International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

COMPARING THE PREDICTIVE PERFORMANCE OF OLS AND 7 ROBUST LINEAR REGRESSION ESTIMATORS ON A REAL AND SIMULATED DATASETS

Sacha Varin
Department of Mathematics
Collège Villamont, Lausanne, Switzerland

Abstract – **Robust regression techniques are relevant tools for investigating data contaminated with influential observations. The article briefly reviews and describes 7 robust estimators for linear regression, including popular ones (Huber M, Tukey's bisquare M, least absolute deviation also called L1 or median regression), some that combine high breakdown and high efficiency [fast MM (Modified M-estimator), fast τ-estimator and HBR (High breakdown rank-based)], and one to handle small samples (Distance-constrained maximum likelihood (DCML)). We include the fast MM and fast τ-estimators because we use the fast-robust bootstrap (FRB) for MM and τ-estimators. Our objective is to compare the predictive performance on a real data application using OLS (Ordinary least squares) and to propose alternatives by using 7 different robust estimations. We also run simulations under various combinations of 4 factors: sample sizes, percentage of outliers, percentage of leverage and number of covariates. The predictive performance is evaluated by cross-validation and minimizing the mean squared error (MSE). We use the R language for data analysis. In the real dataset OLS provides the best prediction. DCML and popular robust estimators give good predictive results as well, especially the Huber M-estimator.**
In simulations involving 3 predictors and n=50, the results clearly favor fast MM, fast τ-estimator and HBR whatever the proportion of outliers. DCML and Tukey M are also good estimators when n=50, especially when the percentage of outliers is small (5% and 10%). With 10 predictors, however, HBR, fast MM, fast τ and especially DCML give better results for n=50. HBR, fast MM and DCML provide better results for n=500. For n=5000 all the robust estimators give the same results independently of the percentage of outliers.
If we vary the percentages of outliers and leverage points simultaneously, DCML, fast MM and HBR are good estimators for n=50 and p=3. For n=500, fast MM, fast τ and HBR provide better results. For n=5000, the 7 robust estimators give the same results. When there are p=10 covariates, fast τ, fast MM, HBR and DCML provide better results for n=50 and n=500. For n=5000, all the robust estimators provide the same results.

Keywords – ***efficiency, high-breakdown, outliers, regression, robust estimators.***

I. INTRODUCTION

Linear regression is one of the most widely-used methods in statistics. OLS estimator is often the default estimator.
However, OLS may be biased by the presence of influential outlying observations. Robust estimators can remain unaffected and provide results that are resistant to influential outlying points.
What constitutes an outlier depends on context. For example a regression outlier is an unusual value of y given the xs. Outlying observations may be errors, or they could have been recorded under exceptional circumstances (Rousseeuw and Hubert, 2011). Typically robust techniques reduce the influence of influential outlying observations on the estimator.
A researcher needs to consider many factors when fitting a robust regression model, for example robustness, efficiency, ease of computation, and transparency. The field of robust statistics is evolving rapidly, so practitioners will want to keep abreast of developments. In our opinion, the priorities for a robust estimator of linear regression are (a) computability, (b) an asymptotic theory for a fairly wide class of distributions, (c) good asymptotic efficiency and (d) a high breakdown point, i. e., mitigation of bias due to the most common types of outliers.

II. LITERATURE REVIEW

In the context of linear regression, we distinguish 2 types of outliers:
- vertical or y-outliers: these are outliers in the response variable;
- leverage points or x-outliers: these are outliers in the space of the explanatory variables.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

These 2 types of outliers have the potential to be influential. A data point is influential if it unduly influences any part of the regression (predicted response, estimated beta coefficients, hypothesis tests) and the regression's predictive accuracy. Y-outliers and x-outliers are not a problem in themselves, but they become a problem if they are influential.

In multiple linear regression influential observations can be identified using Cook's distance which combines leverage and outlyingness. It is a deletion diagnostic that measures the influence of the ith observation if it is removed from the sample. Cook's distance is presumably more suitable for evaluating prediction accuracy than some other influence measures since it shows how far, on average, the *predicted y-value* will move if the observation in question is dropped.

2.1 The Cook's distance

If the i-th observation is deleted, Cook's distance, denoted D_i, is

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) \times MSE} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]$$

where y_i *represents* the *i*th response value, $\hat{y}_i$ represents the *i*th fitted response value, there are k regression parameters, and h_{ii} *is* the leverage.

D_i depends on both the residual, e_i and the leverage, h_{ii}. That is, both the *x* value and the *y* value of the data point play a role in the calculation of Cook's distance. These are guidelines for deciding when D_i is large enough to warrant treating a data point as influential:

- If D_i is greater than 0.5, then the i^{th} data point is worthy of further investigation as it may be influential.
- If D_i is greater than 1, then the i^{th} data point is quite likely to be influential.
- Or, if D_i sticks out like a sore thumb from the other D_i values, it is almost certainly influential (available at https://online.stat.psu.edu/stat462/node/173/).

In our opinion, these guidelines should not be applied rigidly. One should look for cases that stick out and attend first to the largest one (Fox, 1991). Robust techniques are really useful in the presence of influential observations because they downweight these points.

2.2 Robustness properties

The influence function and the breakdown point are the best known measures of robustness. The breakdown point is a global measure of robustness, giving the highest proportion of outliers for which the estimator's bias remains finite (Hubert et al., 2008; Rousseeuw et al., 2004; Yohai, 1987; Yohai and Zamar, 1988; Ruckstuhl, 2016). (But the bias, although finite, is not guaranteed to be small.) In contrast, the influence function, introduced by Hampel (1971), is a local measure of robustness; it evaluates infinitesimal amounts of contamination due to the effect of a single outlier. Huber (1983) states that the bounded influence safeguards against the bias by an infinitesimally small amount of asymmetric contamination since it minimizes the asymptomatic variance at the usually normal model, subject to a bound on the supremum of the model's influence function. Maronna, Martin and Yohai (2006) have recently emphasized that there is a fundamental trade off between high statistical efficiency and small bias in the presence of outliers.

The influence function represents a sort of worst case scenario since in general an estimator with bounded influence may or may not be less biased. The MM-estimator has a 50% breakdown point, the highest possible for a linear, affine-equivariant estimator. On the other hand, our experience is that a bounded influence estimator can tolerate up to 5%-10% of outliers and/or bad leverage points, but generally not more. Therefore we emphasise the importance of the breakdown point criterion for robust models of linear regression.

2.3 Efficiency properties

Statistical efficiency is the sample size required to achieve a given precision compared to OLS at Gaussian distributions. For example the S-estimator, a robust alternative to the standard deviation, has an efficiency of only 28.7%, which means that the S-estimator requires a sample 3.48 (=1/0,287) times as large as the standard deviation needs to achieve the same sampling variance when the data are Gaussian. This illustrates the trade off: robustness is achieved at the cost of a larger sample size (Siegel, 1982).

2.4 Computation properties

Computational complexity describes how the time required to compute an estimate grows as the sample size increases. The complexity is generally expressed as a polynomial (when possible), an exponential, or other function, with lower-order polynomials being asymptotically faster than higher-order polynomials and exponentials. The constant term is generally omitted when describing computational complexity and quadratic complexity will be asymptotically faster than cubic complexity; but for a given sample size the cubic algorithm might be faster depending on the constants (Siegel, 1982).

2.5 Relationship between robustness, efficiency and computability

Increasing efficiency may often compromise robustness. In the case of the MM-estimator, the breakdown point (robustness) is maintained (50%), but the maximum possible bias increases. Conversely, one could use an 80% efficient MM estimator to lower the maximum bias while retaining the 50% breakdown point (Hubert et al., 2008; Rousseeuw et al., 2004). In other words, with a given breakdown point (e.g., 50%), we may increase the efficiency of MM, HBR or τ-estimators; but the price is likely to be a larger bias. In the presence of influential outliers and leverage observations –which can be difficult to detect using scatter plots—the estimated coefficients could be far from the true parameter values. However, if one is

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

confident that the actual fraction of influential outliers and leverage points is not large, the breakdown point can be decreased (i.e. from 50% to 30%), making MM, HBR or τ-estimators easier to compute. Statisticians are currently developing a test that helps set the maximum attainable efficiency for the MM-estimator before the bias becomes excessively high.

