Robust Regression Methods and Outlier Detection

Kai Zhang, Haojia Li, Yiwen Wang

2020/2/21

INTRODUCTION

Linear regression is one of the most important models in statistics. Given the independent observations $(\boldsymbol{X_i}, y_i), i = 1, ..., n$, the linear regression model links the predictors and responses in the following form

$$y_i = \boldsymbol{X_i}^T \boldsymbol{\beta} + \varepsilon_i, \ for\ i = 1, \ldots, n$$

where $\boldsymbol{\beta}$ is the vector consisting of unknown parameters, and ε_i are assumed to be independent identically distributed (i.i.d.) random variables following a normal distribution with zero mean and standard deviation σ independent of the $\boldsymbol{X_i}$. In most of cases, $\boldsymbol{\beta}$ is of interest and the methods of ordinary least square (OLS) is commonly used to estimate $\boldsymbol{\beta}$ because of its theoretically attractive statistical properties (e.g., best linear unbiased estimator). However, data in real world usually fail to meet the assumptions of linear regression model. For example, the response variables can possibly follow a non-normal distribution (e.g., heavy-tailed distribution), or include some extreme value such as outlier. Besides, the high leverage points/values may also exist in predictor variables, which is another form of outlier. Unfortunately, OLS estimate is vulnerable in these cases, especially to outliers in predictors. Such outliers can pull the fitted regression model to themselves and make the estimates biased for the remaining observations. Thus, alternatives to OLS is in need.

In this report, multiple robust estimation methods for regression are introduced. In addition, a simulated study is performed to compare the robustness of different estimation methods to different type of outliers. To evaluate the performance of these estimators, the finite breakdrown point and relative efficiency (Donoho and Huber 1983) are used in this report. The breakdown point is defined as the maximal proportion of arbitrarily extreme observations the estimator can handle before its estimates being distorted so badly (goes to infinity) that no any practical use remains, and it intuitively ranges from 0 to 0.5 so as to be an reasonable estimator composing of more typical observations than atypical ones. The relative efficiency of a regression estimator can be computed by the ratio of variance of the OLS estimator to that of the estimator of interest, in the case where all model's assumptions are satisfied and no outlier exist.

On this basis, robust regression methods including M-estimation, bounded-influence regression and MM-estimation are covered in this report.

M-ESTIMATION

In practice, the response variables frequently don't follow a normal distribution but a heavy-tailed distribution, which can easily generate outliers. M-estimation are robust to this type of outliers by weighting the observations during estimation.

M-estimation estimates $\boldsymbol{\beta}$ by minimizing the sum of residual functions $\rho(x)$ in term of scaled residuals

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{n} \rho(\frac{y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}}{\sigma})$$

where a residual function $\rho(u)$ is required to be a symmetric and monotonic function of $|u|$ with zero value at $u = 0$. For numeric stability, $\rho(u)$ is also designed to have continuous derivative respect to the coefficients, thus the first derivative $\psi(u) = \rho'(u)$ is commonly used to search for the optimal solutions

$$\sum_{i=1}^{n} \psi(\frac{y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}}{\sigma}) x_{ij} = 0, \ for \ j = 0, 1, \ldots, k$$

This equation can be formulated in another way, which is equivalent to the estimate generated by weighted least square (WLS)

$$\sum_{i=1}^{n} \omega(\frac{y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}}{\sigma})(y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}) x_{ij} = 0, \ for \ j = 0, 1, \ldots, k$$

where $w(u)$ is defined as

$$\omega(u) = \frac{\psi(u)}{u}$$

In matrix notation, the above equation can be written as

$$\boldsymbol{XWX\beta} = \boldsymbol{XWY}$$

where $\boldsymbol{W}$ is a $n \times n$ matrix with the i th diagonal element equal to $w(\frac{y_i - \boldsymbol{X_i}^T \boldsymbol{\beta}}{\sigma})$. Furthermore, the WLS solution for the coefficients are $\hat{\boldsymbol{\beta}} = (\boldsymbol{XWX})^{-1}(\boldsymbol{XWY})$. Thus, M-estimation is equivalent to WLS, and achievies robustness by assigning lower weights to those observations with large residuals.

Three typical residual functions $\rho(u)$ are least are Least square function, Huber function and the Tukey's bisquare function, and they are displayed in *Figure 1* with their corresponding $\psi(u)$ function and weight function $w(u)$. Besides, the plots of those functions' curves is also presented in *Figure 2*. Compared to least square function with weight function $w(u) = 1$, the weight value of Huber function and bisquare function fall off from 1 with the absolute residual going up. In term of $\psi(u)$, Huber estimation has a monotonic ψ-function, while the ψ-function for the bisquare estiamtion is re-descending. This allows weight function of bisquare estimation decrease faster than Huber estimation (given same tuning constant k), and reach 0 at the tuning constant k, so that the bisquare estimation offer an larger increase in robustness toward large outliers as well as heavy-tailed distributed response.

