An investigation of GM-estimators for outlier robust regression estimation

Dimitar G. Genov[1] and Nasko R. Atanasov[2]

Abstract - **This paper discusses the class of GM-estimators for outlier robust regression estimation. M-estimators with objective functions of Cauchy, Welsh, Huber, Tukey, Mallows' GM-estimators and Schweppe's GM-estimators are investigated. An accuracy, a convergence and a computational complexity of the algorithms are analyzed. The best model is determined by robust Akaike information criterion.**

Keywords - **Robust estimators, Outliers, Akaike information criterion, Mallows' GM-estimators, Schweppe's GM-estimators.**

I. Introduction

Consider the linear regression model $y_i = x_i^T \beta + \varepsilon_i, i = 1, \ldots, N$, where x_i is a p-dimensional vector. The class of generalized maximum likelihood (GM) type estimators is defined implicity by the condition

$$\sum_{i=1}^{N} x_i \varsigma\left(x_i, \left(y_i - x_i^T \beta\right)/\sigma\right) = 0 . \qquad (1)$$

The parameter σ denotes the scale of ε_i. The function $\varsigma(\cdot,\cdot)$ depends on both the set of regressors (x_i) and the standardized residual. The most important conditions that must be satisfied by $\varsigma(\cdot,\cdot)$, in order for the GM estimator to have nice asymptotic properties are that for all $x \in R^p$ $\varsigma(x,\cdot)$ has to be continuous and continuously differentiable except in a finite number of points, that $\varsigma(x,\cdot)$ has no vertical asymptotes, and that $\varsigma(x,\cdot)$ is odd. Moreover $\mathrm{E}\left(\left(\varsigma(x_i, \varepsilon_i/\sigma)\right)^2 x_i x_i^T\right)$ and $\mathrm{E}\left(\varsigma'(x_i, \varepsilon_i/\sigma) x_i x_i^T\right)$ must exist and be nonsingular, where $\varsigma'(x_i, r) = \partial\varsigma(x_i, r)/\partial r$, and r denotes standardized residual ε_i/σ [2], [6].

The OLS estimator is obtained as a special case of Eq. (1) by setting $\varsigma(x, r) = r$. Also M-estimators are special case of Eq. (1), namely $\varsigma(x,r) = \psi(r)$ for some function ψ satisfying the above regularity conditions [6].

Instead of defining GM estimators as a solition to a first order condition of the type Eq. (1), one can also define them as the minimand of the objective function

$$\sum_{i=1}^{N} \tau\left(x_i, \left(y_i - x_i^T \beta\right)/\sigma\right), \qquad (2)$$

with $\partial\tau(x,r)/\partial r = \varsigma(x,r)$. The focus in this paper, however, is on the definition as aplied by Eq.(1). Note that the OLS estimator is defined by setting $\tau(x,r) = r^2/2$, while the class of M-estimators is obtained by setting $\tau(x,r) = \rho(r)$ with $d\rho(r)/dr = \psi(r)$ [1],[3].

II. Features of Mallows' GM-estimators and Schweppe's GM-estimators

The easiest way to explain the intuition behind GM estimators is by considering the class of Mallows' GM estimators, given by $\varsigma(x,r) = w_x \psi(r)$, with $\psi(r)$ as intrtoduced above (Section I), and $w_x(x)$ a weight function that assigns weights to the vectors of regressors, $w_x : R^p \rightarrow [0,1]$. Using this specification of $\varsigma(\cdot,\cdot)$, Eq. (1) can be rewritten as

$$\sum_{i=1}^{N} w_x(x_i) x_i w_r\left(\left(y_i - x_i^T \beta\right)/\sigma\right)\left(y_i - x_i^T \beta\right) = 0 , \qquad (3)$$

$$w_r(r) = \begin{cases} \psi(r)/r & за\ r \neq 0 \\ 1 & за\ r = 0 \end{cases} . \qquad (4)$$

The functions $\psi(\cdot)$ and $w_x(\cdot)$ can now be chosen such that the weight of ith observation decreases if either $\left(y_i - x_i^T \beta\right)/\sigma$ becomes extremely large (vertical outliers and bad leverage points), or x_i becomes large (leverage points)[6]. In this way, outliers and influential observations automatically receive less weight. For the OLS estimator, $w_x(x) \equiv 1$ and $w_r(r) \equiv 1$, such that all observation receive the same weight.

