Bounded Influence Regression using High Breakdown Scatter Matrices *

Christophe Croux,[1] Stefan Van Aelst,[2] and Catherine Dehon[3]

Abstract

In this paper we estimate the parameters of a regression model using S-estimators of multivariate location and scatter. The approach is proven to be Fisher-consistent, and the influence functions are derived. The corresponding asymptotic variances are obtained and it is shown how they can be estimated in practice. A comparison with other recently proposed robust regression estimators is made.

Keywords: Fisher-Consistency, Influence Function, Robust Regression, S-Estimators, Scatter Matrices.

*Short running head: Robust regression using scatter matrices

[1]Dept. of Applied Economics, KULeuven, Naamsestraat 69, B-3000 Leuven, Belgium.

[2]Dept. of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium. E-mail: Stefan.VanAelst@ua.ac.be

[3]ECARES, Université Libre de Bruxelles, CP-139, Av. F.D. Roosevelt 50, B-1050 Brussels, Belgium.

1 Introduction

Consider the classical regression model

$$y_i = \alpha + \beta^t u_i + \varepsilon_i,$$

$i = 1, \dots, n$ where the error terms $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. and independent of the p-dimensional carriers $u_1, \dots, u_n$. The least-squares (LS) estimators $\hat{\alpha}_{LS}$ and $\hat{\beta}_{LS}$ are defined as the minimizers of the sum of squared residuals

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta^t u_i)^2. \tag{1.1}$$

Since the least squares estimator is very sensitive to the presence of outliers, robust alternatives need to be looked for. Many of these robust regression methods consist of minimizing a robust loss function of the residuals, instead of a quadratic loss function. Main examples here are the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) estimator (Rousseeuw 1984), who can attain the maximum breakdown value. The breakdown value is the smallest fraction of data points that needs to be replaced to carry the estimator arbitrarily far away (for a formal definition, see Rousseeuw and Leroy 1987, page 117). Generalized S-estimators (Croux et al. 1994) and τ-estimators (Yohai and Zamar 1988) combine this high breakdown value with a high efficiency. However, their unbounded influence function is sometimes seen as a drawback.

Another way of robustifying LS consists of robustifying the first order conditions associated to the minimization of (1.1):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta^t u_i)u_i = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta^t u_i) = 0. \tag{1.2}$$

This lead to the construction of M and GM-estimators which are defined as solutions of robustified versions of the first order equations (1.2). Unfortunately, they have no high breakdown point (see e.g. Hampel et al. 1986). To remediate this, MM- (Yohai 1987) and one step GM-estimators (Simpson et al. 1992, Coakley and Hettmansperger 1993, Simpson and Yohai 1998) were proposed. One-step GM-estimators combine a high breakdown point with a bounded influence function. Other high breakdown, bounded influence estimators have recently been proposed by Ferretti et al. (1999) and Chang et al. (1999). The latter paper proposes a class of high breakdown rank regression estimators, extending the class of Generalized R-estimators of Naranjo and Hettmansperger (1994).

In case of the LS-estimator, the solution of the normal equations (1.2) is explicit:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}(u_i - \hat{\mu}_u)(u_i - \hat{\mu}_u)^t\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_y)(u_i - \hat{\mu}_u)\right) \text{ and } \hat{\alpha} = \hat{\mu}_y - \hat{\mu}_u^t\hat{\beta},$$

with $\hat{\mu}_u = \frac{1}{n}\sum_{i=1}^{n} u_i$ and $\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i$. With the use of the empirical covariance matrices $\hat{S}_{uy}$ and $\hat{S}_{uu}$, we may rewrite the above equation as

$$\hat{\beta} = \hat{S}_{uu}^{-1}\hat{S}_{uy} \quad \text{and} \quad \hat{\alpha} = \hat{\mu}_y - \hat{\mu}_u^t\hat{\beta}. \tag{1.3}$$

The idea now is not to robustify the normal equations, but its solutions. Therefore, we will replace the empirical mean and covariance in (1.3) by robust equivalents. Many proposals for robust location and covariance matrices have been made, such as M-estimators (Maronna 1976), the Stahel-Donoho estimator (Stahel 1981), the Minimum Volume Ellipsoid and Minimum Covariance Determinant estimator (Rousseeuw 1984,1985) and S-estimators (Davies 1987, Rousseeuw and Leroy 1987).

Maronna and Morgenthaler (1986) used multivariate M-estimators to insert into (1.3) and showed that the resulting estimators have all the desired equivariance properties. They also gave an expression for the influence function of this approach based on M-estimators, but only for a regression without intercept. Visuri et al. (2001) used rank based covariance matrices and derived results at elliptically symmetric models. In this paper, S-estimators of location and scatter will be used. For a finite sample $\{z_1, \ldots, z_n\} \subset I\!R^{p+1}$ the S-estimates are defined as the couple $(\hat{\mu}, \hat{S})$ which minimizes $\det(S)$ under the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\sqrt{(z_i - \mu)^t S^{-1}(z_i - \mu)}\right) \leq b, \tag{1.4}$$

over all $\mu \in I\!R^{p+1}$ and $S \in \text{PDS}(p+1)$, where $\text{PDS}(p+1)$ is the set of all positive definite symmetric matrices of size $p+1$. The function ρ is chosen by the statistician and b is a selected constant.

At first sight, this approach based on robust covariance matrix estimators seems to be restricted to regression models with elliptically symmetric carrier distribution. Indeed, consistency of robust covariance matrices is always proven under this symmetry assumption. Results on the behavior of $\hat{\alpha}$ and $\hat{\beta}$ using any affine equivariant location and covariance matrix estimators in (1.3) have been obtained by Croux et al. (2001), but only under the severe assumption of elliptical symmetry. In practice, this restriction cannot be retained. Even an ordinary quadratic regression would then not be covered by the hypothesis of the model. An important contribution of this paper is therefore that we prove the approach based on S-estimators to be valid for arbitrary carrier distributions. Moreover, we will show that the resulting estimator is a high breakdown bounded influence estimator, combining good efficiency and robustness properties.

In Section 2 we define the regression functionals based on robust S-estimators of location and scatter. The corresponding influence function is computed in Section 3 and shown to be bounded for ρ functions with bounded derivative. An estimator for the covariance matrix of the estimator is presented in Section 4, where we also construct studentized residuals. Section 5 presents a simulation study to compare the estimator with other existing high breakdown, bounded influence estimators. We compare bias, mean squared error and stability of the estimators under investigation. Section 6 gives a real data example, while Section 7 concludes. The Appendix contains all the proofs.

2 The Functional

The functional form of S-estimators of multivariate location and scatter is defined as follows. Let K be an arbitrary $(p+1)$-dimensional distribution. For our purposes, K represents the joint distribution of the carriers and response variable. Define now the S-estimator $(M(K), S(K))$ as the couple (M, S) which minimizes $\det(S)$ under the constraint

$$\int \rho(\sqrt{(z-M)^t S^{-1}(z-M)}\,)\, dK(z) \leq b, \tag{2.1}$$

over all $M \in I\!R^{p+1}$ and $S \in \mathrm{PDS}(p+1)$. The function ρ satisfies

(R) ρ is even, continuous, non decreasing on $[0, +\infty[$ with $\rho(0) = 0$, and almost everywhere twice differentiable with derivative $\rho' = \psi$.

The constant b satisfies $0 < b < \rho(\infty)$ and determines the breakdown point of the estimator which equals $\min(\frac{b}{\rho(\infty)}, 1 - \frac{b}{\rho(\infty)})$ (see Lopuhaä 1989). The vector $M(K)$ corresponds with the location S-estimator, and $S(K)$ with the scatter S-estimator.

Let u contain the first p components of the variable $z \sim K$ and y the last component, so $z = (u^t, y)^t$. The variable y will be the dependent variable of the regression equation while u contains the explanatory variables. Split up the vector $M(K)$ and matrix $S(K)$ accordingly, that is

$$M(K) = \begin{pmatrix} M_u(K) \\ M_y(K) \end{pmatrix} \quad \text{and} \quad S(K) = \begin{pmatrix} S_{uu}(K) & S_{uy}(K) \\ S_{yu}(K) & S_{yy}(K) \end{pmatrix}.$$

The functional of interest is now defined as $T(K) = (a(K), b(K)^t)^t$ where

$$b(K) = S_{uu}^{-1}(K) S_{uy}(K) \tag{2.2}$$

is called the regression slope functional and

$$a(K) = M_y(K) - b(K)^t M_u(K) \tag{2.3}$$

the intercept functional. One has that $T = (a, b^t)^t$ is regression, scale, and carrier equivariant (Maronna and Morgenthaler, 1986). This means that, using the notation $a(K) = a(u, y)$ and $b(K) = b(u, y)$ whenever $(u^t, y)^t \sim K$,

$$\begin{aligned} a(Au, cy + l^t u + d) &= c\, a(u, y) + d \\ b(Au, cy + l^t u + d) &= (A^{-1})^t (c\, b(u, y) + l) \end{aligned}$$

for every $l \in I\!R^p$, $c, d \in I\!R$ and nonsingular $(p \times p)$ matrix A.