The turning constant k is picked to make a trade-off between robustness and efficiency of the Huber and bisquare function. Smaller k offers more robustness to outliers but sacrifices part of efficiency. In order to attain 95% efficiency compared to OLS in normal case, k is selected to be 1.345σ for Huber estimation and 4.685σ for the bisquare estimation. At those k value, M-estimation can give powerful estimates and is not sensitive to the outliers (in response variables).

Method	*Objective Function*	*Weight Function*
Least-Squares	$\rho_{\mathrm{LS}}(e) = e^2$	$w_{\mathrm{LS}}(e) = 1$
Huber	$\rho_{\mathrm{H}}(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{H}}(e) = \begin{cases} 1 & \text{for } \lvert e\rvert \le k \\ k/\lvert e\rvert & \text{for } \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_{\mathrm{B}}(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ k^2/6 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{B}}(e) = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

Figure 1: Objective function, weight function for least-square, Huber and bisquare function (Source: Fox J, Weisberg S.Robust regression (J). An R and S-Plus companion to applied regression)

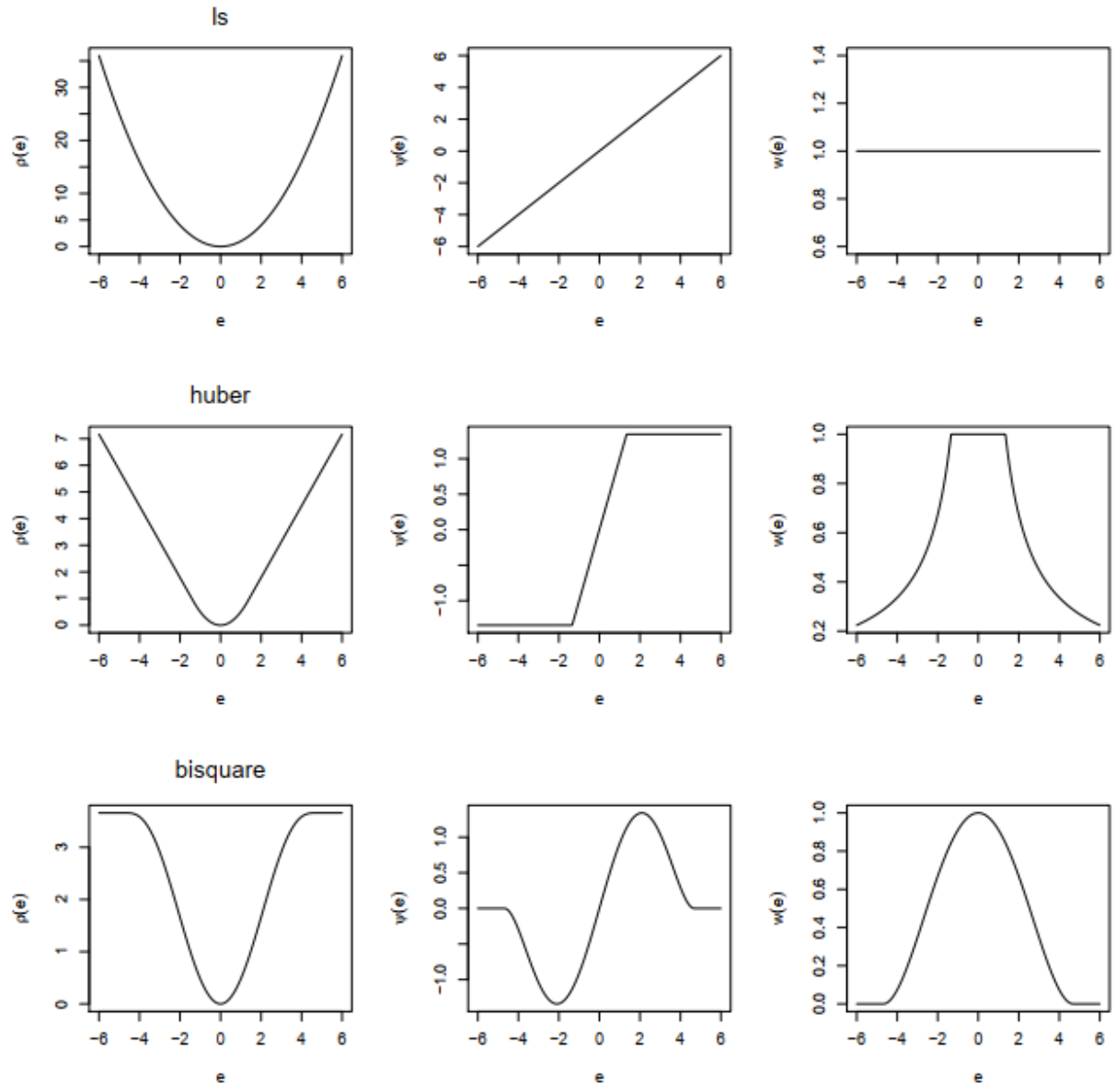

Figure 2: Residual function, objective function, and wieght functions for least-squares, Huber, and bisquare functions. The tunning constants for these graphs are $k = 1.345$ for the Huber function and $k = 4.685$ fro the bisquare.(Source: Fox J, Weisberg S.Robust regression (J). An R and S-Plus companion to applied regression)

In M-estimation, the weights of the ith observation $\omega(\frac{y_i - \boldsymbol{X_i}^T\boldsymbol{\beta}}{\sigma})$ is a function of the residuals, however, residuals are exactly of our interest and unknown before fitting a model. An algorithm called iteratively reweighed least-squares (IRLS) is implemented to search for the optimal solution. In practice, the scale parameter σ in the model is also unknown though it is usually not of our interest. Therefore, σ are commonly estimated by robust scale estimator $\hat{\sigma}$ (e.g., median absolute deviation) before implementing the IRLS algorithm. The IRLS algorithm follows the procedure below:

1. Compute an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, such as OLS estiamte.
2. At each iterantion t, the residuals $\varepsilon_i^{(t-1)} = y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}^{(t-1)}$ for each i th observations, and associated weight $\omega_i^{(t-1)} = \omega(\frac{\varepsilon_i^{(t-1)}}{\hat{\sigma}})$ are computed.
