Topics in Advanced Statistics

C.A. Vriends (440435), W.M. Brus (484505), M. Hofstra (387276), G.J. Leenen (476908)

January 2020

1. Introduction

In this paper we investigate the statistical properties, robustness and predictive performance of three regression estimators. Specifically, we compare the OLS estimator, the LTS estimator and an estimator based on the deterministic MCD covariance matrix (the 'plug-in estimator', or the 'MCD estimator'). We contrast the two robust estimators from a theoretical viewpoint, after which we investigate each estimator's robustness as measured by its empirical influence function (EIF) in the context of a real-life dataset. We then complement the results with a simulation study, in which the theoretical properties of the estimators are illustrated in several scenarios.

The paper proceeds as follows. First, we discuss and compare relevant aspects of the theory behind our estimators. Subsequently, we briefly describe the real-life dataset and apply the estimators to highlight their robustness properties. We then proceed to describe relevant parameters of our simulation study and a selection of important results.

2. Methods

In this section, we elaborate on how one may use robust estimates of location and scatter to construct a robust regression estimator. We also introduce and discuss two examples of such estimators, namely the MCD and LTS estimators.

2.1. Plug-in estimator

The first robust regression estimator employed in this paper is constructed using the deterministic MCD covariance estimator. The regression estimator is easily obtained once we have an estimate of the covariance matrix, since the linear regression entails fitting estimates of the mean μ and covariance Σ of the joint distribution of Y, X in the model $\mathbb{E}[Y_i|X_i] = \alpha + X_i'\beta$, $\beta = \Sigma_{XX}^{-1}\Sigma_{XY}$, $\alpha = \mu_Y - \mu_X'\beta$. If we plug in robust estimators of the location and scatter, we thus obtain a robust regression estimator. The distinctive feature of our plug-in, MCD-based regression estimator is that it is constructed using the subset M of observations that yields the smallest determinant of the covariance matrix.

The main advantages and disadvantages of the plug-in estimator follow directly from the characteristic subset of observations that the estimator uses. As the objective function from which the estimator is derived focusses on minimising the determinant of the covariance matrix, the estimator is naturally more robust to extreme observations, including outliers. By ignoring the observations that contribute most to the volume of the scatter (as measured by the Mahalanobis distance), the estimator is much less sensitive to outliers than estimators that assign a large weight to all observations. In fact, for the appropriate subset size the estimator attains the largest possible breakdown point for an affine equivariant scatter estimate (Lopuhaä et al., 1991). The downside of this is that the use of only a subset of observations generally implies a loss of efficiency in the case where the classical linear model assumptions hold. Whereas non-robust estimators such as OLS use all observations by default, the plug-in estimator ignores some, implying a loss of useful information if there is no contamination[1]. In smaller samples, this may be an important drawback.

[1] Another (less general) way of looking at this is with the Gauss-Markov theorem. If the classical linear model assumptions hold, then the theorem states that OLS is BLUE, and the plug-in estimator must necessarily be inefficient.

Other advantages of the estimator consistency (in a sense to be specified) and asymptotic normality. As advantages, Hubert et al. (2018) note that compared to certain other robust estimators, the MCD estimator boasts asymptotic normality. Further, the location and scatter estimates are Fisher consistent, up to a scaling factor in the latter case (Butler et al., 1993). This result is not only applicable in the case of multivariate normality (Cator and Lopuhaä, 2012), although Hubert et al. (2018) do not state whether this distinguishes the MCD estimator from other robust estimators.

2.2. LTS estimator

The second robust estimator considered is the LTS estimator. As noted in Hubert et al. (2018), there is a strong correspondence between it and the plug-in estimator, especially in the univariate regression setting. To see why, consider that the distinctive feature of the LTS estimator is that is not constructed using the subset of observations that minimises the determinant of the covariance matrix, but instead using the subset that yields the smallest sum of squared residuals. In the univariate case, the minimisation objectives are equivalent, since the determinant of the 1×1 sample covariance matrix $S_H = \frac{1}{h}\sum_{i\in H}(y_i - x_i'\beta)^2$ becomes proportional to (up to scale h) the sum of squared residuals $\sum_{i\in H} r_i^2 = \sum_{i\in H}(y_i - x_i'\beta)^2$. Of course, this does not mean coefficient estimates produced by the two estimators must necessarily be identical, since the reweighted estimators used in practice use different weighting rules.