Consider now the regression model

$$y = \alpha + u^t \beta + \varepsilon$$

where u is the vector of random explicative variables and ε the error term. We suppose that ε is independent of u and that $F(t) = P(\varepsilon \leq t)$ satisfies

(F) The distribution F has a strictly positive, symmetric and unimodal density f.

We denote by H the distribution of $z = (u^t, y)^t$, and call it the model distribution. A regularity condition (to avoid degenerate situations) on the distribution G of the carriers u is that

(G) $P_G(u^t \gamma = \delta) < 1 - \frac{b}{\rho(\infty)}$ for all $\gamma \in I\!R^p \setminus \{0\}$ and $\delta \in I\!R$.

When using a 50% breakdown estimator, this means that not more than half of the mass of the distribution of G is lying on the same hyperplane. For unbounded ρ functions it implies that the distribution of G is not completely concentrated on a hyperplane. A first result is that the functionals a and b defined in (2.3) and (2.2) are Fisher-consistent for the intercept and slope parameters α and β.

Theorem 1. *The functional T is Fisher-consistent for the parameter $\theta = (\alpha, \beta^t)^t$ at the model distribution H, that is*

$$T(H) = \begin{pmatrix} a(H) \\ b(H) \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \theta.$$

Note that no symmetry conditions for the distribution of the carriers have been required. We do require that the distribution of the errors is symmetric which is often assumed in robust linear regression. One of the exceptions is Chang et al. (1999), whose method is based on differences of residuals. Now instead of computing the S-estimator from the sample $\{z_1, \ldots, z_n\}$, we could also compute an S-estimator from the set of differences $\{z_i - z_j \mid 1 \leq i < j \leq n\}$. This corresponds with the Generalized S-estimators of Croux et al. (1994), and yields a method not requiring symmetry of the error terms. Since computing Generalized S-estimators of regression is much more time consuming, we will stick to the class of ordinary S-estimators and keep restriction (F).

3 Influence function

Before deriving the influence function we recall that S-estimators satisfy the following first-order conditions (Lopuhaä 1989):

$$\int w_1(d_K^2(z))(z - M(K))\, dK(z) = 0 \tag{3.1}$$

$$\int w_1(d_K^2(z))(z - M(K))(z - M(K))^t\, dK(z) = \int w_2(d_K^2(z))\, dK(z)\, S(K), \tag{3.2}$$

where the weight functions equal $w_1(t) = \psi(\sqrt{t})/\sqrt{t}$ and $w_2(t) = \frac{\psi(\sqrt{t})\sqrt{t} - \rho(\sqrt{t}) + b}{p+1}$, and $d_H^2(z) = (z - M(H))^t S(H)^{-1}(z - M(H))$ is a squared Mahalanobis distance. It will be shown that w_1 determines the form of the influence function.

The influence function of the functional T at the distribution H measures the effect on T of adding a small mass at $z = (u^t, y)^t$. If we denote the point mass at z by Δ_z and consider the contaminated distribution $H_{\varepsilon,z} = (1-\varepsilon)H + \varepsilon\Delta_z$ then the influence function is given by

$$IF(z; T, H) = \lim_{\varepsilon \downarrow 0} \frac{T(H_{\varepsilon,z}) - T(H)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(H_{\varepsilon,z})\big|_{\varepsilon=0}.$$

(See Hampel et al. 1986.) The next theorem gives an expression for the influence functions of the regression functional b and intercept functional a at a model distribution.

Theorem 2. *Let $y = \alpha + u^t\beta + \varepsilon$, where ε is independent of u, and $\varepsilon \sim F$ satisfies condition (F). Let H be the distribution of $z = (u^t, y)^t$, H_0 the distribution of $(u^t, \varepsilon)^t$ and denote $x = (1, u^t)^t$ and $\theta = (\alpha, \beta^t)^t$. Then the influence function of the functional T at the distribution H is given by*

$$\text{IF}(z; T, H) = C(H_0)^{-1} w_1(d_H^2(z))x(y - x^t\theta) \tag{3.3}$$

where

$$C(H_0) = \int w_1(d^2_{H_0}(z))xx^t\, dH_0(z) + \frac{2}{S_{yy}(H_0)} \int w_1'(d^2_{H_0}(z))y^2xx^t\, dH_0(z). \tag{3.4}$$

Moreover, if the score function $\Lambda_f(t) = -f'(t)/f(t)$ associated to the density f exists, then

$$C(H_0) = \int xx^t w_1(d^2_{H_0}(z))y\Lambda_f(y)\, dH_0(z). \tag{3.5}$$

From the above Theorem, it is seen that the influence function is bounded as soon as w_1 is redescending to zero, which is the case for bounded ψ-functions.

Remark: Let $z \sim H$ the model distribution and denote

$$A = \begin{pmatrix} I_p & 0 \\ -\beta^t & 1 \end{pmatrix} \qquad \text{and} \qquad c = \begin{pmatrix} 0 \\ -\alpha \end{pmatrix}.$$

Then $Az + c \sim H_0$ and by affine equivariance of the S-estimator

$$S(H_0) = AS(H)A^t = \begin{pmatrix} S_{uu}(H) & 0 \\ 0 & S_{yy}(H) - \beta^t S_{uy}(H) \end{pmatrix}.$$

The scale functional $\sigma^2_\varepsilon(H) := S_{yy}(H_0)$ equals therefore

$$\sigma_\varepsilon(H) = \sqrt{S_{yy}(H) - \beta^t S_{uu}(H)\beta}\,. \tag{3.6}$$

Since $\det(A) = 1$ we can rewrite (3.4) as

$$C(H_0) = \int w_1(d^2_H(z))xx^t\, dH(z) + \frac{2}{\sigma^2_\varepsilon(H)} \int w_1'(d^2_H(z))(y - x^t\theta)^2 xx^t\, dH(z), \tag{3.7}$$

which is an expression in terms of the observed distribution H. Equivalently,

$$C(H_0) = \int xx^t w_1(d^2_H(z))(y - x^t\theta)\Lambda_f(y - x^t\theta)\, dH(z). \tag{3.8}$$

4 Estimating the asymptotic variance

At the sample level, we estimate the parameters α, β by $\hat{\alpha} = a(H_n)$ and $\hat{\beta} = b(H_n)$, where H_n is the empirical distribution function of the data $z_i = (x_i^t, y_i)^t$ $(1 \leq i \leq n)$. With $\hat{\mu} = M(H_n)$ and $\hat{S} = S(H_n)$ we retrieve the estimators defined in the introduction (equations (1.3) and (1.4)). Now asymptotic expansions and the asymptotic normality property were obtained for S-estimators of location and scatter by (Lopuhaä 1989, Lopuhaä 1997) under very general regularity conditions. These conditions cover the case when the data z_i are generated from our model distribution H satisfying conditions (F) and (G). Since $\hat{\theta} = (\hat{\alpha}, \hat{\beta}^t)^t$ is a differentiable function of the elements of $(\hat{\mu}, \hat{S})$, the estimator $\hat{\theta}$ will also be asymptotically normal with corresponding asymptotic variance obtained from the influence function by means of

$$\mathrm{ASV}(T,H) = \int \mathrm{IF}(z;T,H)\,\mathrm{IF}(z;T,H)^t\,dH(z)$$

(see e.g. Barndorff-Nielsen and Cox 1989, page 47). Together with expression (3.3) this yields

$$\mathrm{ASV}(T,H) = C(H_0)^{-1}D(H_0)\,C(H_0)^{-1}$$

with

$$D(H_0) = \int w_1^2(d_{H_0}^2(z))y^2xx^t\,dH_0(z) = \int w_1^2(d_H^2(z))(y-x^t\theta)^2xx^t\,dH(z). \tag{4.1}$$

The covariance matrix of $\hat{\theta}$ is now estimated in a natural way by replacing H by H_n in the right hand side of expressions (3.7) and (4.1):

$$\widehat{\mathrm{Cov}(\hat{\theta})} = \frac{1}{n}\widehat{\mathrm{ASV}(T,H)} = \frac{1}{n}\widehat{C(H_0)}^{-1}\widehat{D(H_0)}\,\widehat{C(H_0)}^{-1} \tag{4.2}$$

with

$$\begin{aligned}\widehat{D(H_0)} &= \frac{1}{n}\sum_{i=1}^n w_1^2(d_i^2)r_i^2x_ix_i^t,\\ \widehat{C(H_0)} &= \frac{1}{n}\sum_{i=1}^n\{w_1(d_i^2)+\frac{2}{\hat{\sigma}_{\varepsilon,n}^2}\,w_1'(d_i^2)r_i^2\}x_ix_i^t,\end{aligned}$$

where $x_i = (1, u_i^t)^t$, $r_i = y_i - u_i^t\hat{\beta} - \hat{\alpha}$, $d_i = \sqrt{(z_i-\hat{\mu})^t\hat{S}^{-1}(z_i-\hat{\mu})}$ is the robust Mahalanobis distance of z_i (Rousseeuw and van Zomeren 1990), and

$$\hat{\sigma}_{\varepsilon,n} = \sqrt{\hat{S}_{yy} - \hat{\beta}^t\hat{S}_{uu}^{-1}\hat{\beta}}\,.$$