Moreover it is impractical to compute exact solutions for high breakdown methods like MM, HBR or τ-estimators when the data set contains many regressors. Instead resampling algorithms generate approximate solutions which have been shown to perform acceptably in simulations and on real data sets, but the algorithms' asymptotic properties (e.g., consistency) have not been demonstrated (Olive, 2017). On the other hand, the outlook for speedier and more accurate computation is good. Already the "fast" algorithms are widely available and have been used in this paper to compute the fast-MM and fast τ-estimators. (HBR does not have a fast version). And more recently "deterministic" procedures that dispense with the resampling algorithm have been included in popular statistics platforms, for example the packages 'DetMCD' and 'DetR' for users of R.

2.6 Brief presentation of 7 robust regression estimators

Let us now examine in more detail several of the most commonly used robust estimators: Huber's M-estimator, Tukey's bisquare M-estimator and least absolute deviations as well as the estimators combining high breakdown and high efficiency like fast MM, fast τ-estimator and HBR. We will also review a robust estimator adapted for small sample size (N<100), the DCML.

2.6.1 OLS, Huber and Tukey-M estimators

M-estimates (Huber, 1981) are solutions of the normal equation with appropriate weight functions. They are resistant to unusual outliers, but sensitive to high leverage points.

Using Huber M-estimator the outliers are downweighted but not to zero. In contrast, redescending M-estimators like the Tukey's bisquare M-estimator give extreme observations zero weight (Welsh, 1996). Using a standard notation, we start with the familiar OLS estimator : $\hat{\beta}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} r_i(\beta)^2$ where the function $\arg\min$ returns a vector β for which the sum is minimal, in other words $\arg\min$ refers to minimizing the argument (sum of squared residuals). Huber and Tukey M-estimators are as follows : $\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\frac{r_i(\beta)}{\hat{\sigma}(\beta)})$ where ρ is an appropriate function, which might be squared around zero, but bounded for large (absolute) values with $\sum_{i=1}^{n} \psi(\frac{r_i(\beta)}{\hat{\sigma}(\beta)}) x_{ij} = 0$ (Welsh, 1996). Here is the system of weighted equations $\sum_{i=1}^{n} \omega_i (y_i - x_i^T \beta) x_i = 0$. For OLS regression we get weights for the residuals ω_i=1 for all i. In contrast, the robust methods (Huber and Tukey) attempt to downweight outliers.

2.6.2 L1 estimator

The Least absolute deviation (LAD) also called L1 estimates due to the L1 norm used, was developed by Roger Joseph Boscovich in 1757, nearly 50 years before OLS estimation (Nevitt and Tam, 1998; Birkes and Dodge, 1993). Like Huber's M estimator, the L1 regression is robust with respect to y-outliers but not to leverage points (Gschwandtner and Filzmoser, 2012). Here is the L1 equation: $\hat{\beta}_{L1} = \arg\min_{\beta} \sum_{i=1}^{n} |r_i(\beta)|$. These 3 estimators have 1/n breakdown point. So, as n increases the breakdown point goes towards zero.

2.6.3 MM estimator

The MM estimation was introduced by Yohai (1987) and is a combination of high breakdown value and efficient estimation. The MM-estimate can be found by a three-stage procedure. In the first stage, it computes an initial consistent estimate $\hat{B}_0$ with high breakdown point but possibly low normal efficiency. In the second stage, it computes a robust M-estimate of scale $\hat{\sigma}$ of the residuals based on the initial estimate. In the third stage, it finds an M-estimate $\hat{B}$ starting at $\hat{B}_0$. In practice, LMS (Least median of squares) $\hat{\beta} = \arg\min_{\beta} Med\{(y_i - x_i^T \beta)^2\}$ or S-estimate (Scale estimate) - $\hat{\beta} = \arg\min_{\beta} \hat{\sigma}(r_1(\beta),...,r_n(\beta))$ where $r_i(\beta) = y_i - x_i^T \beta$ and $\hat{\sigma}(r_1(\beta),...,r_n(\beta))$ is the scale M-estimate - with Huber or bisquare functions is typically used as the initial estimate $\hat{B}_0$(Yu et al., 2014). Here is the equation for MM-estimator: $\hat{\beta}_{MM} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\frac{r_i(\beta)}{\hat{\sigma}})$.

Let $\hat{\beta}_0$ be an S estimator, and let $\hat{\sigma}$ be the corresponding M estimator.

2.6.4 τ-estimator

The τ-estimator (Yohai and Zamar, 1988) has been shown to have a good balance between robustness and efficiency. In τ-estimation, unlike MM estimation, there is no global precalculated estimate of scale. Both B and σ are iteratively and alternatively estimated. In the general procedure the function ρ_0 used to estimate scale is chosen to give the

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

maximum breakdown point for regression estimates. On the other hand the function ρ_1 used for estimation of B is chosen to give high efficiency (0.95 is suggested as high) (Riani et al., 2014). Although the τ-estimate of regression is equivalent to an M estimate, it is defined as the minimizer of a particular robust and efficient estimate of the scale of the residuals, the τ-scale estimate $\hat{\sigma}_T$. The equation is as follows: $\hat{x}_m^T = \arg\min_x \hat{\sigma}_\tau^2(r(x))$.

2.6.5 HBR estimator

Some prior simulations have shown that the HBR estimator (Kloke and McKean, 2012; Chang et al., 1999) performs with fair efficiency compared to MM and τ-estimators, but it does not deal well with bad leverage points. Data points which are outliers are downweighted. If all the weights are 1 then HBR is the Wilcoxon norm. High breakdown rank (HBR) estimates are based on a weighted Wilcoxon pseudo-norm. The weights for the HBR estimates make use of the high breakdown minimum covariance determinant (MCD), which is an ellipsoid in *p*-space that covers about half of the data and yet has minimum determinant (McKean and Kloke, 2014). The HBR estimation is a weighted Wilcoxon dispersion function given by : $\|v\|_{HBR} \sum_{i<j} b_{ij} |\upsilon_i - \upsilon_j|$ where $b_{ij} \geq 0$ and $b_{ij} = b_{ji}$. The HBR estimator of β minimizes this objective function, which we denote by $\hat{\beta}_{HBR} = \arg\min \|y - X\beta\|_{HBR}$. The HBR estimator minimizes the convex function $D_{HBR}(\beta)$ which is defined by : $D_{HBR}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij} |(Y_i - Y_j) - (x_i - x_j)'\beta|$. HBR, fast MM and τ-estimators have 50% breakdown point.

2.6.6 DCML estimator

The DCML estimator's breakdown point is at least that of the initial estimator, specifically an MM-estimator which in turn is based on a S-estimator (Maronna et al. 2018). So the breakdown point of the DCML is that of the S-estimator: 50%. The DCML estimator is asymptotically fully efficient and shows a very good balance between efficiency and robustness (Maronna and Yohai, 2015a). Let $\hat{\beta}_0$ be a highly robust estimator, typically an MM-estimator and $\hat{\beta}$ be $\arg\min_\beta L(\beta | r_1, r_2, ... r_n)$. The DCML optimization is : $\min_\beta \sum_{i=1}^{n} r_i^2(\beta)$ subject to $d_{KL,\hat{V}_x}(\hat{\beta}_0, \beta) \leq \delta$ where KL= Kullback-Leibler distance (the natural function from a "true" probability distibution to a "target" probability distribution) and $\hat{v}_x$=(robust estimate).