A disadvantage of Mallows' proposal for GM estimators is that it assigns less weight to both good and bad leverage points, but good leverage points often increase the efficiency of the imployed estimator. As an alternative to Mallows' proposal for GM-estimators, one can consider the proposal of Schweppe. The Schweppe's form of the GM estimator only downweights vertical outliers and bad leverage points, but not good leverage points. This generally increases the efficiency of the Schweppe's estimator over the Mallows' version. The Schweppe's specification of $\varsigma(\cdot,\cdot)$ is given by

[1] Dimitar G. Genov is with the Faculty of Computer System and Automation, Technical University, 9010 Varna, 1 "Studentska" str E-mail:dggenov@yahoo.com

[2] Nasko R. Atanasov is with the Faculty of Computer System and Automation, Technical University, 9010 Varna, 1 "Studentska" str E-mail:nratanasov@yahoo.com

$$\varsigma(x,r) = w_x \psi(r/w_x(x)). \quad (5)$$

Using Eq. (5), Eq. (1) can be written as

$$\sum_{i=1}^{N} x_i w_r \left(\left(y_i - x_i^T \beta \right) / \sigma w_x(x_i) \right) \left(y_i - x_i^T \beta \right) = 0 . \quad (6)$$

Assume that $w_x(\cdot)$ and $\psi(\cdot)$ are chosen such that outliers receive less weight. For a leverage point (y,x), $w_x(x)$ will than be small. The weight for the ith observation in the estimation proces is given by $w_r(\cdot)$ in Eq. (6). Note that this weight may be close to one if the standardized residual is close to zero, irrespective of whether the observation is a leverage point or not. The requirement that the standardized residual is close to zero becomes stricter if $w_x(x_i)$ is small, i.e. if x_i is a leverage point.

The Schweppe's version of GM-estimators also has some practical disadvantages. First, the bias in the Schweppe's estimator may be larger than that of the Mallows' estimator. Second, the Schweppe's estimator more easily displays convergence problem than the Mallows' variant, especialy if strongly redescending specification of ψ are used. Even if no convergence problems arise, moderately bad leverage points tend to have a larger influence on the Schweppe's version of the GM-estimators than on the Mallows' version [4],[6].

If the weights on the regressors $w_x(\cdot)$ are dropped, the class of GM-estimators reduces to the class of M-estimators. Thus, the class of GM-estimators contains the class of maximum likelihood type estimators (M-estimators). Therefore, the class of M-estimators is not dealt with, separately.

In the next part of this paper it will be discuss the problem about specification of $\psi(\cdot)$ and $w_x(\cdot)$. The OLS specification for $\psi(\cdot)$, $\psi(r) = r$, is the most familiar one. OLS-estimator is not robust. The most important reason for this is that the function $\psi(r) = r$ is unbounded. Several forms of bounded ψ functions are suggested in the literature,e.g., the Huber, the Cauchy, the Geman-McClure, the Welsch, the Tuke, the Hampel, the Student t specification, etc.[3],[6].

The Huber's function ψ is given by $\psi(r) = median(-c, c, r)$, where c>0 is a tuning constant[1]. The lower *c*, the more robust is the resulting estimator. As a special case of the Huber's estimator, one can obtain the OLS estimator ($c \rightarrow \infty$) and the least absolute deviations (LAD) estimator ($c \rightarrow 0$). The constant *c* not only determines the robustness of the coresponding estimators, but also its efficiency. For Gaussian ε_i, for example, the efficiency of the estimator is an increasing function of *c*. This illustrates that there is a tradeoff beetween efficiency and robustness. Common value for *c* is 1,345 for the Huber's function. This value produce estimators that have an efficiency of 95% in case ε_i is normally distributed [3].

As a specification for the weight function $w_x(\cdot)$ for the regressors, one usually encounters the specification

$$w_x(x_i) = \sqrt{1 - h_i}, \; i = 1, \ldots, N, \quad (7)$$

with h_i the diagonal elements of the hat matrix H [1],[4]

$$\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T y = Hy . \quad (8)$$

III. The generalized procedure for robust estimation

GM estimators are mostly computed by means of numerical techniques. Most of these techniques employ iteration schemes. Therefore, an initial estimate is required to start up the iteration. A starting value should, preferably, be easy to calculate. From this perspective, the OLS-estimator usually is used.

Once the starting values have been obtained, one can start an iteration scheme for solving Eq. (1). It is, of course, possible to use general techniques for solwing sets of nonlinear equations. The special structure of Eq. (1), however, also allows a different iteration scheme by means of weigthed and ordinary least - squares.

A very important computational aspect concerns the estimation of scale parameter σ. If σ is omitted from Eq. (1), the GM-estimator is not scale invariant, i.e., the estimates would change if both y_i and x_i were multiplied by a constant k>0. The estimate σ, one cannot safely use the ordinary standard deviation, as this estimator is not robust. An often used alternative is the median absolute deviation, defined as

$$MAD\left(\{\varepsilon_i\}_{i=1}^{N}\right) = median\left|\varepsilon_i - median(\varepsilon_i)\right| . \quad (9)$$

The MAD is usually multiplied by 1,4826 to make it consistent estimator of the standard deviation for Gaussian ε_i. The use of a scale equivariant estimator for σ in Eq. (1) renders the GM-estimator for β scale equivariant [6].