Alternatively $C(H_0)$ can be estimated, by using (3.8), as

$$\widehat{C(H_0)} = \frac{1}{n}\sum_{i=1}^n x_ix_i^tw_1(d_i^2)r_i\Lambda_{\hat{f}_n}(r_i), \tag{4.3}$$

which requires however a nonparametric estimate $\hat{f}_n$ of the density f. If f is specified to be $N(0,\sigma^2)$ then we have $\Lambda_f(t) = -\frac{d}{dt}\log f(t) = t/\sigma^2$. The parameter σ can be estimated from the residuals by a consistent scale estimator $\hat\sigma_n(r_1,\ldots,r_n)$. For Gaussian errors (4.3) then results in

$$\widehat{C(H_0)} = \frac{1}{n}\sum_{i=1}^n x_i x_i^t w_1(d_i^2)\frac{r_i^2}{\hat\sigma_n^2}$$

yielding

$$\widehat{\mathrm{Cov}(\hat\theta)} = \hat\sigma_n^4\left(\sum_{i=1}^n w_1(d_i^2)r_i^2 x_i x_i^t\right)^{-1}\left(\sum_{i=1}^n w_1^2(d_i^2)r_i^2 x_i x_i^t\right)\left(\sum_{i=1}^n w_1(d_i^2)r_i^2 x_i x_i^t\right)^{-1}. \tag{4.4}$$

If the function ρ becomes constant for values larger than a certain c^*, then the function w_1 is redescending to zero. It follows that in this case the estimators for $\mathrm{Cov}(\hat\theta)$ are robust since outliers are downweighted to zero in expressions (4.2) and (4.4).

Remark on studentized residuals: The obtained expressions for $\widehat{\mathrm{Cov}(\hat\theta)}$ are also useful for constructing studentized residuals. Using the asymptotic representation

$$\hat\theta = \theta + \frac{1}{n}\sum_{i=1}^n \mathrm{IF}(z_i;T,H) + o_p(n^{-1/2}) = \theta + (n\,C(H_0))^{-1}X^tWe + o_p(n^{-1/2})$$

where $X=(x_1,\ldots,x_n)^t$, $W = diag(w_1(d_H^2(z_1)),\ldots,w_1(d_H^2(z_n)))$ and $e=(\varepsilon_1,\ldots,\varepsilon_n)^t$, a first order approximation of the variance of the residuals can be obtained. Following the approach of McKean et al. (1990,1993) one gets

$$\mathrm{var}(r_i) \doteq \sigma^2\left(1 - 2w_1(d_H^2(z_i))x_i^t(n\,C(H_0))^{-1}x_i + x_i^t(n\,C(H_0))^{-1}X^tW^2X(n\,C(H_0))^{-1}x_i\right)$$

where σ^2 is the variance of the errors, which is supposed to be finite. An estimate for the variance of the residual $r_i = y_i - \hat\theta' x_i = y_i - \hat\beta' u_i - \hat\alpha$ is therefore given by

$$\widehat{\mathrm{var}(r_i)} = \hat\sigma_n^2\left(1 - 2w_1(d_i^2)x_i^t(n\,\widehat{C(H_0)})^{-1}x_i + x_i^t(n\,\widehat{C(H_0)})^{-1}X^t\widehat{W}^2X(n\,\widehat{C(H_0)})^{-1}x_i\right), \tag{4.5}$$

with $\widehat{W} = diag(w_1(d_1^2),\ldots,w_1(d_n^2))$, yielding as studentized residual

$$r_i^* = \frac{r_i}{\sqrt{\widehat{\mathrm{var}(r_i)}}} \tag{4.6}$$

These studentized residuals are adjusted for the variance of the errors and for location, as is the case for studentized residuals from the least squares estimator (see McKean et al 1993).

5 Simulation study

To study the finite-sample behaviour of the regression estimator based on the S-covariance matrix estimator, we compare it with other high breakdown, bounded influence estimators. We consider the one-step GM estimator of Simpson et al. (1992), which uses Mallows weights and Hampel's three-part redescending ψ function and which will be denoted as M1M. Another one-step GM estimator, proposed by Coakley and Hettmansperger (1993), uses Schweppe weights and the Huber ψ function (S1M). We also compare with the High Breakdown Rank (HBR) estimator of (Chang et al. 1999) which is a kind of weighted version of the Wilcoxon rank estimator. All these 3 estimators need an initial estimator, for which the Least Trimmed Squares (LTS) estimator was taken. For computing the weights to downweight leverage points, the Minimum Covariance Determinant (MCD) estimator was chosen. All other tuning constant where selected as suggested in the cited papers. Note that the breakdown point of the three above estimators is determined by the breakdown point of the initial LTS and MCD estimates, which we set equal to 50% (or 25%) such that all estimators under comparison will have the same breakdown point.

To compute the S-estimates of location and scatter we use the Tukey biweight ρ-function

$$\rho_c(t) = \min(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}, \frac{c^2}{6}). \tag{5.1}$$

The constant b determines the breakdown point: taking $b = \rho(\infty)/2$ yields the maximal breakdown point of 50%, while $b = \rho(\infty)/4$ gives a 25% breakdown point estimator. The resulting S covariance based regression estimator, abbreviated as S-CovReg, inherits this breakdown point. The choice of the tuning constant c is arbitrary in this regression setup, but it is customary to select it such that $E_H[\rho(d_H(z))] = b$ for $H = N(0, I_{p+1})$. The function ρ_c is bounded and sufficiently smooth, with an associated weight function w_1 being redescending. For computing the S-estimator of location and scatter, the fast and accurate algorithm of Ruppert (1992) has been used.

To compare the efficiency and robustness of the estimators we simulate under 5 different sampling schemes. In every situation, we generate $m = 1000$ samples of size $n = 100$ according to the regression model

$$y_i = \alpha + \beta_1 u_{i1} + \beta_2 u_{i2} + \beta_3 u_{i3} + \beta_4 u_{i4} + \varepsilon_i,$$

for $i = 1, \ldots, n$. The true regression parameters $\alpha, \beta_1, \ldots, \beta_4$ were set equal 0, without loss of generality due to regression equivariance. First we consider standard normal distributed data, where no outliers are present (NOR). For the second (VO10) and third (V020) sampling scheme contamination is inserted in the samples by generating 10%, respectively 20% of vertical outliers. These vertical outliers correspond to responses generated according to a normal $N(5\chi_{p+1,0.99}, 0.1)$, where $\chi_{p+1,0.99}$ stands for the square root of the 1% upper quantile of a chi-square distribution with $(p + 1)$ degrees of freedom. In the last situations bad leverage points are introduced by generating both the carriers and the response according to a $N(5\chi_{p+1,0.99}, 0.1)$. We consider 10% and 20% of bad leverage points, and name these sampling schemes (BL10) and (BL20).

We focus on the results for the slopes, which are of primary interest (in fact, the intercept can always be estimated afterwards by computing a univariate location estimate from the residuals $y_i - \beta^t u_i$). To compare the estimators we computed the Bias and Mean Squared Error (MSE) for every component $\hat{\beta}_1, \ldots, \hat{\beta}_4$ over all m simulations runs. These numbers are then summarized by $(\text{ave}_{j=1}^4 \text{Bias}(\hat{\beta}_j)^2)^{1/2}$ and $\text{ave}_{j=1}^4 \text{MSE}(\hat{\beta}_j)$, and reported as Bias and MSE in Table 1.

INSERT TABLE 1 ABOUT HERE

A first remark is the bad performance of the LTS estimator, witnessing its weak efficiency properties. However, the LTS is well suited as an initial estimator, as can be seen from the much lower MSE of the S1M, M1M and HBR estimator. Comparing the latter three estimators is difficult: HBR is more efficient than M1M and S1M if no outliers are present, but sacrifices bias in presence of bad leverage points. Let us now compare the estimator S-CovReg with its competitors. For the 50% breakdown point versions of the estimators, the S-covariance based estimate comes out as the best in this simulation study: for Normal data and for data with vertical outliers the S-CovReg estimator is comparable with M1M, S1M, and HBR but in situations with bad leverage points it outperforms the other estimators.

The same conclusion holds for the 25% breakdown point versions of the estimators. Only in the case of 20% outliers the S-CovReg estimator (with 25% breakdown point) does not perform well. However, 20% of outliers occur rarely in practice and if such high level of outliers can be expected choosing a 50% breakdown estimator is more appropriate.

In general we see that the S-CovReg estimator behaves very well compared to the other three high breakdown, bounded influence estimators. Unless the data are expected to contain enormous amounts of outliers, we recommend to use the 25% breakdown S-covariance estimator to obtain a better compromise between robustness and efficiency.