3. Robustness: Empirical Influence Function

In this section we study the robustness of the estimators when applied to a real-life dataset which relates Dutch footballers' market values to their age. To this end, we employ the empirical influence function (EIF), which studies the behaviour of the estimators when one random observation is perturbed. Prerequisite to the validity of this procedure is that the unperturbed data is well-behaved. If it is not, it becomes difficult to determine whether the effect of the perturbation was due to the addition of an artificial data point, or due to the removal of an original data point; if the removed data point was an outlier to begin with, estimator results may change, even when we add an observation that follows the model distribution.

Figure 1 shows the scatterplot of market value versus age. This plot indicates that there are a few players with market values that are much higher than those of other players similar to their age. We first apply the logarithmic transformation to the market values, as the original data displayed in Figure 1 appears skewed. We subsequently employ an approach using tolerance ellipses based on robust estimates of the location and scatter (estimates are constructed using the deterministic MCD algorithm, `alpha` $= 0.75$). We construct the smallest possible volume tolerance ellipse that contains 97.5% of the data, thereby removing the 2.5% of observations with the greatest correlations.

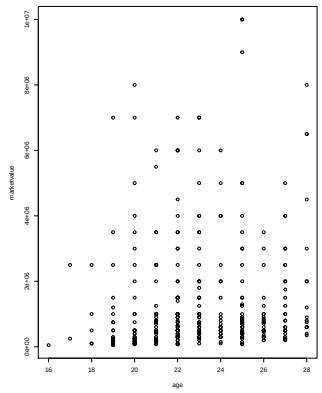


Figure 1: Market value (v) vs. Player age (h)

We proceed to compute the EIF for the variable market value[2]. We choose a random observation from the sample and replace its market value by a new value. Subsequently, we compute the plug-in, LTS and OLS estimates of intercept and slope. We perform this procedure 1,000 times for the variable, gradually increasing the replaced value from zero (as market values are non-negative) to twice the variable's maximum at each step.

In the case of the plug-in and LTS estimators, the proportion of the original cleaned sample that the robust estimators uses to fit the model (`alpha`) must be specified prior to the estimation. As we already took a subset of the original data while removing correlation outliers, we use a relatively high value `alpha` = 0.9.

Figures 2 through 7 show the EIFs for the three different estimators. In each figure, the change in the estimator is plotted on the vertical axis against the change in market value on the horizontal axis. Figure 2 and Figure 5 show that when using OLS, both the intercept and the slope are linearly influenced by a perturbed market value, and that this influence is unbounded. The lack of robustness of the estimator is striking; we can make the estimators deviate arbitrarily far from the original estimate if we make an observation sufficiently large.

Figures 4 through 8 show that the plug-in and LTS estimators have a bounded EIF. As the estimators focus only on a subset of the data, we cannot fit the model with arbitrarily large market values. Beyond a given cutoff point, the perturbed observation falls outside the range of the 90% of observations retained, and the influence of the perturbation becomes zero.

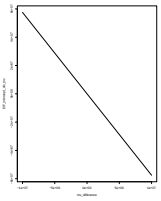


Figure 2: Intercept OLS

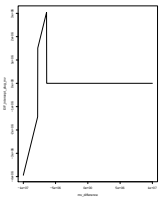


Figure 3: Intercept plug-in

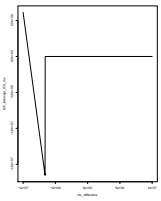


Figure 4: Intercept LTS

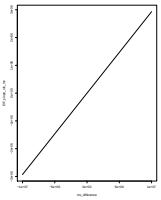


Figure 5: Slope OLS

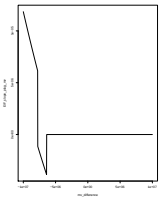


Figure 6: Slope plug-in

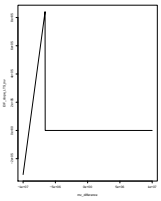


Figure 7: Slope LTS

4. Simulation Study

4.1. Setup

In this section we study the behavior of the OLS, LTS and plug-in estimators in several scenarios. We assess the bias, variance, predictive performance of the estimators across $R = 100$ replications. For the robust estimators, we also inspect the proportion of outliers correctly labelled as such in scenarios where this is relevant.