2.7 In case of small sample size, which estimator to use and why?

There are a few regression estimators adapted for small sample size (N<100). Besides the DCML there is a family of robust regression estimators called bounded residual scale estimators (BRS estimators) (Smucler and Yohai, 2017). They are simultaneously highly robust and highly efficient for small sample size (Smucler and Yohai, 2015). The DCML estimator (Maronna and Yohai, 2015b) is recommended for following reasons: (a) inference (i.e., confidence intervals) is better justified; (b) the estimator can be computed somewhat faster and (c) it has a simpler and more intuitive definition. Thus Maronna et al. (2018) use n=50 for DCML.

2.8 Breakdown point as the criterion for robustness

We emphasize that all these 7 estimators are robust against influential outliers but not necessarily against influential high leverage points. Since our focus is on the predictive performances of these 7 estimators and OLS on real data and in simulations study, we concentrate on the estimators' breakdown point instead of their influence function; in this we follow Davies (1993 ; 1994), who proposes for desirable properties of estimators mainly robustness properties and not efficiency. Nevertheless, we will explore whether robust estimators combining high breakdown and high efficiency (MM, τ, HBR and DCML) really provide better predictive results than classical OLS and other procedures with low breakdown points (Huber, Tukey and L1).

For that we will use a real dataset and run simulations. Concerning the simulations, we first vary only the percentage of outliers and in a second step we vary the percentage of outliers and of leverage simultaneously.

Table 1 below summarizes the breakdown point of the 8 different regression estimators.

Table 1**:** Breakdown point of the different estimators

Estimators	**OLS**	**Huber-M**	**Tukey-M**	**L1**	**fast MM**	**fast τ**	**HBR**	**DCML**
breakdown point	1/n	1/n	1/n	1/n	50 %	50 %	50%	50%

III. MATERIAL AND METHODS

3.1 R software and packages

In this article we use the R software (R Core Team, 2017), the "robustbase", the "MASS", the "quantreg", the "RobPer", the "devtools", the "RobStatTM" and the "boot" packages.

3.2 Application to ATTICA epidemiological data

The dataset ATTICA is an epidemiological study of 731 men and women with no clinical evidence of chronic disease or acute inflammation, aged 18 - 84 years, randomly selected from all areas of the Athens metropolitan region in Greece (Pitsavos et al., 2003). I really thank my colleague, Professor

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

Panagiotakos, who provided this dataset. As this dataset has many medical and health variables it might contain quite a few influential observations. We want to be sure that OLS is the best predictive fit for this dataset as compared to more robust estimators.
Information about age (in years), sex (male/female), systolic (sbp) and diastolic (dbp) blood pressure (in mmHg), physical activity status (at least adequately active/inactive), current smoking (yes/no), educational level (educat) measured by the years of schooling, HDL cholesterol level (HDLC), total cholesterol (TC), body mass index (BMI in Kg/m2) calculated as body weight divided by standing height, fasting blood glucose levels (in mg/dL) and diabetes status (based on glucose levels and/or anti-diabetic medication), was retrieved from the entire database and related to high sensitivity C-reactive protein (CRP) levels (dependent outcome) in order to examine the role of metabolic markers on the chronic systemic inflammation process, in light of confounding factors.

The fitted model is :
$C\text{-}RP = b_o + b_1 X\ age + b_2 X\ sex + b_3 X\ bmi + b_4 X\ educat + b_5 X\ PhysActlevel + b_6 X\ smokingcurrent + b_7 X\ dbp + b_8 X\ Diabetes + b_9 X\ glucose + b_{10} X\ TC + b_{11} X\ HDLC + error$

Table 2 displays some descriptive statistics, the frequency tables of the variables retained (the dependent variable C-RP and all the other explanatory variables) and the graph of the dependent variable.

Table 2**:** Descriptive statistics of C-RP dependent variable and all other independent variables

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age [integer]	Mean (sd) : 43.9 (13.6) min < med < max: 18 < 43 < 84 IQR (CV) : 18 (0.3)	64 distinct values		731 (100%)	0 (0%)
2	sex [factor]	1. man 2. woman	380 (52.0%) 351 (48.0%)		731 (100%)	0 (0%)
3	bmi [numeric]	Mean (sd) : 26 (4.7) min < med < max: 13.9 < 25.4 < 52.5 IQR (CV) : 6.3 (0.2)	517 distinct values		731 (100%)	0 (0%)
4	educat [integer]	Mean (sd) : 12.4 (3.9) min < med < max: 0 < 12 < 36 IQR (CV) : 4 (0.3)	20 distinct values		731 (100%)	0 (0%)
5	PhysActlevel [factor]	1. active 2. adequately activite 3. inactive 4. very active	212 (29.0%) 44 (6.0%) 31 (4.2%) 444 (60.7%)		731 (100%)	0 (0%)
6	smokingcurrent [factor]	1. no 2. yes	318 (43.5%) 413 (56.5%)		731 (100%)	0 (0%)
7	sbp [integer]	Mean (sd) : 122.1 (19) min < med < max: 80 < 120 < 195 IQR (CV) : 20 (0.2)	30 distinct values		731 (100%)	0 (0%)
8	dbp [integer]	Mean (sd) : 78.8 (12) min < med < max: 40 < 80 < 130 IQR (CV) : 15 (0.2)	25 distinct values		731 (100%)	0 (0%)
9	Diabetes [factor]	1. no 2. yes	43 (5.9%) 688 (94.1%)		731 (100%)	0 (0%)
10	glucose [integer]	Mean (sd) : 92.6 (21.2) min < med < max: 49 < 90 < 275 IQR (CV) : 16.5 (0.2)	91 distinct values		731 (100%)	0 (0%)
11	TC [numeric]	Mean (sd) : 192.3 (38.1) min < med < max: 84.2 < 189.6 < 353.6 IQR (CV) : 51 (0.2)	189 distinct values		731 (100%)	0 (0%)
12	HDLC [numeric]	Mean (sd) : 48.5 (14.2) min < med < max: 20 < 47 < 154 IQR (CV) : 19 (0.3)	72 distinct values		731 (100%)	0 (0%)
13	crp [numeric]	Mean (sd) : 2 (2.5) min < med < max: 0 < 1.1 < 15.9 IQR (CV) : 1.9 (1.2)	356 distinct values		731 (100%)	0 (0%)

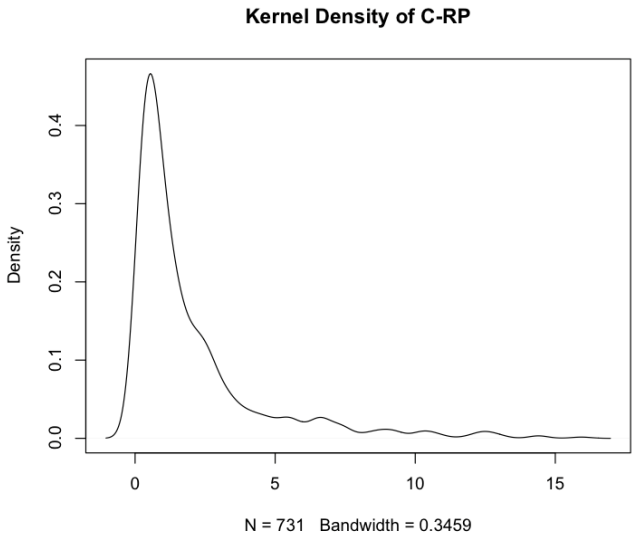

Figure 1: The density plot of the CR-P dependent variable

The dependent variable CR-P is highly right skewed. This skeweness may be due to the presence of some y-outliers as can be seen in Figures 3 and 5. No interaction terms were included in the regression because they did not improve predictive accuracy.