INSERT TABLE 2 ABOUT HERE

It has been shown that high breakdown estimators as Least Median of Squares are very sensitive to slight changes in the data (Hettmansperger and Sheather 1992, Sheather et al. 1997). High breakdown, bounded influence estimators are expected to be much more stable (see e.g. Simpson and Yohai 1998, Chang et al. 1999). To study the stability of the S-covariance estimator we use the simple linear regression model

$$y_i = \alpha + \beta u_i + \varepsilon_i, \qquad i = 1, \dots, 50 \tag{5.2}$$

where both the errors and predictors are i.i.d. standard Gaussian and $\alpha = \beta = 0$. We will slightly change a data point, and compute the effect of this change on the estimate. If the estimator is stable, small changes in the data should lead to small changes in the estimate. For every data point (u_i, y_i), four changes are considered: $u_i \pm 0.1$ and $u_i \pm 0.2$. The maximal effect of any of these 4×50 changes on the estimate of β is then recorded. Table 2 shows the maximal difference for the 25% breakdown point versions of the S-CovReg, M1M, S1M, HBR and LTS estimators over all 100 simulations. First of all, we see that the stability of all bounded influence estimators is much better than for the LTS estimator. Moreover, the S-CovReg and HBR estimators are seen to be much more stable than the one-step GM estimators. A theoretical justification of the stable behavior of S-CovReg can be found in the smooth and bounded character of its influence function (3.3).

Note that high breakdown estimates may have problems in detecting or fitting curvature. See e.g. Cook et al. (1992), McKean et al. (1993), and Chang et al. (1999) for simulations and examples that illustrate the problem. This implies that one has to be careful with the interpretation of residual plots from high breakdown fits. McKean et al. (1996) proposed diagnostics based on a highly efficient estimator and a high breakdown estimator that can expose discrepancies due to curvature in the data.

6 Example

As an example we consider the famous Hawkins-Bradu-Kass data (Hawkins et al. 1984), which is an artificial data set with $n = 75$ and $p = 3$. The first 14 observations are known to be outliers. We used the S-covariance based estimator with Tukey biweight function and 50% breakdown point, since the data set contains a huge amount (almost 20%) of outliers. In

Table 3 (a), we report the estimates obtained with the classical estimator, the S-covariance based estimator and the MM-estimator (Yohai and Zamar 1988) with 50% breakdown point. We have chosen to make a comparison with robust MM-estimators, since this is an established robust regression method with high breakdown point and good efficiency properties. It is standard implemented in S-plus and also reports standard errors and correlations between the estimates (as described in Yohai et al. 1991).

INSERT TABLE 3 ABOUT HERE

We see that the two robust methods give quite similar results, while the classical estimates are very different since they are highly influenced by the outliers. Note that none of the variables is declared as significant by the robust approach, while the second and third slope parameter are significantly different from zero according to the LS method. It is instructive to compare these results with those based on the clean data-set with the 14 artificial outliers deleted. From Table 3 (b), we notice that the results for the method based on the robust covariance matrix hardly change, neither for the estimates, nor for the covariance matrix of the estimates. The MM-estimator appears to be less stable for the correlations between the coefficients. Note that, on the basis of the clean data, LS finds none of the variables to be significant.

Several diagnostic plots can be produced. We will illustrate them for the estimator based on the robust S-covariance matrix defined above. In Figure 1a, the studentized residuals r_i^*, as given by (4.6), are represented versus their index. The scale of the errors $\hat{\sigma}_n$ in (4.5) was estimated by an A-estimator of scale (see Iglewicz 1982, page 417), which has the maximal breakdown point, a redescending influence function (like the regression and intercept estimators) and is standard implemented in S-plus. From Figure 1a we immediately observe that the first 10 observations are not following the linear relation imposed by the majority of the data.

INSERT FIGURE 1 ABOUT HERE

The robust distance d_i of an observation $z_i = (x_i, y_i)$ indicates how far the data point is from the bulk of the data cloud. In Figure 1b the robust distances d_i are compared with the constant c of (5.1). If d_i is bigger than this critical value then $w_1(d_i^2)$ will vanish, resulting in a zero influence on the estimator according to (3.3). We see that all 14 outliers were above this critical value, and therefore are completely downweighted.

To verify whether condition (F) on the residuals is reasonable, a diagnostic plot will be used (cfr. Figure 1c). The solid line is the kernel density estimate $\hat{f}_h(t)$ of the distribution of the residuals. To put emphasis on the central part of the data, the density estimate has been restricted to the interval $[-3\hat{\sigma}_n, 3\hat{\sigma}_n]$. The Gaussian kernel has been used and the bandwidth h was selected using maximum-likelihood cross-validation (see e.g. Härdle 1991, p. 93). Afterwards, a symmetric unimodal version of this density has been added to this plot. It has been constructed as follows: first we computed $\hat{f}_h^s(t_j) = (\hat{f}_h(t_j) + \hat{f}_h(-t_j))/2$ for a grid of equidistant positive points starting from zero. Then a classical monotonic regression algorithm (see e.g. Cox and Cox 1994, page 51) has been applied on the $\hat{f}_h^s(t_j)$ to obtain $\hat{f}_h^{sm}(t_j)$. Putting $\hat{f}_h^{sm}(-t_j) = \hat{f}_h^{sm}(t_j)$ and connecting all the obtained values results in the dashed line of Figure 1c, which is a symmetric and unimodal function. Note that the initial density estimate is reasonably close to the unimodal symmetric version, so it seems reasonable to assume that condition (F) is satisfied. Of course, more formal tests for unimodality and symmetry could be applied.

Finally, a classical QQ-plot is presented in Figure 1d. Once again, since we do not want the outliers to dominate this picture and make the interpretation harder, the plot is based on all residuals with absolute value smaller than $3\hat{\sigma}_n$. Figure 1d suggest that normality will not be rejected. This is confirmed by a Kolmogorov-Smirnoff test (P-value> 0.2).

Supposing normality of the error terms allows us to use formula (4.4) to estimate the covariance matrix of the estimator. Results are reported in Table 4. We see that estimates, standard errors, and correlations between the coefficients remain robust: the outcomes based on the whole data set and just on the clean data are not too different and close to the LS result computed from the data with the outliers deleted.

INSERT TABLE 4 ABOUT HERE

7 Conclusions

In this paper we discussed properties of regression estimators based on high breakdown S-estimators of location and scatter. We proved Fisher consistency of the method, without making the hypothesis of elliptical symmetry on the distribution of the explanatory variables. We derived the influence function, which appears to be bounded for the usual choices of ρ functions in robust statistics. Moreover, it can easily be shown that the resulting regression estimator inherits the breakdown point of the location/scatter S-estimator.

S-estimators of location and scatter have very attractive properties. It was shown that they are asymptotically normal (Davies 1987) with a quite high efficiency also in higher dimensions (Lopuhaä 1989, Croux and Haesbroeck 1999). At the same time they are extremely robust and have a smooth influence function. Moreover, there exist very fast algorithms to compute them (Ruppert 1992, Woodruff and Rocke 1994), even in high dimensions. They seem to be an excellent choice as robust covariance matrix estimators in multivariate analysis. In the context of principal components analysis, they have been successfully applied by Croux and Haesbroeck (2000) and in discriminant analysis by Croux and Dehon (2001).

Although many robust regression approaches have already been proposed in the literature, we think that the approach based on robust covariance matrices merits to be added to the list of the better robust regression estimators. Let us mention some important advantages. First of all, the estimator combines good robustness (high breakdown point and bounded influence function) and good efficiency properties. The simulation study has shown the good performance of the estimator in terms of efficiency and stability properties, in comparison with other recent high breakdown, bounded influence regression estimators. Also, a robust estimate for the covariance matrix of the estimator is available. This allows to construct reliable standard errors around their robust estimates, which is important in practice but often neglected in the robustness literature (among the exceptions are the bounded influence regression estimators proposed by Chang et al. 1999 and Ferretti et al. 1999). Moreover, the method is quite simple and easy to explain: once the S-estimator of scatter is computed, which can be done using the fast algorithm of Ruppert (1992), the estimators of regression are explicitly computable without extra work. Finally, a similar approach can be applied to more general regression models, like multivariate regression (Rousseeuw et al. 2000) and calibration models (Cheng and Van Ness 1997).

Acknowledgment: We wish to thank the referee for his remarks, which lead to a considerable improvement of the paper.

Appendix

To prove Theorem 1 we will use the following lemma.

Lemma 1. *If the function ρ satisfies condition (R) and the distribution F satisfies condition (F), then the function*

$$\lambda_{\sigma,c}(t) = \int \rho(\sqrt{(y-t)^2\sigma + c}\,)\, dF(y)$$

is symmetric and increasing on $[0, +\infty[$ for every $\sigma > 0$, $c \geq 0$. Moreover, for $c < c^* = \inf\{t > 0 \,|\, \rho(t) = \rho(\infty)\}$, $\lambda_{\sigma,c}(t)$ is strictly increasing on $[0, +\infty[$.