In our simulation study, we generate a one-dimensional dependent variable $y_i = 3x_i + \eta_i$. We choose to use a single regressor as in that case visual inspection of the data and assessment of e.g. bias is most straightforward. The model distribution of $y_i|x_i$ is a standard normal, i.e. for non-outlier observations we have that x_i and η_i are drawn from independent standard normals. We do this procedure for a training set and for a test set, which consist of an equal number of observations N. The precise choice of N differs

[2]We explicitly omit an analysis for the age variable, as replacing ages of football players aged 16-28 by random values is illogical.

between scenario, depending on which aspect of an estimator the scenario aims to illustrate. In each scenario, the test set contains only observations generated from the model distribution to ensure a fair assessment of the estimators' predictive performance, which we measure by the RMSE.

In scenarios where the training set contains outliers, these outliers occur at a rate ε and may be one of four types. The first distinction between outlier types is whether an outlier can be classified as a vertical outlier, in which case only the disturbance η_i is perturbed, or as a bad leverage point, in which case both x_i and η_i are perturbed. These scenarios are mutually exclusive, i.e. we do not simultaneously generate vertical outliers and bad leverage points. The second distinction depends on the contamination distribution from which η_i is drawn. We consider cases where η_i is drawn from a symmetric distribution (with respect to the mean of $y_i|x_i$) and the case where it is drawn from an asymmetric distribution. We somewhat arbitrarily opt for a $Unif(-3,3)$ as the symmetric distribution, and the asymmetric distribution is an inverse gamma $IG(1,1)$. Both ensure that the probability of an observation relatively far from zero (the mean of $y_i|x_i$ under the model distribution) is large.

The extremity of outliers is controlled using two scaling factors ψ_h and ψ_v. The former controls the leverage of observations, while the latter controls the extremity of a vertical outlier. All in all, the DGP may be summarised as follows:

$$y_i = \begin{cases} 3x_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0,1), & \text{for uncontaminated observations,} \\ 3\psi_h x_i + \psi_v \eta_i, \quad \eta_i \sim \mathcal{U}(-3,3), & \text{for symmetrically contaminated observations,} \\ 3\psi_h x_i + \psi_v \eta_i, \quad \eta_i \sim IG(1,1), & \text{for asymmetrically contaminated observations.} \end{cases} \tag{1}$$

For vertical outliers we thus have that $\psi_h = 1$, while for bad leverage points we set $\psi_h > 1$.

4.2. Scenarios

To evaluate different aspects of the performance of the estimators, we run five scenarios. The sample size, contamination level, outlier extremity and leverage extremity differ amongst the scenarios, depending on which aspect of an estimator the scenario aims to illustrate. In all scenarios where there are outliers, we inspect the cases where they are generated from a symmetric and asymmetric distribution separately.

4.2.1. Scenario 1: No contamination

As explained in section 2, both the MCD and LTS estimator are calculated using a subset of h instead of the full sample of N observations. Consequently, the estimators are expected to be less efficient, as there is a loss of valuable information in case of no contamination. Our first scenario therefore compares the estimators in the setting without contamination, i.e. $\varepsilon = 0$. To make the scenario as favourable as possible for OLS, we further set $N = 50$, as for larger sample sizes the loss of information among the robust estimators becomes small.

4.2.2. Scenario 2: One extreme outlier

The breakdown point of the OLS estimator is the lowest among the estimators at a single observation. To show this empirically, one vertical outlier is introduced ($\varepsilon = \frac{1}{N}$) and the outlier extremity is set at $\psi_v = 1000$. Due to extremity of the outlier compared to the model distribution, the OLS estimator is expected to break immediately. Contrarily, the MCD and LTS should perform well in this setting as their breakdown points are much larger than $\frac{1}{N}$.

In this scenario and in all subsequent scenarios, we increase the sample size to $N = 500$. This permits incorporating a larger number of outliers without breaking the estimators, which is conducive to our study of outlier detection rates in scenarios 4 and 5. The sample size is arguably also more representative of the sample sizes encountered in microeconometric studies.