3. The updated WLS estimates is computed by $\hat{\boldsymbol{\beta}}^{(t)} = (\boldsymbol{X}\boldsymbol{W}^{(t-1)}\boldsymbol{X})^{-1}(\boldsymbol{X}\boldsymbol{W}^{(t-1)}\boldsymbol{Y})$, where $\boldsymbol{W}^{(t-1)}$ is the weight matrix from previous iteration. Then, go back to step 2 until the estiamte goes to convergency.

Both the Huber estimation and bisquare estimation are robust alternative to OLS estimation. However, similar to OLS, these two M-estimators are vulnerable to outliers in predictors variables (high-leverage point). Even one "bad" leverage point can contort the fitted model to infinity, thus the finite breakdown point for the class of M-estimators is $1/n$.

BOUNDED INFLUENCE SQUARE

Though M-estimation is robust to y-axis outliers, their finite breakdown point is $1/n$ because their vulnerability to x-axis outliers. Bounded influence square method (e.g. least trimmed square, least median square) is another type of estimation method, which provides estimator with high breakdown point.

The standardized sensitivity curve (SC) is

$$SC_n(x_0) = \frac{\hat{\theta}_{n+m}(x_1, \ldots, x_n, x_0, \ldots, x_0) - \hat{\theta}_n(x_1, \ldots, x_n)}{m/(n+m)}$$

where $x_1, \ldots, x_n$ are observations sampled from a specific distribution, a set of identical values x_0 contaminate the sample and are allowed to range on the whole line, and $\hat{\theta}_n$ is the estimator with n observations. The sensitivity curves measures how bad the estimate will be distorted by different x_0. If the sensitivity curve is bounded, it indicates the estimator can tolerate m outliers (both predictors and responses) in total $n+m$ sample size (finite breakdown point $\geq m/(m+n)$). Rousseeuw (1987) showed that any regression equivariant estimator has breakdown point less than $([(n-p)/2]+1)/n$, where n is the number of observations and p is the number of predictors. And least trimmed square (LTS) and least median square (LMS) regression are two typical methods that can attain that upper bound.

The least trimmed square (LTS) estimate is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \sum_{i=1}^{q} (y_{(i)} - \boldsymbol{X}_{(i)}^T\boldsymbol{\beta})^2$$

where $(y_{(1)} - \boldsymbol{X}_{(1)}^T\boldsymbol{\beta}) \leq \cdots \leq (y_{(q)} - \boldsymbol{X}_{(q)}^T\boldsymbol{\beta}) \leq \cdots \leq (y_{(n)} - \boldsymbol{X}_{(n)}^T\boldsymbol{\beta})$. The robustness of LTS is determined by the proportion of observations are trimmed when fitting the model. Rousseeuw (1987) showed that the upper bound of breakdown point $([(n-p)/2]+1)/n$ can be attained by LTS regerssion when $q = [n/2]+[(p+1)/2]$. Therefore, LTS regression is of very high breakdown point and robust to multiple types of outliers. However, the variance of LTS estimator is also very large, especially when too many observations are trimmed.