3.3 Model evaluation

3.3.1 Cross-validation and error evaluation metric

The evaluation metric retained for predictive accuracy is the Mean squared error (MSE). $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

where n is the number of data points, y_i represents observed value and $\tilde{y}_i$ represents predicted values.
To assess prediction accuracy, the holdout cross-validation (training set=66% and testing set=34%) is run 1000 times to examine the variability of the out-of-bag accuracy value. The cross-validation predictive accuracy values are averaged and then the model that minimizes the Mean squared error (MSE) is deemed the best (Varin, 2020). We also calculate the 95% BCa bootstrap confidence intervals around the MSE value (Efron, 1987; DiCiccio and Efron, 1996).

3.4 Unbiased prediction

Minimizing anything else than the mean squared error (MSE) may lead to biased prediction (Varin, 2020). Indeed, if we want an unbiased prediction, then the MSE is the only criterion that will be minimized by the true value in expectation**;** and we will simply need to accept that skewed or highly variable conditional outcomes have a strong influence on the conditional expectation. We note that the possible inflation in MSE is due to the presence of outliers (Varin, 2020).

IV. RESULTS AND DISCUSSIONS

In this section an application based on the aforementioned data will be performed to compare OLS and the 7 different robust regressions estimators in their predictive performance.

4.1 Residuals diagnostic plots of linear model

No (multi)collinearity is found, the variance inflation factor is not higher than 2 and the regression function is linear. There is a problem of heteroscedasticity of the variance and of non-Gaussian distribution of the residuals, a strong positive skew in the residuals (Figure 2).

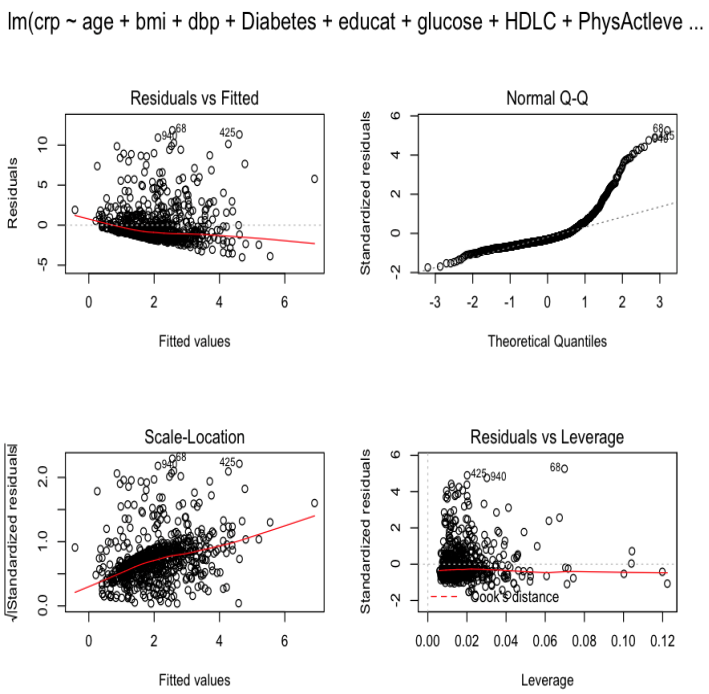

Figure 2: OLS standard diagnostic plots. Dots indicate observed individuals participated in the ATTICA study

Unbiased point predictors require only that the errors have zero mean (James et al., 2017), which is guaranteed by the inclusion of intercepts in the models. We use the BCa nonparametric bootstrap (Efron, 1987; DiCiccio and Efron, 1996) to generate confidence intervals and prediction intervals that don't depend on mean/variance assumptions. The BCa bootstrap is second-order correct and works with skewed distributions and biased statistics (Varin, 2020).

Figure 3 shows many outliers, many leverages and just a few outliers and leverages, but it is not known *a priori* if any of them is influential. According to Figure 4, the Cook's distance plot reveals one clearly influential observation (n°42). We certainly advise the use of graphics and the scrutiny of the points with "values of D that are substantially larger than the rest"; it is indeed both an influential y-outlier and leverage (Figure 3). More precisely, the residual of this point (11.84) is 5 times larger than the residual standard deviation of the model (2.34). The rest of the observations cannot be considered influential points.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

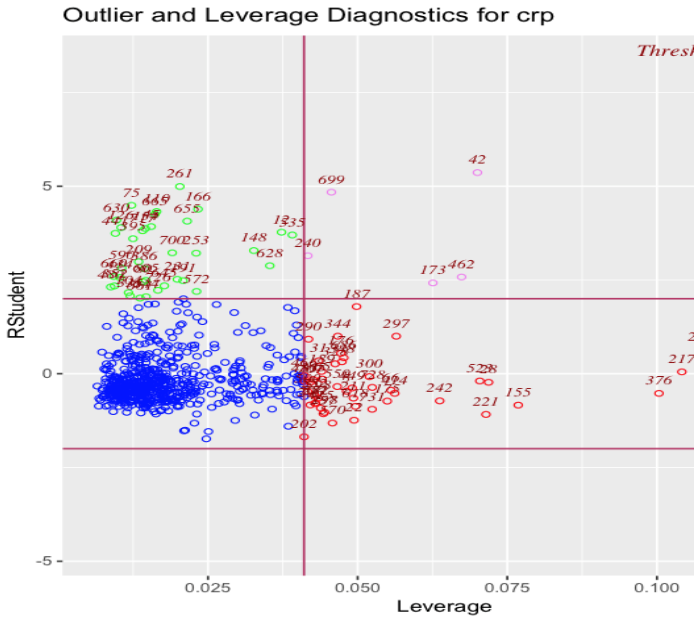

Figure 3: Outlier and leverage diagnostic plot for CRP variable. Dots indicate observed individuals participated in the ATTICA study

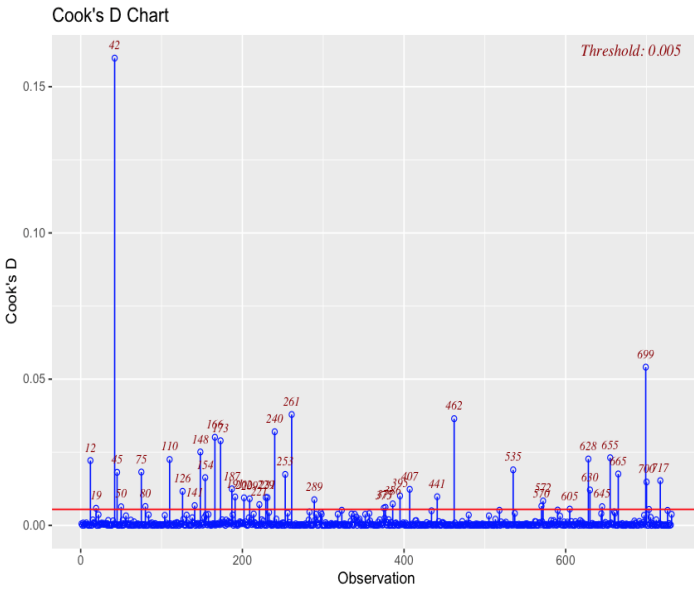

Figure 4: Cook's distance plot showing the influential observations; numbers indicate observed individuals participated in the ATTICA study

