \min
\]

Where \( q = \left[ n (1 - \alpha) + 1 \right] \), and \( \alpha \) is the proportion of trimming that is chosen in practical issues. \( r_{(1)} \left( \hat{\beta} \right)^2 \leq \cdots \leq r_{(q)} \left( \hat{\beta} \right)^2 \) are ordered squared residuals. Using \( q = \left( \frac{n}{2} \right) +1 \) ensures that the estimator has a breakdown point of BP = 0.5, and convergence rate of \( n^{-\frac{1}{2}} \) (Rousseeuw 1983). Although highly resistant to outliers, LTS suffers badly in terms of very low efficiency, which is about 0.08, relative to OLS estimators (Stromberg, et al. 2000). The reason that LTS estimates call attentions to us is that it can be used as initial estimate of the residuals to calculate S1S GM-estimates, which is much more efficient and robust estimator.
3.4 S-Estimates

S-estimates (Rousseeuw and Yohai 1984) are defined by

\[ \hat{\sigma} \left( r_1 \left( \hat{\beta} \right), \cdots, r_n \left( \hat{\beta} \right) \right) = \min, \]  

(3.9)

where \( \hat{\sigma} \left( r_1 \left( \hat{\beta} \right), \cdots, r_n \left( \hat{\beta} \right) \right) \) is the scale M-estimate which is defined as the solution of

\[ \frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{r_i \left( \hat{\beta} \right)}{\hat{\sigma}} \right) = \delta, \]  

(3.10)

Differentiating the equation (3.10) yields

\[ \frac{1}{n} \sum_{i=1}^{n} \psi \left( \frac{r_i \left( \hat{\beta} \right)}{\hat{\sigma}} \right) = \delta, \]  

(3.11)

where \( \delta \) is taken to be \( E_{\Phi} \left[ \rho \left( r \right) \right] \), where \( \Phi \) is the standard normal distribution, and \( \psi \) function is replaced with an appropriate weight function. For the biweight scale, S-estimates with a breakdown point of \( \text{BP} = 0.5 \) has an asymptotic efficiency of 0.29 under the assumption of normally distributed errors (Maronna, Martin and Yahai 2006).

3.5 GS-Estimates

Croux et al. (1994) propose GS-estimates in an attempt to improve the low efficiency of S-estimators. GS-estimates are defined as

\[ \hat{\beta} = \arg \min_{\beta} S_n(\beta) \]  

(3.12)

where

\[ S_n(\beta) = \sup \left\{ S > 0; \frac{1}{2} \left( \sum_{i<j}^{n} \rho \left( \frac{r_i - r_j}{S} \right) \right) \geq k_{n,p} \right\} \]  

(3.13)
If $\rho(x) = I(|x| \geq 1)$ and $k_{n,p} = \binom{n}{2} - \binom{h_p}{2} + 1 / \binom{n}{2}$ with $h_p = \frac{n+p+1}{2}$, where $p$ is the number of regression parameters, then we have $S_n(\beta) = Q_n(\beta)$.

A special case of GS-estimator is the Least Quartile Difference (LQD) estimator, which is defined as

$$\hat{\beta} = \arg\min_{\beta} Q_n(r_1, \ldots, r_n)$$

(3.14)

where

$$Q_n = \{|r_i - r_j| ; i < j\}^{\binom{h_p}{2}}_{\binom{n}{2}}$$

(3.15)

This means that $Q_n$ is the $\binom{h_p}{2}$th order statistic among the $\binom{n}{2}$ elements of the set $\{|r_i - r_j| ; i < j\}$.

GS-estimates have a breakdown point as high as S-estimates. In order to improve the efficiency of GS-estimators, Ferretti et al. (1999) propose generalized $\tau$ estimates, which downweight the points with high leverage. This method achieves an efficiency of 0.75 under the Gauss-Markov assumption.

**3.6 MM-Estimates**

First proposed by Yohai (1987), MM-estimators have become increasingly popular and are perhaps now the most commonly employed robust regression technique. The MM-estimates can be found by a three-stage procedure. In the first stage, compute an initial consistent estimate $\hat{\beta}_0$ with high BP but possibly low normal efficiency. In the second stage, compute a robust M-estimate of scale $\hat{\sigma}$ of the residuals based on the initial estimate. In the third stage, find an M-estimate $\hat{\beta}$ using an IRLS procedure starting at $\hat{\beta}_0$.