Proof of Lemma 1: The symmetry of $\lambda_{\sigma,c}$ follows from the symmetry of F:

$$\lambda_{\sigma,c}(-t) = \int \rho\left(((y+t)^2\sigma + c)^{1/2}\right)\, dF(y) = \int \rho\left(((-y+t)^2\sigma + c)^{1/2}\right)\, dF(y) = \lambda_{\sigma,c}(t).$$

Now $\lambda_{\sigma,c}$ has a positive derivative $\lambda'_{\sigma,c}(t)$ on $]0, +\infty[$ which can be seen as follows:

$$\begin{aligned}
\lambda'_{\sigma,c}(t) &= \frac{\partial}{\partial t} \int \rho\left(((y-t)^2\sigma + c)^{1/2}\right)\, dF(y) \\
&= -\sigma \int \frac{(y-t)}{((y-t)^2\sigma + c)^{1/2}}\, \psi\left(((y-t)^2\sigma + c)^{1/2}\right)\, f(y)\, dy \\
&= -\sigma \left\{ \int_{-\infty}^{t} \frac{(y-t)}{((y-t)^2\sigma + c)^{1/2}}\, \psi\left(((y-t)^2\sigma + c)^{1/2}\right)\, f(y)\, dy \right. \\
&\quad \left. + \int_{t}^{\infty} \frac{(y-t)}{((y-t)^2\sigma + c)^{1/2}}\, \psi\left(((y-t)^2\sigma + c)^{1/2}\right)\, f(y)\, dy \right\}.
\end{aligned}$$

By transforming the integration variables in these last two integrals, we obtain

$$\begin{aligned}
\lambda'_{\sigma,c}(t) &= -\sigma \left\{ -\int_0^{\infty} \frac{s}{s^2\sigma + c}\, \psi\left((s^2\sigma + c)^{1/2}\right)\, f(t-s)\, ds \right. \\
&\quad \left. + \int_0^{\infty} \frac{s}{s^2\sigma + c}\, \psi\left((s^2\sigma + c)^{1/2}\right)\, f(t+s)\, ds \right\} \\
&= \sigma \int_0^{\infty} \frac{s}{s^2\sigma + c}\, \psi\left((s^2\sigma + c)^{1/2}\right)\, [f(t-s) - f(t+s)]\, ds.
\end{aligned}$$

For every $s, t > 0$ we have $f(t-s) - f(t+s) > 0$ from the unimodality of F. Condition (R) ensures that $\psi\left((s^2\sigma + c)^{1/2}\right) \geq 0$ implying that $\lambda'_{\sigma,c}(t) \geq 0$. Moreover, if $c < c^*$, then $\{s > 0 \,|\, \psi\left((s^2\sigma + c)^{1/2}\right) > 0\}$ has a non-zero Lebesgue measure, so that in this case $\lambda'_{\sigma,c}(t) > 0$ for $t > 0$. □

Proof of Theorem 1: First of all, due to equivariance, we may suppose that $\alpha = 0$ and $\beta = 0$, so $y = \varepsilon$. Lopuhaä (1989) has shown that a solution $(M(H), S(H))$ of problem (2.1) always exists. It is now sufficient to prove that $M_y(H) = 0$ and $S_{uy}(H) = 0$, which will imply immediately that $T(H) = 0 = (\alpha, \beta^t)^t$. Denote $M \equiv M(H)$, $S \equiv S(H)$ and

$$S^{-1} = \begin{pmatrix} S^{uu} & S^{uy} \\ S^{yu} & S^{yy} \end{pmatrix}$$

where $0 < S^{yy} < \infty$ since S is a positive definite matrix. Suppose that (i) $S^{uy} \neq 0$ or (ii) ($S^{uy} = 0$ and $M_y \neq 0$). With $\tilde{S}^{uu} = S^{uu} - (S^{uy}S^{yu})/S^{yy}$, define $\tilde{S}$ by

$$\tilde{S}^{-1} = \begin{pmatrix} \tilde{S}^{uu} & 0 \\ 0 & S^{yy} \end{pmatrix} \text{ and put } \tilde{M} = \begin{pmatrix} M_u \\ 0 \end{pmatrix}.$$

Now by definition of $T(H) = (M(H), S(H))$, and using independence of y and u, we may write

$$b \geq \iint \rho(\alpha_u(y)^{1/2})\, dF(y)\, dG(u)$$

with

$$\alpha_u(y) = (u - M_u)^t S^{uu}(u - M_u) + (y - M_y)^2 S^{yy} + 2(u - M_u)^t S^{uy}(y - M_y)$$

With $t(u) = M_y - \frac{(u-M_u)^t S^{uy}}{S^{yy}} \in I\!R$, we have that $\alpha_u(y) = (y - t(u))^2 S^{yy} + (u - M_u)^t \tilde{S}^{uu}(u - M_u)$. From Lemma 1 it follows that the function

$$t \to \int \rho\left(((y - t)^2 S^{yy} + (u - M_u)^t \tilde{S}^{uu}(u - M_u))^{1/2}\right) dF(y)$$

is symmetric and increasing on $[0, +\infty[$. Therefore, it holds for every u that

$$\int \rho(\alpha_u(y)^{1/2})\, dF(y) \geq \int \rho\left((y^2 S^{yy} + (u - M_u)^t \tilde{S}^{uu}(u - M_u))^{1/2}\right) dF(y),$$

with strict inequality if $t(u) \neq 0$ and $c_u < c^*$, where $c_u = \sqrt{(u - M_u)^t \tilde{S}^{uu}(u - M_u)}$ and c^* defined as in Lemma 1. Denote $A = \{u \,|\, t(u) = 0\}$ and $B = \{u \,|\, c_u \geq c^*\}$. Since for all $u \in B$, $\alpha_u(y)^{1/2} \geq c^*$ for every y, we have that

$$b \geq E_H[\rho(\alpha_u(y)^{1/2})] \geq E_H[\rho(\alpha_u(y)^{1/2}) I(u \in B)] = \rho(\infty) P(B).$$

If $P_G(A \cup B) = 1$, then we would have that $P(A) \geq 1 - P(B) \geq 1 - \frac{b}{\rho(\infty)}$ contradicting hypothesis (G), since A forms a hyperplane in $I\!R^p$. Therefore, we have $P_G(A \cup B) < 1$ and

$$\begin{aligned} b \geq \iint \rho(\alpha_u(y)^{1/2})\, dF(y)\, dG(u) &> \iint \rho\left((y^2 S^{yy} + (u - M_u)^t \tilde{S}^{uu}(u - M_u))^{1/2}\right) dF(y)\, dG(u) \\ &= \int \rho\left(((z - \tilde{M})^t \tilde{S}^{-1}(z - \tilde{M}))^{1/2}\right) dH(z), \end{aligned}$$

while at the same time $\det(S) = \det(\tilde{S})$. Therefore there exists a constant $c < 1$ such that $E_H\left[\rho\left(((z-\tilde{M})^t(c\tilde{S})^{-1}(z-\tilde{M}))^{1/2}\right)\right] \leq b$ while $\det(c\tilde{S}) = c^{p+1}\det(\tilde{S}) < \det(S)$, hereby contradicting the definition of $T(H) = (M(H), S(H))$. We conclude that case (i) and (ii) are excluded, and therefore $M_y = 0$ and $S^{uy} = 0$ (which implies $S_{uy} = 0$). $\square$

To prove Theorem 2 we need the following two lemmas.

Lemma 2. *From the first order condition* (3.2) *for the scatter matrix functional* S*, it follows that*

$$\text{(A.1)}\qquad \left[\frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\, y^2\, \tilde{u}\, dH_0(z)\right]\mathrm{IF}(z;M_y,H_0)+$$
$$\left[\int w_1(d^2_{H_0}(z))\,\tilde{u}\tilde{u}^t\, dH_0(z)+\frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\, y^2\,\tilde{u}\tilde{u}^t\, dH_0(z)\right]S^{-1}_{uu}(H_0)\mathrm{IF}(z;S_{uy},H_0)=$$
$$w_1(d^2_{H_0}(z))y\tilde{u},$$

where $\tilde{u} = u - M_u(H_0)$.

Proof of Lemma 2: Consider the contaminated distribution $H_\varepsilon = (1-\varepsilon)H_0 + \varepsilon\Delta_z$. Lopuhaä (1989) has shown that a solution of problem (2.1) exists for contaminated distributions of this type when ε is sufficiently small. From the (u, y) component of equation (3.2) we obtain

$$(1-\varepsilon)\int w_1(d^2_{H_\varepsilon}(z))(u-M_u(H_\varepsilon))(y-M_y(H_\varepsilon))\, dH_0(z)+\varepsilon\, w_1(d^2_{H_\varepsilon}(z))(u-M_u(H_\varepsilon))(y-M_y(H_\varepsilon))=$$
$$(1-\varepsilon)\int w_2(d^2_{H_\varepsilon}(z))\, dH_0(z)\, S_{uy}(H_\varepsilon)+\varepsilon\, w_2(d^2_{H_\varepsilon}(z))\, S_{uy}(H_\varepsilon).$$