We proposed the generalized procedure for robust estimation includes the sequent stage, listed below:

- choice of estimator's type;
- choice of method for estimation;
- calculation of the model's parameters with chosen set of structures;
- choice of the best structure from this set, with applying of robust Akaike information criterion;
- test for first order autocorrelation in the residuals;
- final choice of the model, after analysis of results from applying of different estimation methods and estimator's types.

IV. Model selection and model diagnostics

The best model among competing multivariate models is determined by robust Akaike information criterion [2]

$$AICR = 2LFR + 2p . \quad (10)$$

The parameter *p* denotes the number of estimated parameters. The robust loss function LFR depends on both the number of observations (N) and the objective function $\rho(r_i)$ of the standardized residuals

$$LFR = N^{-1}\sum_{i=1}^{N}\rho(r_i). \qquad (11)$$

AICCR (robust Akaike information criterion corrected) is used when the ratio $N/p < 40$ [5]

$$AICCR = AICR + \frac{2p(p+1)}{N-p-1}. \qquad (12)$$

The Durbin-Watson test for first order autocorrelation in the residuals is used

$$DW = \sum_{i=2}^{N}(r_i - r_{i-1})^2 \Big/ \sum_{i=1}^{N} r_i^{\,2}. \qquad (13)$$

V. Simulation investigations

It is created an applied software in MATLAB with realization of the sequent varieties:

- estimator's type – Mallows' GM, Schweppe's GM, M-estimator;
- estimation methods – modified residuals, modified weigths, pseudo observations, Huber-Kleiner's method;
- objective function - Huber, Cauchy, Geman-McClure, Welsch, Tuke, "Fair", L_p, L_1.
- scale estimation – MAD or ordinary standard deviation.

Simulation research is made with the model:

$$y = -68 + 13x_1 + 23x_2. \qquad (14)$$

Number of the observations is 100. To the output model signal is added Gaussian white noise and the ratio noise/signal is 7,76%. There are simulated 10 additional outliers and as a result the ratio noise/signal increases up to 32%.

The accuracy of the estimations $\hat{b}_j,\ j = 0,\ldots,p$ is defined by relative mean-squared error

$$Qb = \sqrt{\sum_{j=0}^{p}\left(b_j - \hat{b}_j\right)^2 \Big/ \sum_{j=0}^{p} b_j^2}\ . \qquad (15)$$

The parameters $b_j,\ j = 0,\ldots,p$ denote the real values of model's parameters.

Criterion for stopping the iterative procedure is

$$\varepsilon = \left|\hat{b}^i - \hat{b}^{i-1}\right| \le 0{,}0001. \qquad (16)$$

Computational complexity of the algorithm is estimated by the cumulative number of floating point operations into MATLAB (flops).

At the estimation of the parameters there is used general regressive model from the type

$$y = b_0 + \sum_{i=1}^{p} b_i x_i + \sum_{\substack{i=1 \\ j>i}}^{p} b_{ij} x_i x_j + \sum_{i=1}^{p} b_{ii} x_i^{\,2}. \qquad (17)$$

There are made a lot of investigations with the mentioned in Section V eight objective functions. The best results in aspect of convergence and accuracy is given by Huber's function and it is used in the next investistigation Eq. (18).

$$\rho(r) = \begin{cases} r^2/2 & if \quad |r| \le c \\ c|r| - c^2/2 & if \quad |r| > c \end{cases}. \qquad (18)$$

Table I

Choice of estimator's type and method

			Q_b	iter	flops
M-estimator	Modified residuals	MAD	.0034	7	13703
		std	.0076	6	78202
	Modified weights	MAD	.0034	6	36744
		std	.0076	5	95422
	Pseudo observetions	MAD	.0034	7	14843
		std	.0076	6	78667
	Huber-Kleiner	MAD			
		std	.0034	8	83893
Mallows' GM	Modified residuals	MAD	.0034	7	79436
		std	.0075	6	79767
	Modified weights	MAD	.0034	6	102143
		std	.0075	5	96674
	Pseudo observetions	MAD	.0034	7	80576
		std	.0075	6	80232
	Huber-Kleiner	MAD			
		std	.0032	9	87859
Schweppe's GM	Modified residuals	MAD	.0034	7	79436
		std	.0075	6	79767
	Modified weights	MAD	.0034	6	102143
		std	.0075	5	96674
	Pseudo observetions	MAD	.0034	7	80576
		std	.0075	6	80232
	Huber-Kleiner	MAD			
		std	.0033	7	83070

The results from experiments are given in Table I. The relative mean-squared error (Q_b), number of iterations for reaching the preassigned accuracy ε (iter) and the computational complexity (flops) are given in accordance with the type of estimators (Mallows' GM, Schweppe's GM and M-estimator), the used method for estimation (modified residuals, modified weigths, pseudo observations, Huber-Kleiner's method) and the chosen scale estimators of residuals (median absolute deviation – MAD or ordinary standard deviation - std).