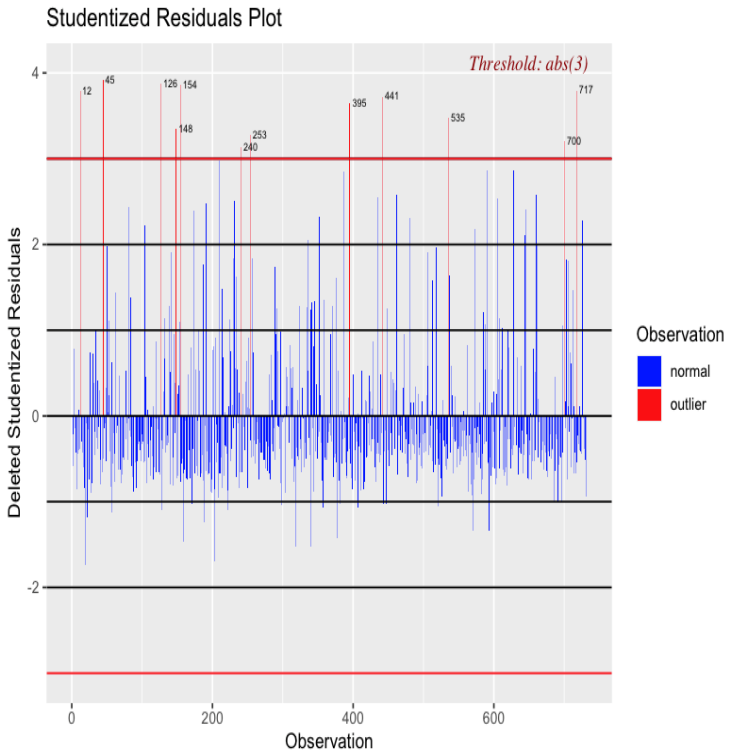

Figure 5: Deleted studentized residuals plot showing the y-outliers; numbers indicate observed individuals participated in the ATTICA study

4.2 Discussion

Since point 42 is the salient influential observation, a high breakdown estimator isn't required. However, a classical robust estimator like Huber, Tukey or L1 could be worthwhile. Table 3 shows that the estimators with the best prediction results (the lower MSE) are OLS and the popular estimators: Huber, Tukey, and L1. DCML is good as well but not the other high breakdown point estimators. In this case, with only one influential observation among 731, OLS regression still provides the best predictive result.

Table 3 : Mean squared error (MSE) value for the 7 different robust estimators and OLS. In bold the lower MSE. In brackets the 95% BCa bootstrap confidence intervals. Number of iterations B=5,000

Estimators	**DCML**	**MM**	**Huber**	**Tukey**	**L1**	**fast τ**	**HBR**	**OLS**
MSE	6.14 (4.9; 7.6)	6.39 (4.9; 7.9)	5.75 (4.8; 7.1)	6.32 (5.1; 7.9)	6.13 (5; 7.6)	7.06 (5.7; 8.7)	6.58 (5.1; 8)	**5.4 (4.4; 6.7)**

OLS and standard robust estimators perform better, the use of highbreakdown estimators is absolutely counterproductive here. But are the differences meaningful? We compute and

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

present in brackets the 95% confidence intervals for the MSE using the nonparametric second-order accuracy BCa bootstrap with 5000 replications. Given the overlapping intervals, there is thus no statistical evidence that any estimator fits better and provides more accurate predictive results than the others at the 5% level (Varin, 2016).

V. SIMULATION EXPERIMENTS

Simulation studies may be a better alternative for objectively comparing the predictive performances of these 7 robust regression estimators and OLS. We want to figure out which estimator performs better (minimizing MSE) in what circumstances (Varin, 2020).

5.1 Design of the simulation

We thus consider four different factors: n, sample size from 50 to 5000; the percentage of outliers, from 5% to 30%; the percentage of leverage from 5% to 30% and the number of covariates, p, 3 and 10. In the first step only the percentage of outliers varies, and there are no leverage points. The second step varies the percentage of outliers and the percentage of leverage simultaneously.

More precisely, for the simulated model with 3 variables, 2 of them (x and a) follow a uniform distribution on intevals $[0,5]$ and 1 variable (z) follows a Normal distribution $N(2;3)$. The theoretical model is the following: $y = 0.1 \times b - 0.5 \times z - a + 10$.

Regarding the simulated model with 10 variables, 5 variables (x, a, c, e and g) are uniformaly distributed on $[0,5]$ intervals and 5 (z, b, d, f and h) are normally distributed $N(2;3)$. The theoretical model is the following:

$y = 0.1 \times b - 0.5 \times z - a + 1.5 \times c + 2 \times d + 3 \times e - 4 \times f + 3.5 \times g - 4.5 \times h + 2.5 \times i + 10$.

We start by creating the values of covariates according to predefined distributions, i.e. uniform $[0,5]$ and Normal $N(2;3)$ distributions. Then we calculate "y" according to the theoretical models, one with 3 and the other with 10 covariates. Finally, we generate "y_obs" in our case according to a Normal distribution (mean equals to our theoretical model and standard deviation equals 0.1). In order to add outliers, a proportion of "y_obs" (5%, 10% and 30%) has been replaced by the same Normal distribution but with a standard deviation of 0.5, or 5 times greater. For the leverage observations, the Normal distribution $N(2;3)$ of a covariate has been replaced with a standard deviation of 6, or 2 times greater.

While fitting the model with 10 variables, we will not be able to present the 95% BCa confidence intervals because too many replications are needed to calculate the acceleration constant ”a”. Only the MSE is reported.

In the appendix we present the 24 tables. In tables 4 through 9 with the sample sizes (n=50, n=500, n=5000) we cross the results for the number of covariates (p=3; p=10) and the percentage of outliers (5%; 10% and 30%). There is no leverage. In table 10 to 18 we cross the percentages of outliers with the percentages of leverages (5%, 10% and 30%) simultaneously for p=3 and in table 19 to 27 for p=10. The MSE are reported with the smaller MSE numbers are in bold print. In brackets we report the 95% BCa bootstrap confidence intervals based on 1,000 iterations.

5.2 Discussion of the simulation results

Considering only the outliers, the simulation results are clearly in favour of fast MM, fast τ-estimator and HBR whatever the percentage of outliers when there are 3 predictors in the model. DCML and Tukey are good estimators in case of small sample size (n=50) especially when the percentage of outliers is small (5% and 10%). Let us remark that when the outliers are not too many (5%) and the sample size is small (n=50), the OLS estimator provides good results for p=3 and p=10.

More precisely if the number of predictors is 3, for n=50 the fast τ-estimator and the fast MM are better when the percentage of outliers is high (30%).

For n=500, the fast MM and fast τ-estimator are good estimators when the percentage of outliers is low (5% and 10%). For higher percentages of outliers (30%), HBR is better.

For n=5000, the fast MM and fast τ-estimator provide better results; and the fast τ-estimator is good when the percentage of outliers is 10%.

However, with 10 predictors in the model, HBR, fast MM, fast τ-estimator and especially DCML give better results for n=50. HBR, fastMM and DCML provide better results for n=500. For n=5000 they all give the same results.

More precisely, for n=50 HBR and DCML are good estimators when the percentage of outliers is low (5% and 10%). Fast MM is good for higher percentages of outliers (30%).

For n=500, DCML is good for small percentage of outliers (5%). For large percentages of outliers, fast MM and HBR are good estimators.

For n=5000, they all give the same results. As the sample size increase (n=5000) and the model is fitted with many predictors, in our case p=10, the robust estimators give equivalent MSE. Let us remark that OLS does not perform well.

Now, crossing the percentage of outliers and the percentage of leverages, we remark that when the sample size is small (n=50), whatever the percentage of outliers (5%, 10% or 30%) with 5% and 10% of leverage observations, DCML provides the best results. With 30% of leverage and small percentages of outliers (5% and 10%) DCML is still the best estimator. With 30% of outliers, fast MM and HBR are good estimators.

For n=500 whatever the percentage of outliers with 5% of leverage, fast MM and HBR provide the best results. With 10% of leverage, with 5% and 10% of outliers, fast MM and fast τ are good estimators. With 30% of outliers, HBR gives the best predictive results. With 30% of leverage and a small amount of outliers (5% and 10%), fast MM, fast τ and HBR

provide the best results. With 30% of outliers, HBR is the best estimator. For n=5000, the 7 robust estimators give the same results. Again here, let us remark that OLS does not perform well.

Considering p=10 covariates, 5% of leverage and 5% of outliers in the model, for n=50 fast τ and HBR are the best estimators. With 10% of outliers, DCML is the best and with 30% of outliers it is the HBR estimator. With 10% of leverage and 5% of outliers, fast MM and HBR are the best estimators. With 10% and 30% of outliers, HBR gives the best results. With 30% of leverages, 5% and/or 30% of outliers, DCML and HBR provide the best results. With 10% of outliers fast MM provide better results.