In practice, LMS or S-estimation with Huber or bisquare weights is typically engaged as an initial estimate $\hat{\beta}_0$. let $\rho_0(r) = \rho_1 \left( \frac{r}{k_0} \right)$, $\rho(r) = \rho_1 \left( \frac{r}{k_1} \right)$, and each of the $\rho$-function
must be bounded. The scale estimate $\hat{\sigma}$ satisfies

$$
\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) = 0.5,
$$

(3.16)

Maronna, Martin and Yahai (2006) shows that the asymptotic BP of $\hat{\sigma}$ is 0.5. If the $\rho$-function is biweight, then $k_0 = 1.56$ ensures the estimator has the asymptotic BP = 0.5. Recall that an M-estimate requires the absolute minimum of

$$
L(\beta) = \sum_{i=1}^{n} \rho \left( \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right).
$$

(3.17)

Let $\rho$ satisfy $\rho \leq \rho_0$. Yohai (1987) shows that if $\hat{\beta}$ is such that $L(\hat{\beta}) \leq (\hat{\beta}_0)$, then $\hat{\beta}$’s BP is not less than that of $\hat{\beta}_0$. Furthermore, the breakdown point of the MM-estimate depends only on $k_0$ and the asymptotic variance of the MM-estimate depends only on $k_1$. We can choose $k_1$ in order to attain the desired normal efficiency without affecting its breakdown point. In order that $\rho \leq \rho_0$, we must have $k_1 \geq k_0$; the larger $k_1$, the higher the efficiency at the normal distribution.

Maronna, Martin and Yahai (2006) provides the values of $k_1$ corresponding to various efficiencies of the biweight $\rho$-function.

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>3.14</td>
<td>3.44</td>
<td>3.88</td>
<td>4.68</td>
</tr>
</tbody>
</table>

However, Yohai (1987) indicates that MM-estimates with larger values of $k_1$ are more sensitive to outliers than the estimates corresponding to a smaller value of $k_1$. Therefore, an MM-estimate with bisquare function and efficiency 0.85 ($k_1 = 3.44$) computed starting from a bisquare S-estimate is recommended (Maronna, Martin and Yahai 2006).
3.7 Generalized M-Estimates

The general GM class of estimators is defined by

\[
\sum_{i=1}^{n} w_i \psi \left( \frac{r_i(\hat{\beta})}{v_i \sigma} \right) x_i = 0, \tag{3.18}
\]

Where \( \psi \) is the score function (as in the case of M-estimation, this is the typically the Huber or biweight function), and the weight \( w_i \) and \( v_i \) initially depend on the model matrix \( X \) from an initial OLS regression fitted to the data but are updated iteratively.

3.7.1 Mallows GM-estimate

This GM-estimate proposed by Mallows (Mallows 1975 or Krasker and Welsch 1982) depends only on the \( w_i \) weights with \( w_i = \sqrt{1 - h_i} \), that is, \( v_i = 1 \) in equation (3.18). The weight \( w_i \) ensures that the observations with high leverage receive less weight than observations with small leverage. However, even ”good ” leverage points that fall in line with the pattern in the bulk of the data are down-weighted, resulting in a loss of efficiency.

3.7.2 Schweppe GM-estimate

Schweppe’s solution (introduced in Handschin et al. 1975) adjusts the leverage weights according to the size of the residual \( r_i \). In order to achieve this result, the \( w_i \) weights are defined in the same way as for Mallows, \( w_i = \sqrt{1 - h_i} \), but now \( v_i = w_i \) (see Chave and Thomson 2003). Carroll and Welsh (1988) suggested that the Schweppe estimator is not consistent when the errors are asymmetric.

The breakdown point for both Mallows and Schweppe GM-estimates never exceed \( 1/(p + 1) \), where \( p \) is the number of parameters estimated by the model.
3.7.3 S1S GM-estimate

In an attempt to overcome problems of Mallows and Schweppe’s solution GM-estimators, Coakley and Hettmansperger (1993) introduced Schweppe one-step estimate (S1S), which extends from the original Schweppe estimator. The advantage of this estimator over the original is that the leverage weights consider where observations fit with the bulk of the data. In other words, it considers whether the observations are "good" or "bad" leverage points, giving less weight to the latter. S1S estimator is defined as solutions to the equation (3.19), the resulting estimator is

\[
\hat{\beta} = \hat{\beta}_0 + \left[ \sum_{i=1}^{n} \psi' \left( \frac{r_i \left( \hat{\beta}_0 \right)}{\hat{\sigma} w_i} \right) x_i x'_i \right]^{-1} \times \sum_{i=1}^{n} \hat{\sigma} w_i \psi \left( \frac{r_i \left( \hat{\beta}_0 \right)}{\hat{\sigma} w_i} \right) x_i. \tag{3.19}
\]

The choices of \(\hat{\beta}_0, w_i, \psi\) function and \(\hat{\sigma}\) are chosen in practical issues.