Differentiating both sides of the above equation w.r.t. ε and evaluating at 0 yields

$$-\int w_1(d^2_{H_0}(z))\tilde{u}(y-M_y(H_0))\, dH_0(z)+\int w_1'(d^2_{H_0}(z))\frac{\partial}{\partial\varepsilon}d^2_{H_\varepsilon}(z)_{|\varepsilon=0}\,\tilde{u}(y-M_y(H_0))\, dH_0(z)-$$
$$\int w_1(d^2_{H_0}(z))\mathrm{IF}(z;M_u,H_0)(y-M_y(H_0))\, dH_0(z)-\int w_1(d^2_{H_0}(z))\tilde{u}\mathrm{IF}(z;M_y,H_0)\, dH_0(z)+$$
$$w_1(d^2_{H_0}(z))\tilde{u}(y-M_y(H_0)) = -\int w_2(d^2_{H_0}(z))\, dH_0(z)\, S_{uy}(H_0)+w_2(d^2_{H_0}(z))\, S_{uy}(H_0)+$$
$$\int w_2'(d^2_{H_0}(z))\frac{\partial}{\partial\varepsilon}d^2_{H_\varepsilon}(z)_{|\varepsilon=0}\, dH_0(z)\, S_{uy}(H_0)+\int w_2(d^2_{H_0}(z))\, dH_0(z)\,\mathrm{IF}(z;S_{uy},H_0).$$

where $\tilde{u} = u - M_u(H_0)$. From (3.2) it follows that the first term on the left hand side equals the first term on the right hand side in the above equation. Since $M_y(H_0) = 0$ and $S_{uy}(H_0) = 0$ and using that

$$\frac{\partial}{\partial\varepsilon}d^2_{H_\varepsilon}(z)_{|\varepsilon=0} = (z-M(H_0))^t\mathrm{IF}(z;S^{-1},H_0)(z-M(H_0)) - 2(z-M(H_0))^tS^{-1}(H_0)\mathrm{IF}(z;M,H_0)$$

the previous equation becomes

$$\text{(A.2)}\quad \int w_1'(d^2_{H_0}(z))(z-M(H_0))^t\mathrm{IF}(z;S^{-1},H_0)(z-M(H_0))\tilde{u}y\,dH_0(z) -$$
$$2\int w_1'(d^2_{H_0}(z))(z-M(H_0))^t S^{-1}(H_0)\mathrm{IF}(z;M,H_0)\tilde{u}y\,dH_0(z) -$$
$$\int w_1(d^2_{H_0}(z))y\,dH_0(z)\ \mathrm{IF}(z;M_u,H_0) - \int w_1(d^2_{H_0}(z))\tilde{u}\,dH_0(z)\ \mathrm{IF}(z;M_y,H_0) +$$
$$w_1(d^2_{H_0}(z))\tilde{u}y = \int w_2(d^2_{H_0}(z))\,dH_0(z)\,\mathrm{IF}(z;S_{uy},H_0).$$

Now $\mathrm{IF}(z;S^{-1},H_0) = -S(H_0)^{-1}\mathrm{IF}(z;S,H_0)S(H_0)^{-1}$ (this inequality follows immediately from matrix derivation rules see e.g. (Pullman 1976, page 120)), so

$$\text{(A.3)}\quad (z-M(H_0))^t\mathrm{IF}(z;S^{-1},H_0)(z-M(H_0)) =$$
$$(z-M(H_0))^t S(H_0)^{-1}\mathrm{IF}(z;S,H_0)S(H_0)^{-1}(z-M(H_0)) =$$
$$\tilde{u}^t\mathrm{IF}(z;S_{uu}^{-1},H_0)\tilde{u} + \mathrm{IF}(z;S_{yy},H_0)(y/S_{yy}(H_0))^2 + 2\tilde{u}^t S_{uu}^{-1}(H_0)\mathrm{IF}(z;S_{uy},H_0)y/S_{yy}(H_0)$$

since $M_y(H_0)=0$ and $S_{uy}(H_0)=0$. Also $d^2_{H_0}(z) = y^2/S_{yy}(H_0) + \tilde{u}^t S_{uu}^{-1}(H_0)\tilde{u}$. Therefore, using (A.3), the first integral in expression (A.2) can be split up into three parts. The first part equals

$$\int \tilde{u}^t\mathrm{IF}(z;S_{uu}^{-1},H_0)\tilde{u}\left\{\int w_1'(y^2/S_{yy}(H_0)+\tilde{u}^t S_{uu}^{-1}(H_0)\tilde{u})y\,dF(y)\right\}\tilde{u}\,dG(u) = 0,$$

since the inner integral is zero thanks to symmetry of F. For the same reason we have for the second part

$$\int\left\{\int w_1'(y^2/S_{yy}(H_0)+\tilde{u}^t S_{uu}^{-1}(H_0)\tilde{u})(y/S_{yy}(H_0))^2 y\,dF(y)\right\} dG(u)\mathrm{IF}(z;S_{yy},H_0) = 0.$$

Therefore, the first integral of equation (A.2) becomes

$$\text{(A.4)}\quad \int w_1'(d^2_{H_0}(z))(z-M(H_0))^t\mathrm{IF}(z;S^{-1},H_0)(z-M(H_0))\tilde{u}y\,dH_0(z) =$$
$$\frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\,y^2\,\tilde{u}\tilde{u}^t\,dH_0(z)\ S_{uu}^{-1}(H_0)\mathrm{IF}(z;S_{uy},H_0).$$

The second integral of equation (A.2) can be split up into two parts by using

$$(z-M(H_0))^t S^{-1}(H_0)\mathrm{IF}(z;M,H_0) = \tilde{u}^t S_{uu}^{-1}(H_0)\mathrm{IF}(z;M_u,H_0) + \mathrm{IF}(z;M_y,H_0)y/S_{yy}(H_0).$$

The first part equals

$$\int \tilde{u}^t S_{uu}^{-1}(H_0)\mathrm{IF}(z;M_u,H_0)\left\{\int w_1'(y^2/S_{yy}(H_0)+\tilde{u}^t S_{uu}^{-1}(H_0)\tilde{u})y\,dF(y)\right\}\tilde{u}\,dG(u) = 0,$$

by symmetry of F. Therefore, the second integral of equation (A.2) reduces to

$$\text{(A.5)} \qquad \int w_1'(d^2_{H_0}(z))(z - M(H_0))^t S^{-1}(H_0)\mathrm{IF}(z; M, H_0)\tilde{u}y \, dH_0(z) = \frac{1}{S_{yy}(H_0)} \int w_1'(d^2_{H_0}(z)) \, y^2 \tilde{u} \, dH_0(z) \; \mathrm{IF}(z; M_y, H_0).$$

The integral in the third term of expression (A.2) also equals zero by symmetry of F. From the u component of (3.1) it follows that also the fourth term of (A.2) equals zero. Using the (u, u) component of (3.2) the right hand term of (A.2) can be rewritten as

$$\text{(A.6)} \qquad \int w_2(d^2_{H_0}(z)) \, dH_0(z) \, \mathrm{IF}(z; S_{uy}, H_0) = \int w_1(d^2_{H_0}(z))\tilde{u}\tilde{u}^t \, dH_0(z) \; S^{-1}_{uu}(H_0)$$

Substituting (A.4), (A.5), and (A.6) in expression (A.2) yields the desired result. □

Starting from the y component of (3.1), with similar computations as in Lemma 2, the next lemma can be proven.

Lemma 3. *From the first order condition* (3.1) *for the location functional* M*, it follows that*

$$\text{(A.7)} \qquad \left[\int w_1(d^2_{H_0}(z)) \, dH_0(z) + \frac{2}{S_{yy}(H_0)} \int w_1'(d^2_{H_0}(z)) \, y^2 \, dH_0(z)\right] \; \mathrm{IF}(z; M_y, H_0) + \frac{2}{S_{yy}(H_0)} \int w_1'(d^2_{H_0}(z)) \, y^2 \, \tilde{u}^t \, dH_0(z) \; S^{-1}_{uu}(H_0)\mathrm{IF}(z; S_{uy}, H_0) = w_1(d^2_{H_0}(z))y$$

where $\tilde{u} = u - M_u(H_0)$*.*

Using Lemma 2 and lemma 3 given above we can prove Theorem 2.

Proof of Theorem 2: We first derive the influence function at H_0. We write

$$T(H_0) = \begin{pmatrix} a(H_0) \\ b(H_0) \end{pmatrix} = Q(H_0)^{-1} \begin{pmatrix} M_y(H_0) \\ S^{-1}_{uu}(H_0)S_{uy}(H_0) \end{pmatrix},$$

with

$$Q(H_0) = \begin{pmatrix} 1 & M_u^t(H_0) \\ 0 & I_p \end{pmatrix},$$

and I_p the identity matrix. Since $M_y(H_0) = 0$ and $S_{uy}(H_0) = 0$ it follows that

$$\text{(A.8)} \qquad \mathrm{IF}(z; T, H_0) = Q(H_0)^{-1} \begin{pmatrix} \mathrm{IF}(z; M_y, H_0) \\ S^{-1}_{uu}(H_0)\mathrm{IF}(z; S_{uy}, H_0) \end{pmatrix}.$$

We can combine Lemma 2 and Lemma 3 in one single equation:

(A.9)

$$\left[\int w_1(d^2_{H_0}(z))\,\tilde{x}\tilde{x}^t\,dH_0(z) + \frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\,y^2\,\tilde{x}\tilde{x}^t\,dH_0(z)\right]\begin{pmatrix}\mathrm{IF}(z;M_y,H_0)\\ S_{uu}^{-1}(H_0)\mathrm{IF}(z;S_{uy},H_0)\end{pmatrix} = w_1(d^2_{H_0}(z))y\tilde{x}$$

with $\tilde{x} = (1, \tilde{u}^t)^t$ and where we used $\int w_1(d^2_{H_0}(z))\,\tilde{u}\,dH_0(z) = 0$ (which follows from (3.1)). Together with expression (A.8) this yields