Table II

Model selection and model diagnostics

	M1	M2	M3	M4	M5	M6
b_o	-67,28	-73,04	-68,05	-68,04	-67,91	-68,19
b_1		12,77	12,93	12,94	12,96	13,02
b_2	22,68		22,77	22,78	22,79	22,91
b_{12}				0,171	0,118	0,192
b_{11}					-0,135	-0,103
b_{22}						0,252
LFR	35,74	107,4	5,63	5,553	5,481	5,406
AICR	75,49	218,8	17,26	19,11	20,96	22,81
AICCR	75,62	218,7	17,51	19,53	21,60	23,71
DW	1,88	2,065	2,40	2,402	2,408	2,415

Table II displays the estimation parameters, the model selection and diagnostics. The data are simulated with model Eq. 14. Mallows' GM-estimator with modified weights method and MAD as scale estimator is applied. The best model is

$$y = -68.05 + 12.93x_1 + 22.77x_2 \, .$$

The corresponding values of AICR and AICCR are minimal and DW=2,4. Therefore, there is not first order autocorrelation in the residuals and the estimations are unbiassed.

VI. Conclusion

From the implemented research then can be made some conclusions:

- the minimum error (Q_b=0,0032) is achieved with Mallows' GM-estimator with Huber-Kleiner's method, but This method gives the slowest convergence;
- the rate of convergence, at every forms of the robust estimators, is the greatest with modified weights method, but this method also gives the greatest computational complexity;
- modified residuals method gives the least computational complexity;
- the accuracy with MAD scale estimators is nearly twice as big as than the ordinary standart deviation one;
- the usage of the studentized residuals (it is not given in the Taable I) leads to considerably increasing of the computational complexity of the algorithms (for Schweppe's GM with Huber-Kleiner's method - flops=83070, but for Schweppe's GM with Huber-Kleiner's method with studentised residuals - flops=2664134);
- The equal error (Q_b=0,0034 or Q_b=0,0075) is dued to the application of the same objective function in every estimators – Huber's function Eq. (18).

VII. Appendix

Cosider the problem for investigation of micro motors. There is not completed design theory in this area of knowledge. Therefore, some characteristics must be defined experimentally. This leads to creation of mathematical model.

Factors, their values and steps of variation are given in Table III.

Table III
Micro motor

EP55/110A	L_{rotor} [mm]	L_{stator} [mm]	W_{rotor} [number of wind]	$L_{wavstator}$ [mm]
factors	X_1	X_2	X_3	X_4
Base level	40	40	54	167
step	2,5	2,5	10	5
Upper level	42,5	42,5	64	172
Lower level	37,5	37,5	44	162

It is realized Full Factors Experiment with $N=2^4$. The motor's efficiency is a function of factors $\eta = \eta(x_1, x_2, x_3, x_4)$ and is obtained by applying of Mallows' GM-estimator with modified weight method and MAD for scale estimation.

The best model is

$$\begin{aligned} \eta = 49.2 - 5.49x_1 + 0.97x_2 + 5.32x_3 - \\ - 0.28x_4 + 0.73x_1x_2 - 4.97x_1x_3 \end{aligned} \, .$$

Indicators, according which the best model is chosen, are: LFR=0,3271; AICR=14,7941; AICCR=28,7941; DW=1,9244.

References

[1] Хьюбер П., Робастность в статистике, Москва, Мир, 1984.

[2] Хампель Ф., и др., Робастность в статистике. Подход на основе функций влияния, Москва, Мир, 1989.

[3] Zhang Zh., M-estimators, www-sop.inria.fr/robotvis/personnel/zzhang/Publis/Tutorial-Estim/node24

[4] Chave A., Thomson D, A bounded influence regression estimator based on the statistics of the hat matrix, J. Roy. Stat. Soc., Series C (Appl. Statist.), 52, 307-322, 2003.

[5] Godinez-Dominguez E., Freire J., Information-theoretic approach for selection of spatial and temporal models of community organization, MARINE ECOLOGY PROGRESS SERIES, Vol.253:17-24, 2003.

[6] Lukas A., A brief intraduction to robust statistics, www.staff.feweb.vu.nl/alucas/thesis.