The S1S estimator takes initial estimates of the residuals and the scale of the residuals from a regression with a high breakdown point rather than from an OLS regression, as is the case with the GM-estimators developed before it. Using Rousseeuw’s LTS estimator for the initial estimates gives a breakdown point of \(BP = 0.5\) and results in 0.95 efficiency relative to OLS estimators under the Gauss-Markov assumptions. The method is also different from the Mallows and Schweppe estimators in that once the initial estimates from the LTS regression are included, final M-estimates are calculated in a single step rather than iteratively.

3.8 R-Estimates

R-estimators (Jaeckel 1972) minimizes the sum of some score of the ranked residuals

\[
\sum_{i=1}^{n} a_n \left( R_i \right) r_i = \min, \tag{3.20}
\]
where $R_i$ represents the rank of the $i$th residual $r_i$, and $a_n(\cdot)$ is a monotone score function that satisfies

$$
\sum_{i=1}^{n} a_n(i) = 0. \tag{3.21}
$$

Many choices of the score function have been proposed. Although R-estimators are scale equivariant which is an advantage relative to M-estimators, they have some disadvantages. One problem is that the optimal choice of the score function is unclear. The second problem is that most of R-estimators have a breakdown point of $BP = 1/n \to 0$. The bounded influence R-estimator proposed by Naranjo and Hettmansperger (1994) has a fairly high efficiency (0.9 - 0.95) when the errors have normal distribution, however, it is proved that their breakdown point is not more than 0.2.

### 3.9 REWLS Estimator

REWLS estimator (Gervini and Yohai 2002) is much more attractive than other robust estimators due to its simultaneously attaining high breakdown point and efficiency under normal errors. With a pair of initial robust estimates of regression and scale, $T_{0n}$ and $S_n$ respectively, the standardized residuals are defined as

$$
r_i = \frac{y_i - x'_iT_{0n}}{S_n}
$$

A large value of $|r_i|$ would suggest that $(x_i, y_i)$ is an outlier. Rousseeuw and Leroy (1987) define

$$
w_i = \begin{cases} 
1 & \text{if } |r_i| < t_0 \\
0 & \text{if } |r_i| \geq t_0 
\end{cases} \tag{3.22}
$$

with $t_0 = 2.5$ and compute a weighted Least-squares estimator $T_{1n} = (X'WX)^{-1}X'WY$, where $W = diag(w_1, \cdots, w_n)$ and $Y = (y_1, \cdots, y_n)'$. Let the empirical distribution
function of the standardized absolute residual be

\[
F_n^+(t) = \frac{1}{n} \sum_{i=1}^{n} I(|r_i| \leq t).
\]

As a measure of proportion of outliers in the sample we then define

\[
d_n = \max_{i > i_0} \left\{ F^+ (|r|_{(i)}) - \frac{(i - 1)}{n} \right\}^+
\]

(3.23)

where \{\cdot\}^+ denotes positive part, \(F^+\) denotes the distribution of \(|X|\) when \(X \sim F\). \(|r|_{(1)} \leq \ldots \leq |r|_{(n)}\) are the order statistic of the standardized absolute residuals and \(i_0 = \max \{i : |r|_{(i)} < \eta\}\), where \(\eta\) is some large quantile of \(F^+\). Rousseeuw and Leroy choose a reasonable \(\eta = 2.5\). Thus those \([nd_n]\) observations with largest standardized absolute residuals are eliminated. The adaptive cut-off value is

\[
t_n = \min \left\{ t : F_n^+(t) \geq 1 - d_n \right\},
\]

(3.24)

that is \(t_n = |r|_{(i_n)}\) with \(i_n = n - [nd_n]\). Observe that \(i_n > i_0\) and \(t_n > \eta\). With this adaptive cut-off value, the weights are defined as \(w_i = w(|r|/t_n)\) and the REWLSE is

\[
T_{1n} = \begin{cases} 
(X'WX)^{-1}X'WY & \text{if } S_n > 0 \\
T_{0n} & \text{if } S_n = 0.
\end{cases}
\]