4.2.3. Scenario 3: Testing the breakdown point

In this scenario the number of outliers is chosen to be exactly one below the breakdown point of the MCD estimator, namely $\varepsilon = \frac{1}{N}\lfloor\frac{(N-p+1)}{2}\rfloor - \frac{1}{N}$. Outliers are modelled as vertical outliers and outlier extremity is

set to $\psi_v = 10^{15}$, to ensure the estimators would break if their breakdown points were lower in reality. This works because, if the subset selection procedures fail to exclude even a single outlier j, the value of x_j will make Σ_{XX}^{-1} (numerically) undefined[3].

Then, the number of outlying observations is increased by one, i.e. we set $\varepsilon = \frac{1}{N}\lfloor\frac{(N-p+1)}{2}\rfloor$. In theory, if we also set $N = 501$, the MCD estimator should break, while the LTS estimator should still work as it has a breakdown point of $\varepsilon = \frac{1}{N}\left(\lfloor\frac{(N-p)}{2}+1\rfloor\right)$ for appropriate subset size $h = \lfloor\frac{N}{2}\rfloor + \lfloor\frac{p+1}{2}\rfloor$ (Rousseeuw and Van Driessen, 2006). In numbers, at $N = 501$ and $p = 1$, the LTS estimator breaks down at 251 outliers, while the MCD estimator breaks down at 250 outliers; our choice of ε would beget exactly 250 outliers.

However, the `ltsReg()` function in `R` permits subset sizes between $h = \lfloor\frac{N+p+1}{2}\rfloor$ for `alpha` $= 0.5$ and $h = N$ for `alpha` $= 1$. As we use `ltsReg()` in this paper, we cannot create the scenario where the MCD estimator breaks but the LTS estimator does not. The best we can do is to test whether the MCD breaks down here.

4.2.4. Scenario 4: Detection of bad leverage points

In this scenario we compare outlier detection rates of the MCD and LTS estimators when the training data are contaminated with bad leverage points. Specifically, we introduce bad leverage points at a rate $\varepsilon = 45\%$ and set the leverage and outlier extremity both to 10. The choice of outlier and leverage extremities ensures that the outliers differ meaningfully from the model distribution, but also that the outliers are not guaranteed to all be detected by the robust methods.

The aim of this scenario is to compare outlier detection rates when the simulation design is arguably more favourable for the MCD estimator, which detects outliers using the Mahalanobis distance. By modelling bad leverage points instead of vertical outliers, we model the outliers as having a large Mahalanobis distance. It is therefore expected that the MCD estimator performs better than the LTS estimator in regard to outlier detection.

4.2.5. Scenario 5: Detection of vertical outliers

In this scenario we revert the argument made in scenario 4. The LTS estimator detects based on the standardized residuals, and we expect it to outperform the MCD estimator when outliers are best characterised as having large residuals. We again set $\varepsilon = 45\%$, but now normalise the leverage extremity ($\psi_h = 1$) while retaining the outlier extremity ($\psi_v = 10$).

4.3. Results

4.3.1. Scenario 1: No contamination

In scenario 1 there is no contamination. As can be seen in Figure 8, the OLS coefficient estimates exhibit the least variability amongst the estimators. Table 1 shows the OLS estimates have a standard deviation of 0.085, noticeably smaller than those of the LTS estimator (0.130) and the MCD estimator (0.249). The latter estimator struggles visibly compared to OLS and LTS, and Figure 8 shows that across 100 replications the estimator has a tendency to underestimate the slope in this small sample size.

We conclude that OLS outperforms the robust estimators in this scenario, in line with our expectations. The OLS slope estimates are closest to the true value on average and exhibit the least variability. In terms of prediction, all estimators produce an RMSE in line with the irreducible error ($Var(\eta_i) = 1$). This was to be expected, as the test set contains no outliers. The MCD estimator trails behind the other estimators as its slope estimate is furthest off on average.

[3]In theory we should not be able to run OLS in this case, as that estimator does not exclude these observations. However, unlike our MCD estimator, OLS is not manually programmed by inverting Σ_{XX} and still works as a result.