For n=500, with 5% of leverages and 5% of outliers, the fast MM estimator is good. With 10% of outliers, fast τ is the best estimator. With 30% of outliers fast MM and HBR are the best estmators. With 10% of leverages and 5% of outliers, fast MM is the best estimator. With 10% of outliers fast τ and DCML are the best ones. HBR and DCML are good estimators for 30% of outliers. With 30% of leverages and 5% of outliers fast MM is the best one. Fast τ is good when the percentage of outliers is 10% and DCML is good for 30% of outliers. Again here, OLS does not perform well.

For n=5000 the 7 robust estimators provide the same results. OLS does not provide good results.

We emphasize once more that according to the 95% nonparametic BCa bootstrap confidence intervals no one of these 7 robust estimators provides better predictive results at 5% level.

VI. CONCLUSION

In this article we briefly describe and compare 7 robust regression and OLS estimators while fitting linear regression models. MSE is our criterion. Based on the real data set results, we can say that OLS and the 3 popular estimators perform better than the high breakdown estimators. The main reason is that there is just one influential observation in 731 samples. On the contrary, our simulations experiments demonstrated that the 4 high breakdown estimators have overall best predictive outcomes. In terms of prediction performances especially with a large percentage of outliers (30%) and whatever the percentage of leverage observations, the high breakdown estimators are quite impressive. OLS and the popular robust procedures – Huber, Tukey and L1 – are less attractive due especially to their low breakdown point in the presence of influential leverage. However we reiterate that according to their overlapped 95% confidence intervals, no one of these 7 robust estimators performs better at level 5% in any situation whether in the real data set or in simulations experiments. These results clearly show that the most important thing is to choose a robust estimator that we know how to use, no need to look for new unknown robust estimators. In the end they appear to provide about the same predictive performance.

In conclusion, the selection of a robust estimator for linear regression should balance robustness, efficiency and computability. From this point of view, fast MM and the other highbreakdown estimators are a good choice. But unless the influential outliers are really far from the majority of the data, fast MM and the other highbreakdown estimators may not be optimal predictors. Indeed, when the majority of the data overlap in some sense with the influential outliers, we have often observed that fast MM and the other highbreakdown estimators can suffer from masking, meaning that one or more outliers are labelled as good cases by some criterion. Another negative point to the high breakdown estimators is their computation time using R software as compared to the more popular estimators.

Acknowledgement

The author thanks Professors Jean-Pierre Asselin de Beauville, Louis-Marc Bourdeau and Eric Blankmeyer for constructive and valuable comments, guidance of the paper presentation and careful reading.

Conflict of interest

The author declares no conflict of interest.

VII. REFERENCES

1 Birkes, D., Dodge, YD. (1993). Alternative methods of regression. Wiley: New-York.

2 Chang, W., Mckean, J., Naranj, J. and Sheathe, S. (1999). High-breakdown rank regression. *Journal of the American Statistical Association*, 94, 205-219.

3 Davies, PL. (1993). Aspects of robust linear regression. *Ann. Statist*, (21), 1843-1899.

4 Davies, PL. (1994). Desirable properties, breakdown and efficiency in the linear regression model. *Statist. Probab. Lett.*, (19), 361-370.

5 DiCiccio, TJ., Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, Vol. 11, (3), 189-228.

6 Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, Vol. 82, (397), 171-185.

7 Fox, J. (1991). *Regression diagnostics: An introduction*. Sage Publications.

8 Gschwandtner, M., Filzmoser, P. (2012). Computing Robust Regression Estimators: Developments since Dutter (1977). *Austrian Journal of Statistics*, 41(1), 45–58.

9 Hampel, FRA. (1971). General qualitative definition of robustness. *Annals of Mathematical Statistics*, 42(6), 1887-1896.

10 Huber, PJ. (1981). Robust Statistics. John Wiley : New York.

11 Huber, PJ. (1983). Minimax Aspects of Bounded-Influence Regression. *Journal of the American Statistical Association*, 78 (381), 66-72.

12 Hubert, M., Rousseeuw, PJ., Van Aelst, S. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science*, 23(1), 92-119.

13 James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R. New York: Springer, 7th Printing.

14 Kloke, JD., McKean, JW. (2012). Rfit: Rank-based Estimation for Linear Models. *The R Journal*, Vol. 4/2.

15 Maronna, RA., Martin, RD. and Yohai, VJ. (2006). Robust Statistics, Theory and Methods. Wiley Series in Probility and Statistics. John Wiley & Sons, Ltd.

16 Maronna, RA., Yohai, VJ. (2015a). Robust and efficient estimation of high dimensional scatter and location. *ArXiv:1504.03389v1.mathST*, 1-33.

17 Maronna, RA., Yohai, VJ. (2015b). High finite-sample efficiency and robustness based on distance-constrained maximum likelihood. *Journal Computational Statistics & Data Analysis*, 83, 262-274.

18 Maronna, RA., Martin, RD., Yohai, VJ. and Salibian-Barrera, M. (2018). Robust Statistics. Theory and Methods, Wiley, 2nd Edition.

19 McKean, JW., Kloke, JD. (2014). Efficient and adaptive rank-based fits for linear models with skew-normal errors. *Journal of Statistical Distributions and Applications*, (1), 1-18.

20 Nevitt, J., Tam, HPA. (1998). Comparison of robust and nonparametric estimators under the simple linear regression model. *Multiple Linear Regression Viewpoint*, 25, 54-69.

21 Olive, D. (2017). Robust Multivariate Analysis. Springer.

22 Pitsavos, C., Panagiotakos, DB., Chrysohoou, C. and Stefanadis, C. (2003). Epidemiology of cardiovascular risk factors in Greece: aims, design and baseline characteristics of the ATTICA study. *BMC Public Health*, 3:32.

23 R Core Team. R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ; 2017, URL https://www.R-project.org/.

24 Riani, M., Ceroli, A., Atkinson, AC. and Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics*, (8), 646–677.

25 Rousseeuw, PJ., Van Aelst, S., Van Driessen, K. and Agulló, J. (2004). Robust multivariate regression. *Technometrics*, 46, 293-305.

26 Rousseeuw, P., Hubert, M. (2011). Robust statistics for outlier detection, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 73-79.

27 Ruckstuhl, A. (2016). Robust Fitting of Parametric Models Based on M-Estimation. *WBL Applied Statistics. Robust Fitting Techniques*, 1-64.

28 Siegel, AF. (1982). Robust regression using repeated medians. *Biometrika*, 69(1), 242-244.

29 Smucler, E., Yohai, VJ. (2015). Highly robust and highly finite sample efficient estimators for the linear model. *Modern Nonparametric, Robust and Multivariate Methods*, 91-108.

30 Smucler, E., Yohai, VJ. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, (111), 116-130.

31 Varin, S. (2016). Étude quantitative des données basée sur les statistiques et le logiciel R. Approches descriptive et inférentielle. Hallennes-lez-Haubourdin : The bookedition.com.