(3.25)

If the initial regression and scale estimators with \(BP = 0.5\) are chosen, the breakdown point of the REWLSE estimates is not less than 0.5. Furthermore, the REWLS estimates are asymptotically equivalent to the OLS estimates and hence asymptotically efficient when the errors are normally distributed.
3.10 Robust Regression based on regularization of case-specific parameters

Let $z_i$ be the indicator variable taking 1 for the $i$th observation and 0 otherwise, and $\gamma = (\gamma_1, \ldots, \gamma_n)^T$ be the coefficients of the case indicators. The robust estimate of $\beta$ is achieved by minimizing

$$L(\beta, \gamma) = \frac{1}{2} \{ Y - (X\beta + \gamma) \}^T \{ Y - (X\beta + \gamma) \} + \sum_{i=1}^{n} p_\lambda(|\gamma_i|),$$

where $p_\lambda(|\cdot|)$ is the penalty function which depends on the regularization parameter $\lambda$. The penalty function may take a $L_1$, $L_2$ or Smoothly Clipped Absolute Deviation (SCAD) penalty.

The minimizer $\hat{\gamma} = \text{sgn}(r)(|r| - \lambda)_+$, that is,

$$\hat{\gamma} = \begin{cases} 0 & \text{if } |r_i| \leq \lambda; \\
 y_i - x_i^T \hat{\beta} & \text{if } |r_i| > \lambda. \end{cases}$$

The adjusted residuals are defined as $r_i^* = y_i - x_i^T \hat{\beta} - \hat{\gamma}_i$. The $\gamma$-adjusted loss is equivalent to truncated square error loss which is $(y - x^T \beta)^2$ if $|y - x^T \beta| \leq \lambda$, and is $\lambda^2$ otherwise. Appropriate selection of $\lambda$ needs to be combined with the iterative algorithm.

4 Discussion

When used in combination with more resistant estimators, new estimators result that both highly resistant to outliers and highly efficient. M-estimation using the residuals from an initial highly resistant LTS fit, for example, leads to the S1S GM-estimator, which is highly resistant to both residual outliers and high leverage observations, and
### Table 1: Breakdown Points and Asymptotic Efficiencies of Various Regression Estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Breakdown Point</th>
<th>Asymptotic Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LMS</td>
<td>0.5</td>
<td>0.37</td>
</tr>
<tr>
<td>LTS</td>
<td>0.5</td>
<td>0.08</td>
</tr>
<tr>
<td>S-estimates</td>
<td>0.5</td>
<td>0.29</td>
</tr>
<tr>
<td>GS-estimates</td>
<td>0.5</td>
<td>0.75</td>
</tr>
<tr>
<td>MM-estimates</td>
<td>0.5</td>
<td>0.85</td>
</tr>
<tr>
<td>GM-estimates(S1S)</td>
<td>0.5</td>
<td>0.95</td>
</tr>
<tr>
<td>REWLS-estimates</td>
<td>0.5</td>
<td>1.00</td>
</tr>
<tr>
<td>Low BP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM-estimates(Mallows,Schwepppe)</td>
<td>1/(p + 1)</td>
<td>0.95</td>
</tr>
<tr>
<td>Bounded R-estimates</td>
<td>&lt; 0.2</td>
<td>0.90-0.95</td>
</tr>
<tr>
<td>M-estimates</td>
<td>1/n</td>
<td>0.95</td>
</tr>
<tr>
<td>LAD</td>
<td>1/n</td>
<td>0.64</td>
</tr>
<tr>
<td>OLS</td>
<td>1/n</td>
<td>1.00</td>
</tr>
</tbody>
</table>

yet maintains an efficiency of about 0.95 relative to the OLS estimator. Computing M-estimates from the residuals from LMS or S-estimation leads to the efficient and robust MM-estimators. Considering both high breakdown point and efficiency, S1S GM-estimates and MM-estimates with $k_1 = 3.44$ are preferred.

### References


Chave and Thomson 2003


Croux et al 1994
Ferretti et al 1999
Handschin et al 1975


Naranjo and Hettmenisperger 1994


Rousseeuw and Croux 1993
Rousseeuw and Yohai 1984


Stromberg et al 2000