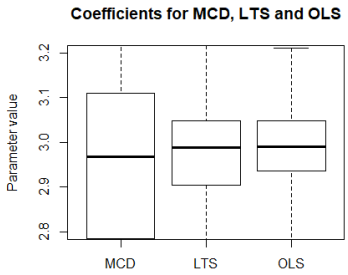


Figure 8: Boxplot of the coefficients of the MCD, LTS and OLS estimators in scenario 1

4.3.2. Scenario 2: One extreme outlier

In Table 1, the mean, variance and RMSE for the estimators are presented in the presence of one large outlier, which is generated from either a uniform or an inverse gamma distribution. The impact of the asymmetric distribution on the OLS estimator is clearly reflected in the mean slope estimate. Whereas the symmetric outlier still permits a reasonable estimate on average (2.62), the asymmetry completely throws OLS off target (7.04). This is not to say OLS performs well even with a symmetric outlier, because at a standard deviation of 2.58 the probability of a slope estimate close to 3.00 is quite small if we only have a single sample. This is clearly reflected in Figures 9a and 9b, where the interquartile range of the OLS estimates reaches well beyond the vertical borders of the boxplots. Noteworthy in these plots is that the *median* OLS estimate is actually close to 3.00 under both the symmetric and asymmetric distribution; the replications in which a bad draw of η occurs, which are subsequently exacerbated by $\psi_v = 1000$, have much less effect on the median OLS estimate than the mean OLS estimate across all replications. This is a testament to the robustness of the median as a location estimator.

The MCD and LTS estimators perform well, with mean estimates of 2.99-3.00 and standard deviations of approximately 0.05 under both the symmetric and the asymmetric distributions. The robust estimators deal with the outlier easily, and are noticeably more accurate than in scenario 1 now that the sample size has increased from 50 to 500. LTS is again more accurate than MCD, with a smaller standard deviation. Both robust methods retain the irreducible error as RMSE.

(a) Continuous uniform

(b) Inverse gamma

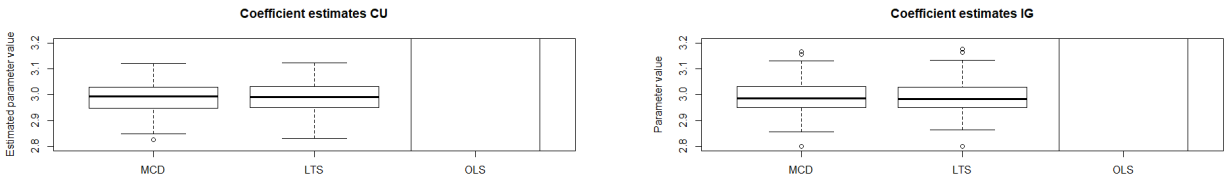
Figure 9: Boxplot of the estimated coefficients of the estimators in scenario 2

4.3.3. Scenario 3: Testing the breakdown point

Figures 10a and 10b show that the MCD and LTS are both still able to handle $(\lfloor \frac{(N-p+1)}{2} \rfloor - 1)$ outliers, even if $\psi_v = 10^{15}$, whether they are symmetrically distributed or not. Contrarily, OLS is off by orders of magnitude, as is clearly reflected in the mean and standard deviation shown in Table 1. The table further shows that the outlier detection rates of the MCD and LTS estimators are 100% on average (precise detection

rules are explained below), both in case of a symmetric and an asymmetric distribution. This is unsurprising given the extremity of the outliers.

As expected, both the MCD and LTS estimator diverge when one additional outlier is added (i.e. $\varepsilon = \lfloor \frac{(N-p+1)}{2} \rfloor$). No results are shown since the regression methods fail.



(a) Continuous uniform

(b) Inverse gamma

Figure 10: Boxplot of the estimated coefficients of the estimators in scenario 3

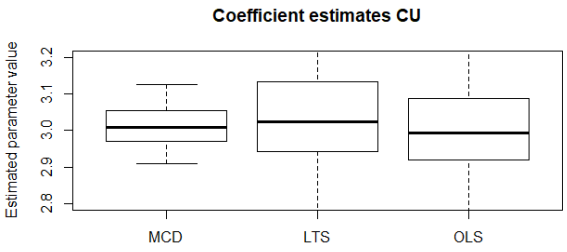
4.3.4. Scenarios 4 and 5: Outlier detection

Scenario 4 and 5 are evaluated simultaneously, as they were both designed to inspect outlier detection rates of the robust estimators. As outlier detection rules, we employ the weighting schemes given on slide 30 of lecture 2 (for MCD) and slide 20 of lecture 3 (for LTS). In the case of MCD this method is not exact, as the assumption that the Mahalanobis distances are χ^2-distributed holds only asymptotically. However, we ignore this fact here. As explained in Hardin and Rocke (2005), the χ^2-based rule generally labels more observations as outliers than the finite-sample rule would. Since we are only interested in relative performances, the χ^2-based rule therefore leads to the same conclusions if the MCD estimator is less effective at outlier detection that the LTS estimator[4].