(A.10) $$\mathrm{IF}(z;T,H_0) = Q(H_0)^{-1}A(H_0)^{-1}w_1(d^2_{H_0}(z))\tilde{x}y$$

where

(A.11) $$A(H_0) = \int w_1(d^2_{H_0}(z))\,\tilde{x}\tilde{x}^t\,dH_0(z) + \frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\,y^2\,\tilde{x}\tilde{x}^t\,dH_0(z)$$

Now we can easily check that $Q(H_0)^t\tilde{x} = (1, u^t)^t = x$. It follows from (A.10) and $Q(H_0)^{-1} = -Q(H_0)$ that

(A.12)
$$\begin{aligned}\mathrm{IF}(z;T,H_0) &= -Q(H_0)^{-1}A(H_0)^{-1}Q(H_0)^t w_1(d^2_{H_0}(z))xy\\ &= (Q(H_0)^t A(H_0)Q(H_0))^{-1}w_1(d^2_{H_0}(z))xy\\ &= C(H_0)^{-1}w_1(d^2_{H_0}(z))xy\end{aligned}$$

where $C(H_0) = Q(H_0)^t A(H_0)Q(H_0)$ equals, using (A.11),

(A.13) $$C(H_0) = \int w_1(d^2_{H_0}(z))\,xx^t\,dH_0(z) + \frac{2}{S_{yy}(H_0)}\int w_1'(d^2_{H_0}(z))\,y^2\,xx^t\,dH_0(z).$$

Using integration by parts, expression (A.13) can be rewritten as (3.5).

Finally, note that if $z = (u^t, y)^t \sim H$, then $Az + c \sim H_0$ with $A = \begin{pmatrix} I_p & 0\\ -\beta^t & 1\end{pmatrix}$, and $c = \begin{pmatrix} 0\\ -\alpha\end{pmatrix}$. By equivariance of the functional T we have $T(H) = T(H_0) + \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$ and therefore

(A.14) $$\mathrm{IF}(z;T,H) = \mathrm{IF}(Az + c;T,H_0) = \mathrm{IF}((u, y - x^t\theta);T,H_0)$$

Due to the affine equivariance of the S-estimator, we have

$$\begin{aligned}d^2_H(z) &= (z - M(H))^t S(H)^{-1}(z - M(H))\\ &= (Az - (M(H_0) - c))^t S(H_0)^{-1}(Az - (M(H_0) - c))\\ &= d^2_{H_0}(Az + c)\end{aligned}$$

for all $z \in I\!R^{p+1}$. Therefore, it follows from (A.14) and (A.12) that

$$\mathrm{IF}(z;T,H) = C(H_0)^{-1}w_1(d^2_H(z))x(y - x^t\theta).$$

□

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, London.

Chang, W.H., McKean, J.W., Naranjo, J.D. and Sheather, S.J. (1999), High-breakdown rank regression, *Journal of the American Statistical Association,* **94**, 205-219.

Cheng, C. and Van Ness, J.W. (1997), Robust calibration, *Technometrics,* **39**, 401-411.

Coakley, C.W. and Hettmansperger, T.P. (1993), A bounded influence, high breakdown, efficient regression estimator, *Journal of the American Statistical Association,* **88**, 872-880.

Cook, R.D., Hawkins, D.M. and Weisberg, S. (1992), Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits, *Journal of the American Statistical Association,* **87**, 419-424.

Cox, T.F. and Cox, M.A.A. (1994), *Multidimensional Scaling,* Chapman and Hall, London.

Croux, C. and Dehon, C. (2001), Robust linear discriminant analysis using S-estimators, *The Canadian Journal of Statistics*, **2**9, 473-492

Croux, C., Dehon, C., Van Aelst, S. and Rousseeuw, P. (2001), Robust estimation of the conditional median function at elliptical models, *Statistics and Probability Letters,* **51**, 361–368.

Croux, C. and Haesbroeck, G. (1999), Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of Multivariate Analysis,* **71**, 161–190.

Croux, C. and Haesbroeck, G. (2000), Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, *Biometrika,* **87**, 603–618

Croux, C., Rousseeuw, P.J. and Hössjer, O. (1994), Generalized S-estimators, *Journal of the American Statistical Association,* **89**, 1271-1281.

Davies, L. (1987), Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, *The Annals of Statistics,* **15**, 1269-1292.

Ferretti, N., Kelmansky, D., Yohai, V.J. and Zamar, R.H. (1999), A class of locally and globally robust regression estimates, *Journal of the American Statistical Association,* **94**, 174-188.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* John Wiley and Sons, New York.

Härdle, W. (1991), *Smoothing Techniques With Implementation in S,* Springer-Verlag, New York.

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984), Location of several outliers in multiple regression data using elemental sets, *Technometrics,* **26**, 197-208.

Hettmansperger, T.P. and Sheather, S.J. (1992), A cautionary note on the method of least median squares, *The American statistician,* **46**, 79-83.

Iglewicz, B. (1982), Robust scale estimates, *Understanding Robust and Exploratory Data Analysis*, (eds. D.C. Hoaglin, F. Mosteller and J.W. Tukey), 404-431, John Wiley and Sons, New York.

Lopuhaä H.P. (1989), On the relation between S-estimators and M-estimators of multivariate location and covariance, *The Annals of Statistics,* **17**, 1662-1683.

Lopuhaä, H. P. (1997), Asymptotic expansion of S-estimators of location and covariance, *Statistica Neerlandica,* **5**1 220–237.

Maronna, R. (1976), Robust M-estimates of multivariate location and scatter, *The Annals of Statistics,* **4**, 51-67.

Maronna, R. and Morgenthaler, S. (1986), Robust regression through robust covariances, *Communications in Statistics: Theory and Methods,* **15**, 1347-1365.

Naranjo, J. D. and Hettmansperger, T. P. (1994), Bounded influence rank regression, *Journal of the Royal Statistical Society Series B,* **56**, 209–220.

McKean, J.W., Naranjo, J.D. and Sheather, S.J. (1996), Diagnostics to detect differences in robust fits of linear models, *Computational Statistics,* **11**, 223-243.

McKean, J.W., Sheather, S.J. and Hettmansperger, T.P. (1990), Regression Diagnostics for rank-based methods, *Journal of the American Statistical Association,* **85**, 1018-1028.

McKean, J.W., Sheather, S.J. and Hettmansperger, T.P. (1993), The use and interpretation of residuals based on robust estimation, *Journal of the American Statistical Association,* **88**, 1254-1263.

Pullman, N.J. (1976), *Matrix Theory and its Applications: Selected Topics,* Marcel Dekker, New York.

Rousseeuw, P.J. (1984), Least median of squares regression, *Journal of the American Statistical Association,* **79**, 871–880.

Rousseeuw, P.J. (1985), Multivariate estimation with high breakdown point, *Mathematical statistics and applications, Vol. B*, (eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz), 283-297, Reidel, Dordrecht.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley-Interscience, New York.

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K. and Agulló, J. (2000), Robust multivariate regression, Technical Report, University of Antwerp.

Rousseeuw, P.J. and van Zomeren, B.C. (1990), Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association,* **85**, 633-651.

Ruppert, D. (1992), Computing S-estimators for regression and multivariate location/dispersion, *Journal of Computational and Graphical Statistics,* **1**, 253-270.

Sheather, S.J., McKean, J.W. and Hettmansperger, T.P. (1997), Finite sample stability properties of the least median of squares estimator, *Journal of Statistical Computation and Simulation,* **58**, 371-383.

Simpson, D.G., Ruppert, D. and Carroll, R.J. (1992), On one-step GM estimates and stability of inferences in linear regression, *Journal of the American Statistical Association,* **87**, 439-450.

Simpson, D.G. and Yohai, V.J. (1998), Functional stability of one-step GM-estimators in approximately linear regression, *The Annals of Statistics,* **26**, 1147-1169.

Stahel, W.A. (1981), Robust Estimation: Infinitesimal optimality and covariance matrix estimators, PhD. Thesis (in German), ETH, Zurich.

Visuri, S., Koivunen, V., Möttönen, J., Ollila, E. and Oja, H. (2001), Affine equivariant multivariate rank methods, to appear in *Journal of Statistical Planning and Inference.*

Woodruff, D.L. and Rocke, D.M. (1994), Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association,* **89**, 888-896.

Yohai, V.J. (1987), High breakdown point and high efficiency robust estimates for regression, *The Annals of Statistics,* **15**, 642-656.

Yohai, V.J., Stahel, W.A. and Zamar, R.H. (1991), A procedure for robust estimation and inference in linear regression, *Directions in Robust Statistics and Diagnostics,* (eds. W. Stahel and S. Weisberg), 365-374, Springer-Verlag, New York.

Yohai, V.J. and Zamar, R.H. (1988), High breakdown-point estimates of regression by means of the minimization of an efficient scale, *Journal of the American Statistical Association,* **83**, 406-413.