32 Varin, S. (2020). Comparing the performances of Generalized additive models, Multivariate adaptive regression splines and polynomial linear models on a real and simulated datasets. *International Journal of Multidisciplinary Sciences and Advanced Technology. Vol. 1, (6), 10-35.*

33 Welsh, AH. (1996). Aspects of Statistical Inference. Wiley series in probability and statistics.

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

35 Yohai, VJ., Zamar, RH. (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.

36 Yu, C., Yao, W. and Bai, X. (2014). Robust Linear Regression: A Review and Comparison, 1-27, from : https://arxiv.org/pdf/1404.6274.pdf

VIII. APPENDIX

Table 4: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (from 5% to 30%) for sample size n=50. The standard deviation of outliers is 5 times greater. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.027 (0.01; 0.07)	0.021 (0.012; 0.05)	0.076 (0.04; 0.143)

Huber	0.029 (0.01; 0.09)	0.07 (0.03; 0.16)	0.14 (0.088; 0.284)
Tukey	**0.012 (0.006; 0.03)**	0.022 (0.014; 0.04)	0.12 (0.06; 0.18)
L1	0.028 (0.01; 0.08)	0.07 (0.03; 0.16)	0.09 (0.05; 0.18)
FastTau	0.027 (0.01; 0.07)	0.07 (0.028; 0.19)	**0.039 (0.025; 0.065)**
HBR	0.027 (0.01; 0.07)	0.06 (0.026; 0.15)	0.12 (0.062; 0.17)
DCML	0.027 (0.013; 0.08)	**0.019 (0.01, 0.04)**	0.048 (0.032; 0.083)
OLS	0.026 (0.012; 0.06)	0.021 (0.012; 0.05)	0.048 (0.031; 0.082)

Table 5: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (from 5% to 30%) for sample size n=500. The standard deviation of outliers is 5 times greater. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.016 (0.01; 0.02)**	0.033 (0.0215; 0.0443)	0.08 (0.06; 0.1)
Huber	0.021 (0.016; 0.03)	0.04 (0.026; 0.053)	0.09 (0.07; 0.12)
Tukey	0.024 (0.014; 0.035)	0.04 (0.026; 0.053)	0.09 (0.07; 0.12)
L1	0.02 (0.015; 0.027)	0.04 (0.03; 0.05)	0.08 (0.06; 0.1)
FastTau	0.02 (0.015; 0.035)	**0.028 (0.02; 0.038)**	0.08 (0.06; 0.1)
HBR	0.018 (0.013; 0.023)	0.04 (0.02; 0.06)	**0.07 (0.05; 0.09)**
DCML	0.02 (0.015; 0.035)	0.03 (0.02; 0.041)	0.08 (0.06; 0.1)
OLS	0.024 (0.018; 0.035)	0.034 (0.025; 0.051)	0.08 (0.06; 0.11)

Table 6: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (from 5% to 30%) for sample size n=5000. The standard deviation of outliers is 5 times greater. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.021 (0.018; 0.022)**	0.034 (0.03; 0.037)	**0.078 (0.073; 0.083)**
Huber	0.023 (0.021; 0.026)	0.035 (0.032; 0.038)	0.082 (0.077; 0.088)
Tukey	0.022 (0.019; 0.024)	0.033 (0.029; 0.036)	0.082 (0.077; 0.088)
L1	0.022 (0.019; 0.024)	0.033 (0.029; 0.036)	0.082 (0.077; 0.088)
FastTau	0.022 (0.02; 0.025)	**0.031 (0.034; 0.037)**	0.084 (0.083; 0.087)
HBR	0.023 (0.022; 0.026)	0.034 (0.032; 0.036)	0.079 (0.083; 0.085)
DCML	0.022 (0.02; 0.025)	0.04 (0.036; 0.041)	0.079 (0.072; 0.085)
OLS	0.026 (0.025; 0.027)	0.045 (0.043; 0.046)	0.085 (0.08; 0.091)

Table 7: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (from 5% to 30%) for sample size n=50. The standard deviation of outliers is 5 times greater. Lower MSE is in bold.

	5%	10%	30%
FastMM	0.017	0.018	**0.026**
Huber	0.018	0.043	0.066
Tukey	0.016	0.022	0.068
L1	0.012	0.02	0.068
FastTau	0.03	0.046	0.053
HBR	**0.009**	0.023	0.051
DCML	0.014	**0.017**	0.042
OLS	0.014	0.021	0.054

Table 8: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (from 5% to 30%) for sample size n=500. The standard deviation of outliers is 5 times greater. Lower MSE is in bold.

	5%	10%	30%
FastMM	0.019	0.039	0.097
Huber	0.019	0.04	0.078
Tukey	0.028	0.041	0.078
L1	0.028	0.041	0.078
FastTau	0.022	**0.029**	0.1
HBR	0.02	0.033	**0.077**
DCML	**0.018**	0.04	0.078
OLS	0.022	0.04	0.089

Table 9: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (from 5% to 30%) for sample size n=5000. The standard deviation of outliers is 5 times greater. Lower MSE is in bold.

	5%	10%	30%
FastMM	0.021	0.033	0.08
Huber	0.022	0.033	0.08
Tukey	0.021	0.033	0.08
L1	0.021	0.033	0.08
FastTau	0.021	0.033	0.08
HBR	0.021	0.033	0.08
DCML	0.021	0.033	0.08
OLS	0.024	0.036	0.084

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 9-23
Published Online March 2021 in IJEAST (http://www.ijeast.com)

Table 10: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.025 (0.008; 0.085)	0.06 (0.02; 0.12)	0.08 (0.04; 0.16)
Huber	0.023 (0.01; 0.06)	0.03 (0.015 ; 0.06)	0.06 (0.04 ; 0.1)
Tukey	0.023 (0.01; 0.06)	0.03 (0.01 ; 0.07)	0.07 (0.03; 0.2)
L1	0.024 (0.01; 0.05)	0.03 (0.01; 0.07)	0.07 (0.03; 0.18)
FastTau	0.027 (0.007; 0.06)	0.03 (0.01; 0.07)	0.07 (0.03; 0.17)
HBR	0.023 (0.01; 0.06)	0.03 (0.01; 0.07)	0.07 (0.04; 0.14)
DCML	**0.022 (0.009; 0.07)**	**0.02 (0.01; 0.03)**	**0.03 (0.02; 0.05)**
OLS	0.023 (0.01; 0.06)	0.03 (0.01; 0.06)	0.06 (0.04; 0.11)

Table 11: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.02 (0.01; 0.06)	0.05 (0.02; 0.12)	0.07 (0.04; 0.16)
Huber	0.02 (0.01; 0.05)	0.03 (0.02 ; 0.06)	0.06 (0.04 ; 0.09)
Tukey	0.02 (0.008; 0.07)	0.03 (0.01 ; 0.08)	0.07 (0.03; 0.2)
L1	0.02 (0.009; 0.06)	0.03 (0.01; 0.08)	0.07 (0.03; 0.18)
FastTau	0.02 (0.01; 0.07)	0.03 (0.01; 0.07)	0.07 (0.04; 0.13)
HBR	0.02 (0.007; 0.05)	0.03 (0.01; 0.07)	0.07 (0.04; 0.12)
DCML	**0.01 (0.008; 0.02)**	**0.02 (0.009; 0.04)**	**0.01 (0.06, 0.26)**
OLS	0.02 (0.007; 0.06)	0.03 (0.01; 0.07)	0.07 (0.03; 0.2)

Table 12: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.02 (0.01; 0.06)	0.06 (0.02; 0.12)	**0.07 (0.04; 0.15)**
Huber	0.02 (0.008; 0.03)	0.03 (0.01 ; 0.06)	0.09 (0.04 ; 0.18)
Tukey	0.02 (0.008; 0.07)	0.03 (0.01 ; 0.07)	0.09 (0.03; 0.2)
L1	0.02 (0.009; 0.07)	0.03 (0.01; 0.07)	0.09 (0.03; 0.18)
FastTau	0.02 (0.008; 0.07)	0.03 (0.01; 0.07)	0.13 (0.1; 0.17)
HBR	0.02 (0.007; 0.05)	0.03 (0.01; 0.07)	**0.07 (0.04; 0.12)**
DCML	**0.01 (0.007; 0.03)**	**0.02 (0.009; 0.06)**	0.09 (0.04; 0.18)
OLS	0.02 (0.008; 0.04)	0.07 (0.03; 0.14)	0.11 (0.06; 0.2)