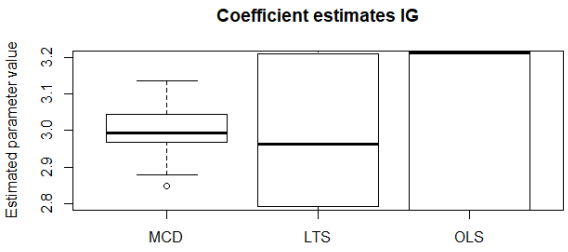
Table 1 shows this surprisingly holds in both scenario 4 and scenario 5. In terms of slope estimate variability, the MCD estimator outperforms the LTS estimator in scenario 4 with a standard deviation that is 2-3 times lower. As Figures 11a and 11b show, this result is especially pronounced when the outliers are drawn from an inverse gamma distribution. However, in terms of outlier detection it underperforms the LTS estimator. This result holds whether the outliers are generated from a symmetric or an asymmetric distribution, although both methods are better able to label the outliers when they are generated from an $IG(1,1)$.

The use of bad leverage points as opposed to vertical outliers does not result in MCD outperforming LTS on outlier detection, but scenario 5 shows that it is nonetheless conducive to the absolute performance of the method. When bad leverage points are placed by vertical outliers, the performance of MCD drops noticeably; 8 to 10% depending on which contamination distribution is used. In the case of LTS the difference between detecting bad leverage points and vertical outliers is marginal, with less than one standard deviation between the mean detection rates. Instead, Figures 12a and 12b show that the biggest improvement for the LTS estimator pertains to the variability of its slope estimates, which are half $[\mathcal{U}(1,1)]$ or even a quarter $[IG(1,1)]$ of what they were in the presence of bad leverage points.

[4]In scenario 3 nothing changes either due to the extremity of the outliers.

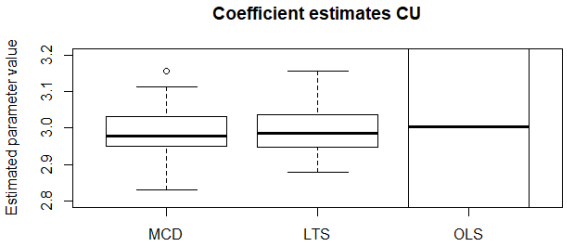


(a) Continuous uniform

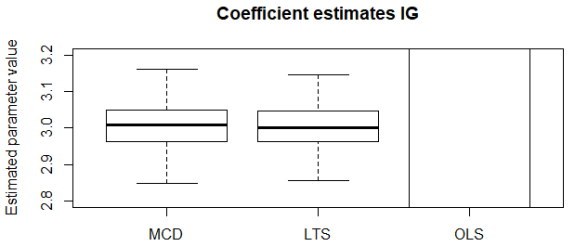


(b) Inverse gamma

Figure 11: Boxplot of the estimated coefficients of the estimators in scenario 4



(a) Continuous uniform



(b) Inverse gamma

Figure 12: Boxplot of the estimated coefficients of the estimators in scenario 5

Table 1: Simulation Results

Results of all simulation scenarios. ε, ψ_h and ψ_v denote contamination level, leverage extremity and outlier extremity. $\hat{\beta}$ and φ respectively denote the slope estimate and the outlier detection rate. In scenario 3, we write the contamination level as $\varepsilon^*_{MCD,-1}$ to indicate that the number of outlying observations was set one below the breakdown point of the MCD estimator.