Table 1: MSE and Bias for the S-CovReg, M1M, S1M, HBR and LTS estimators of the slope parameters in a regression model with $p = 4$ and $n = 100$, once for the 50% breakdown point and once for the 25% breakdown point versions of these estimators. Samples were generated without outliers (NOR), with 10% and 20% of vertical outliers (VO10 and VO20) and with 10% and 20% of bad leverage points (BL10 and BL20).

	MSE					Bias ($\times 10^{-3}$)				
BDP=50%	NOR	VO10	VO20	BL10	BL20	NOR	VO10	VO20	BL10	BL20
S-CovReg	1.51	1.51	2.05	1.46	1.57	5.8	2.0	4.4	3.3	5.0
M1M	1.92	1.79	1.60	4.56	7.57	3.2	2.3	3.5	106.1	179.0
S1M	1.63	2.24	3.34	5.65	8.04	3.1	5.1	5.0	44.9	35.1
HBR	1.16	1.60	2.46	6.91	7.54	4.1	4.1	4.6	231.1	249.4
LTS	6.98	6.81	5.74	12.45	12.68	7.6	4.3	7.1	225.4	250.0
BDP=25%										
S-CovReg	1.14	1.93	63.66	1.20	7.42	4.6	3.4	14.3	3.6	249.4
M1M	1.20	1.29	1.41	2.87	4.82	4.2	2.0	3.8	120.7	184.5
S1M	1.19	1.86	3.11	3.02	4.26	4.1	5.2	5.1	15.2	17.2
HBR	1.15	1.64	2.56	7.09	7.57	4.1	4.2	4.6	236.8	249.4
LTS	3.51	2.84	1.92	9.66	10.22	5.1	1.2	4.3	234.6	250.0

Table 2: Maximal difference of the slope estimate under slight changes of the data points, over all 100 simulations, for the S-CovReg, M1M, S1M, HBR and LTS estimator.

	S-CovReg	M1M	S1M	HBR	LTS
Maximum	0.065	0.125	0.140	0.077	0.689

Table 3: Estimates of the intercept and regression parameters for (a) the Hawkins-Bradu-Kass data and (b) the clean Hawkins-Bradu-Kass data. Standards errors are reported between parenthesis, correlations between estimated coefficients are in the right panel of the table. The estimators considered are the Least Squares (LS) estimator, the estimator based on the robust S-estimator of location/scatter, and an MM-estimator.

(a)

$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$
LS-estimator				$\hat{\beta}_1$	-0.637		
-0.388	0.239	-0.335	0.383	$\hat{\beta}_2$	-0.180	-0.084	
(0.405)	(0.255)	(0.015)	(0.125)	$\hat{\beta}_3$	0.470	-0.540	-0.775
Robust Covariance Based				$\hat{\beta}_1$	-0.360		
-0.018	0.097	0.004	-0.130	$\hat{\beta}_2$	-0.635	-0.009	
(0.226)	(0.079)	(0.078)	(0.077)	$\hat{\beta}_3$	-0.386	-0.316	-0.086
MM-estimator				$\hat{\beta}_1$	-0.648		
-0.181	0.081	0.040	-0.052	$\hat{\beta}_2$	-0.164	-0.084	
(0.114)	(0.073)	(0.044)	(0.039)	$\hat{\beta}_3$	0.426	-0.487	-0.795

(b)

$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$
LS-estimator				$\hat{\beta}_1$	-0.456		
-0.010	0.062	0.012	-0.107	$\hat{\beta}_2$	-0.527	-0.031	
(0.190)	(0.067)	(0.066)	(0.069)	$\hat{\beta}_3$	-0.481	-0.102	-0.123
Robust Covariance Based				$\hat{\beta}_1$	-0.328		
-0.021	0.123	-0.001	-0.147	$\hat{\beta}_2$	-0.672	-0.020	
(0.253)	(0.087)	(0.088)	(0.082)	$\hat{\beta}_3$	-0.371	-0.354	-0.058
MM-estimator				$\hat{\beta}_1$	-0.463		
-0.011	0.062	0.012	-0.107	$\hat{\beta}_2$	-0.533	-0.008	
(0.245)	(0.086)	(0.086)	(0.090)	$\hat{\beta}_3$	-0.451	-0.127	-0.149

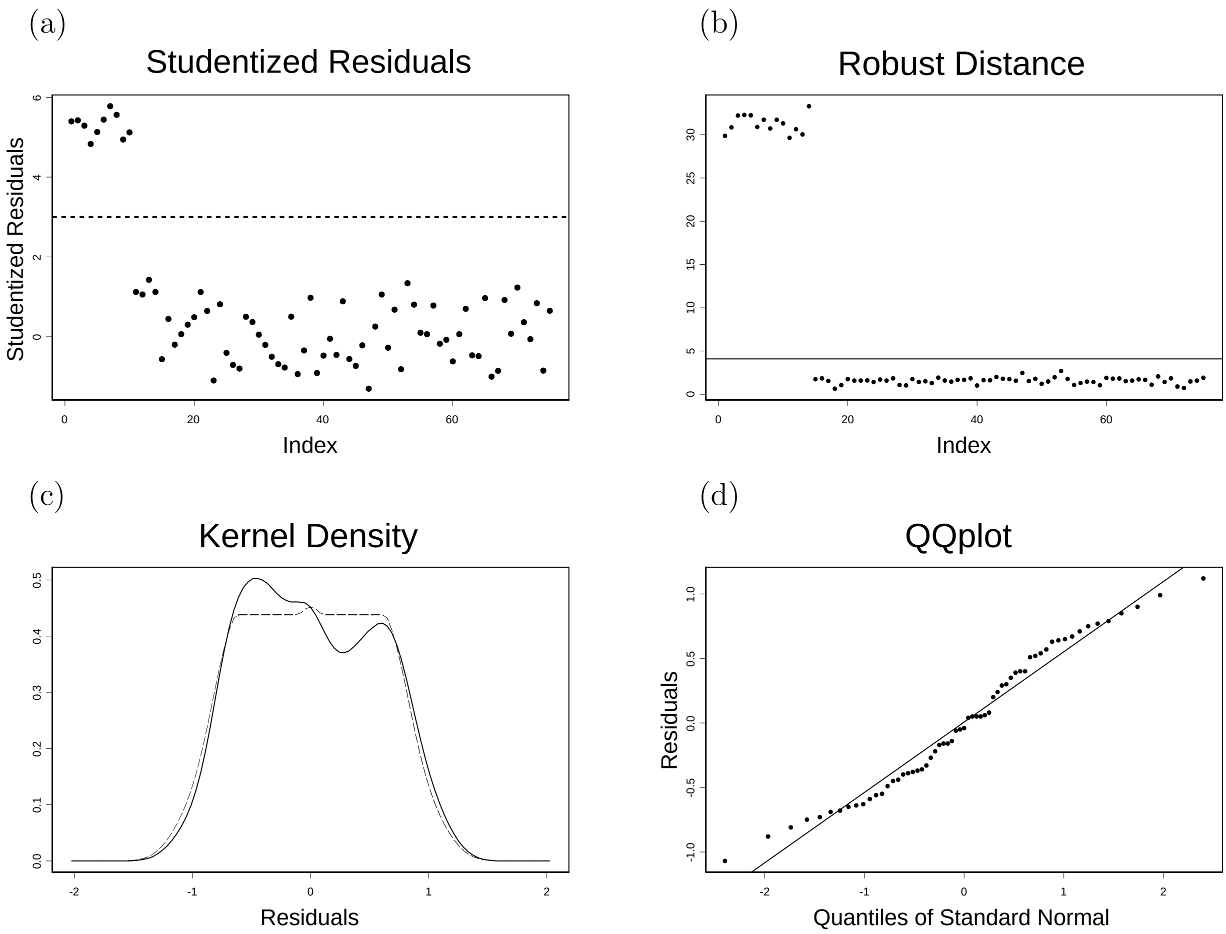
Table 4: Estimates of the intercept and regression parameters for the Hawkins-Bradu-Kass data by the method based on a robust covariance S-estimator, as in Table 3. Now, for computing standards errors and correlations between estimated coefficients, the hypothesis of normality was used.

$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$
Complete data set				$\hat{\beta}_1$	-0.538		
-0.018	0.097	0.004	-0.130	$\hat{\beta}_2$	-0.457	-0.028	
(0.297)	(0.102)	(0.101)	(0.111)	$\hat{\beta}_3$	-0.572	0.063	-0.114
Clean data set				$\hat{\beta}_1$	-0.530		
-0.021	0.123	-0.001	-0.147	$\hat{\beta}_2$	-0.461	-0.002	
(0.231)	(0.080)	(0.079)	(0.089)	$\hat{\beta}_3$	-0.534	0.003	-0.150

Figure caption

Figure 1: Diagnostic plots for the residuals of the regression estimator based on an S-estimator of multivariate location/scatter for the Hawkins-Bradu-Kass data: (a) studentized residuals (b) robust distances versus their index (c) kernel-based density estimate (solid line) and its symmetric, unimodal version (dashed line) (d) QQ-plot of the residuals.

Figure 1


(a)
Studentized Residuals
Studentized Residuals
Index
(b)
Robust Distance
Index
(c)
Kernel Density
Residuals
(d)
QQplot
Residuals
Quantiles of Standard Normal