Table 13: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.017 (0.016; 0.018)**	**0.03 (0.02; 0.04)**	**0.07 (0.06; 0.08)**
Huber	0.025 (0.024; 0.026)	0.04 (0.03 ; 0.05)	0.09 (0.08 ; 0.1)
Tukey	0.019 (0.018; 0.02)	0.04 (0.03 ; 0.05)	0.08 (0.07; 0.09)
L1	0.019 (0.018; 0.02)	0.04 (0.03; 0.05)	0.08 (0.07; 0.09)
FastTau	0.018 (0.017; 0.02)	0.04 (0.03; 0.05)	**0.07 (0.06; 0.08)**
HBR	0.017 (0.016; 0.018)	**0.03 (0.02; 0.03)**	0.08 (0.07; 0.09)
DCML	0.02 (0.019; 0.021)	0.04 (0.03; 0.05)	0.09 (0.08; 0.1)
OLS	0.021 (0.02; 0.022)	0.04 (0.03; 0.05)	0.09 (0.08; 0.1)

Table 14: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.02 (0.01; 0.03)**	**0.03 (0.02; 0.04)**	0.08 (0.07; 0.09)
Huber	0.03 ()	0.04 (0.03 ; 0.05)	0.09 (0.08 ; 0.1)
Tukey	0.03	0.04	0.08

	()	(0.03 ; 0.05)	(0.07; 0.09)
L1	0.03 ()	0.04 (0.03; 0.05)	0.08 (0.07; 0.09)
FastTau	**0.02 (0.01; 0.03)**	**0.03 (0.02; 0.04)**	0.08 (0.07; 0.09)
HBR	0.02 (0.01; 0.03)	0.04 (0.03; 0.05)	**0.06 (0.05; 0.07)**
DCML	0.03 (0.02; 0.04)	0.04 (0.03; 0.05)	0.09 (0.08; 0.1)
OLS	0.03 (0.02; 0.04)	0.04 (0.03; 0.05)	0.09 (0.08; 0.1)

Table 15: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.018 (0.017; 0.019)**	**0.04 (0.03; 0.05)**	0.08 (0.07; 0.09)
Huber	0.02 (0.01; 0.03)	0.05 (0.04; 0.06)	0.09 (0.08 ; 0.1)
Tukey	0.02 (0.01; 0.03)	0.05 (0.04 ; 0.06)	0.08 (0.07; 0.09)
L1	0.02 (0.01; 0.03)	0.05 (0.04; 0.06)	0.08 (0.07; 0.09)
FastTau	**0.018 (0.017; 0.019)**	0.05 (0.04; 0.06)	**0.06 (0.05; 0.07)**
HBR	**0.018 (0.017; 0.019)**	**0.04 (0.03; 0.05)**	0.08 (0.07; 0.09)
DCML	0.02 (0.01; 0.03)	0.05 (0.04; 0.06)	0.1 (0.09; 0.11)
OLS	0.02 (0.01; 0.03)	0.05 (0.04; 0.06)	0.09 (0.08; 0.1)

Table 16: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.022**	**0.034**	**0.08**
Huber	**0.022**	**0.034**	**0.08**
Tukey	**0.022**	**0.034**	**0.08**
L1	**0.022**	**0.034**	**0.08**
FastTau	**0.022**	**0.034**	**0.08**
HBR	**0.022**	**0.034**	**0.08**
DCML	**0.022**	**0.034**	**0.08**
OLS	0.026	0.037	0.084

Table 17: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.022**	**0.033**	**0.08**
Huber	**0.022**	**0.033**	**0.08**
Tukey	**0.022**	**0.033**	**0.08**
L1	**0.022**	**0.033**	**0.08**
FastTau	**0.022**	**0.033**	**0.08**
HBR	**0.022**	**0.033**	**0.08**
DCML	**0.022**	**0.033**	**0.08**
OLS	0.024	0.036	0.1

Table 18: Robust regression estimators MSE results according to the number of covariates (p=3), the percentage of outliers (5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.02**	**0.03**	**0.08**
Huber	**0.02**	**0.03**	**0.08**
Tukey	**0.02**	**0.03**	**0.08**
L1	**0.02**	**0.03**	**0.08**
FastTau	**0.02**	**0.03**	**0.08**
HBR	**0.02**	**0.03**	**0.08**
DCML	**0.02**	**0.03**	**0.08**
OLS	0.023	0.034	0.085

Table 19: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.016	0.02	0.044
Huber	0.017	0.018	0.044
Tukey	0.008	0.02	0.045
L1	0.009	0.018	0.043
FastTau	**0.007**	0.04	0.05
HBR	**0.007**	0.018	**0.04**
DCML	0.009	**0.016**	0.06
OLS	0.02	0.022	0.1

Table 20: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.007**	0.022	0.066

Huber	0.017	0.018	0.14
Tukey	0.008	0.02	0.046
L1	0.009	0.018	0.043
FastTau	0.018	0.04	0.06
HBR	**0.007**	**0.011**	**0.042**
DCML	0.009	0.018	0.043
OLS	0.02	0.028	0.050

Table 21: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=50. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	0.0155	**0.01**	0.07
Huber	0.017	0.018	0.14
Tukey	0.007	0.02	0.05
L1	0.009	0.018	0.04
FastTau	0.022	0.048	0.08
HBR	**0.007**	0.018	**0.04**
DCML	**0.007**	0.015	**0.04**
OLS	0.024	0.028	0.08

Table 22: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.019**	0.04	**0.076**
Huber	0.021	0.04	0.08
Tukey	0.03	0.04	0.08
L1	0.03	0.04	0.08
FastTau	0.025	**0.03**	0.09
HBR	0.027	0.04	**0.076**
DCML	0.03	0.04	0.08
OLS	0.04	0.05	0.1

Table 23: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.019**	0.036	0.09
Huber	0.021	0.039	0.096
Tukey	0.028	0.041	0.078
L1	0.028	0.041	0.078
FastTau	0.023	**0.03**	0.085
HBR	0.028	0.04	**0.075**
DCML	0.02	**0.03**	**0.075**
OLS	0.027	0.041	0.1

Table 24: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=500. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.019**	0.036	0.09
Huber	0.021	0.033	0.096
Tukey	0.028	0.04	0.079
L1	0.028	0.04	0.078
FastTau	0.023	**0.03**	0.085
HBR	0.028	0.04	0.077
DCML	0.02	0.034	**0.074**
OLS	0.029	0.042	0.1

Table 25: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 5%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.02**	**0.03**	**0.08**
Huber	**0.02**	**0.03**	**0.08**
Tukey	**0.02**	**0.03**	**0.08**
L1	**0.02**	**0.03**	**0.08**
FastTau	**0.02**	**0.03**	**0.08**
HBR	**0.02**	**0.03**	**0.08**
DCML	**0.02**	**0.03**	**0.08**
OLS	0.024	0.038	0.084

Table 26: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers (5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 10%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.021**	**0.035**	**0.08**
Huber	**0.021**	**0.035**	**0.08**
Tukey	**0.021**	**0.035**	**0.08**
L1	**0.021**	**0.035**	**0.08**
FastTau	**0.021**	**0.035**	**0.08**
HBR	**0.021**	**0.035**	**0.08**
DCML	**0.021**	**0.035**	**0.08**
OLS	0.025	0.039	0.085

Table 27: Robust regression estimators MSE results according to the number of covariates (p=10), the percentage of outliers

(5%, 10% and 30%) for sample size n=5000. The percentage of leverage (x-outliers) is 30%. Lower MSE is in bold. In brackets, the 95% BCa bootstrap CIs based on B=1000 iterations

	5%	10%	30%
FastMM	**0.022**	**0.036**	**0.082**
Huber	**0.022**	**0.036**	**0.082**
Tukey	**0.022**	**0.036**	**0.082**
L1	**0.022**	**0.036**	**0.082**
FastTau	**0.022**	**0.036**	**0.082**
HBR	**0.022**	**0.036**	**0.082**
DCML	**0.022**	**0.036**	**0.082**
OLS	0.028	0.04	0.09