	Scenario 1		
	$\mathcal{N}(0,1)$		
	ε	ψ_h	ψ_v
	0%	1	1
	Plugin	LTS	OLS
Mean $\hat{\beta}$	2.9514	2.9715	2.9922
SD $\hat{\beta}$	0.2488	0.1300	0.0849
RMSE	1.009	0.9868	0.9818
Mean φ			
SD φ			

	Scenario 2						**Scenario 3**					
	$\mathcal{U}(-3,3)$			$IG(1,1)$			$\mathcal{U}(-3,3)$			$IG(1,1)$		
	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v
	$1/N$	1	10^3	$1/N$	1	10^3	$\varepsilon^*_{MCD,-1}$	1	10^{15}	$\varepsilon^*_{MCD,-1}$	1	10^{15}
	Plugin	LTS	OLS	Plugin	LTS	OLS	Plugin	LTS	OLS	Plugin	LTS	OLS
Mean $\hat{\beta}$	2.9923	2.9959	2.6241	2.992	2.998	7.0355	2.9916	2.9917	-4.30E+12	2.9919	2.9921	2.18E+13
SD $\hat{\beta}$	0.0693	0.0458	2.5755	0.0592	0.0387	28.8195	0.0608	0.0608	5.23E+13	0.0648	0.0656	7.49E+15
RMSE	1.0057	1.0049	1.8803	0.9985	0.9975	3.1915	0.9952	0.9952	4.28E+13	1.0015	1.0016	2.61E+15
Mean φ							1.0000	1.0000		1.0000	1.0000	
SD φ							0.0346	0.0361		0.0400	0.0346	

	Scenario 4						**Scenario 5**					
	$\mathcal{U}(-3,3)$			$IG(1,1)$			$\mathcal{U}(-3,3)$			$IG(1,1)$		
	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v	ε	ψ_h	ψ_v
	45%	10	10	45%	10	10	45%	1	10	45%	1	10
	Plugin	LTS	OLS	Plugin	LTS	OLS	Plugin	LTS	OLS	Plugin	LTS	OLS
Mean $\hat{\beta}$	3.0103	3.0296	2.9993	2.9996	3.0013	5.9014	2.9877	2.9896	2.9867	3.0081	3.0053	-0.1722
SD $\hat{\beta}$	0.052	0.1273	0.1378	0.06	0.2232	38.1499	0.0608	0.0566	0.413	0.0693	0.064	23.0202
RMSE	0.9973	1.004	1.0045	1.0007	1.0235	4.088	0.9976	0.9971	1.0722	1.0009	1.0007	11.2317
Mean φ	0.8298	0.867		0.8587	0.8935		0.7549	0.8707		0.7493	0.9057	
SD φ	0.0300	0.0265		0.0283	0.0316		0.0374	0.0265		0.0387	0.0300	

5. Conclusion

In this paper we analysed several properties of the OLS estimator, LTS estimator and the plug-in estimator. We that as the LTS estimator and plug-in estimator only use a subset of the data to fit the model, they are more robust than the OLS estimator. However, in case of no contamination, the robust estimators suffer from a loss in efficiency because they do not use all the available data to estimate parameters. Subsequently, we used data on Dutch footballers and their age to compare the EIFs of the estimators. We computed the EIF for the intercept and the slope for all three estimators. We observe that when we replace a random observation, the OLS estimators deviate arbitrarily far from the original estimators, whereas the plug-in and LTS estimator have a bounded EIF. This is evidence for the robustness of the plug-in and LTS estimators. Finally, we conducted a simulation to compare these estimators in different scenarios. We confirmed that in the case of no contamination, OLS is the most efficient. In all scenarios with contamination, our results confirm the robustness of the plug-in and LTS estimator.

All in all, the results of the investigation of the EIF and those of the simulation study correspond. Both studies confirm the robustness of the plug-in and LTS estimator. This, together with the simulation evidence that OLS outperforms the robust estimators in terms of efficiency in an uncontaminated sample, is fully in line with theory.

References

Butler, R. W., Davies, P. L., Jhun, M., 1993. Asymptotics for the minimum covariance determinant estimator. Ann. Statist. 21, 1385–1400.

Cator, E. A., Lopuhaä, H. P., 2012. Central limit theorem and influence function for the mcd estimators at general multivariate distributions. Bernoulli 18, 520–551.

Hardin, J., Rocke, D. M., 2005. The distribution of robust distances. Journal of Computational and Graphical Statistics 14, 928–946.

Hubert, M., Debruyne, M., Rousseeuw, P. J., 2018. Minimum covariance determinant and extensions. WIREs Computational Statistics 10, e1421.

Lopuhaä, H. P., Rousseeuw, P. J., et al., 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19, 229–248.

Rousseeuw, P. J., Van Driessen, K., 2006. Computing lts regression for large data sets. Data Mining and Knowledge Discovery 12, 29–45.