And the least median square (LMS) is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} Med\left\{(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2\right\}$$

Similar to LTS regression, LMS can also attain the breakdown point $([(n-p)/2]+1)/n$ (Rousseeuw, 1987). Unfortunately, LMS regression's efficiency is very low as well (0.37 at normal case, Rousseeuw and Croux 1993).

MM-ESTIMATION

M estimate is high-efficient and robust to outliers in response variables, but it is vulnerable to bad leverage points, so the finite breakdown point is $1/n$. Besides, a preliminary estimate of scale parameter σ is required when computing IRLS algorithm to find the solution. On the contrary, bounded influence square methods can resist any type of outliers (both in response and predictors variables) with high breakdown point, but they are low-efficient at normal case. Thus, there is a trade-off between robustness and efficiency between the two. However, MM-estimate (Yohai 1987) achieve both a high breakpoint and high efficiency at the normal distribution, while does not require a scale estimate preliminarily.

Assume bounded residual function ρ (e.g., bisquare function), and $\rho_0(u) = \rho(u/c_0)$ and $\rho_1(u) = \rho(u/c_1)$. Moreover, $c_0 \leq c_1$ holds to ensure $\rho_0 \geq \rho_1$. The computation of MM-estimate follows the following steps:

1. Compute an initial high breakdown value estimate $\hat{\boldsymbol{\beta}}^{(0)}$ which may not efficient at the normal case. The initial estimate can be LTS and LMS estimates, so no previous scale parameter is included.

2. The scale σ is estiamted by M estimation with residuals computed from the initial $\hat{\boldsymbol{\beta}}^{(0)}$, so the estiamte $\hat{\sigma}$ is the solution of the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}^{(0)}}{\hat{\sigma}}\right) = 0.5$$

3. Use ρ_0-function as the residual function, $\hat{\boldsymbol{\beta}}^{(0)}$ as the initial estimate, and $\hat{\sigma}$ as the scale estiamte, IRLS algorithm is used to find the point to minimize

$$L(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\rho_1\left(\frac{y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}}{\hat{\sigma}}\right)$$

Yohai (1987) showed that, given ρ_0 and ρ_1, if $\hat{\boldsymbol{\beta}}$ is such that $L(\hat{\boldsymbol{\beta}}) \leq L(\hat{\boldsymbol{\beta}}^{(0)})$, then $\hat{\boldsymbol{\beta}}$ is consistent and its breakdown point is not less than that of $\hat{\boldsymbol{\beta}}^{(0)}$. Fortunately, it is also proved that $L(\hat{\boldsymbol{\beta}})$ decreases at each iteration within the IRLS algorithm, which makes the condition $L(\hat{\boldsymbol{\beta}}) \leq L(\hat{\boldsymbol{\beta}}^{(0)})$ always ensure. In this way, $\hat{\boldsymbol{\beta}}$ should have a high breakdown point. Moreover, because $\hat{\boldsymbol{\beta}}$ is still a M-estimator essentially, it can also achieve a high efficiency as well. As a result, the MM-estimate achieve the high breakdown point and high efficiency under normal case simutaneously.

Application and Discussion

In this section, we compare the performance of the robust regression methods (LSE, Huber, Bisquare, MM-estimation and LTS) dealing with different type of outliers (and non-normal respsonses) by simulating the corresponding dataset. In the simulation study, we have the response variable Y and the predictors X with sample size equal 60, and the true $\beta_1 = 3$.

$$y = 3x + \varepsilon,\ where\ x \sim N(5, 2^2)$$

- Case 1: $\varepsilon \sim N(0,1)$ – standard normal distribution
- Case 2: $\varepsilon \sim t3$ – t-distribution with degree of freedom 3
- Case 3: $\varepsilon \sim t1$ – t-distribution with degree of freedom 1
- Case 4: $\varepsilon \sim 0.9N(0,1) + 0.1N(0, 20^2)$ – contaminated normal mixture
- Case 5: $\varepsilon \sim N(0,1)$ with 10% identical outliers in y direction (where we let 10% of y's equal to 60)
- Case 6: $\varepsilon \sim N(0,1)$ with 10% identical high leverage outliers (where we let 10% of x's subject to N(30,1) and y's equal to 200)

The *Figure 3* shows the simulated residual distribution for all the cases.

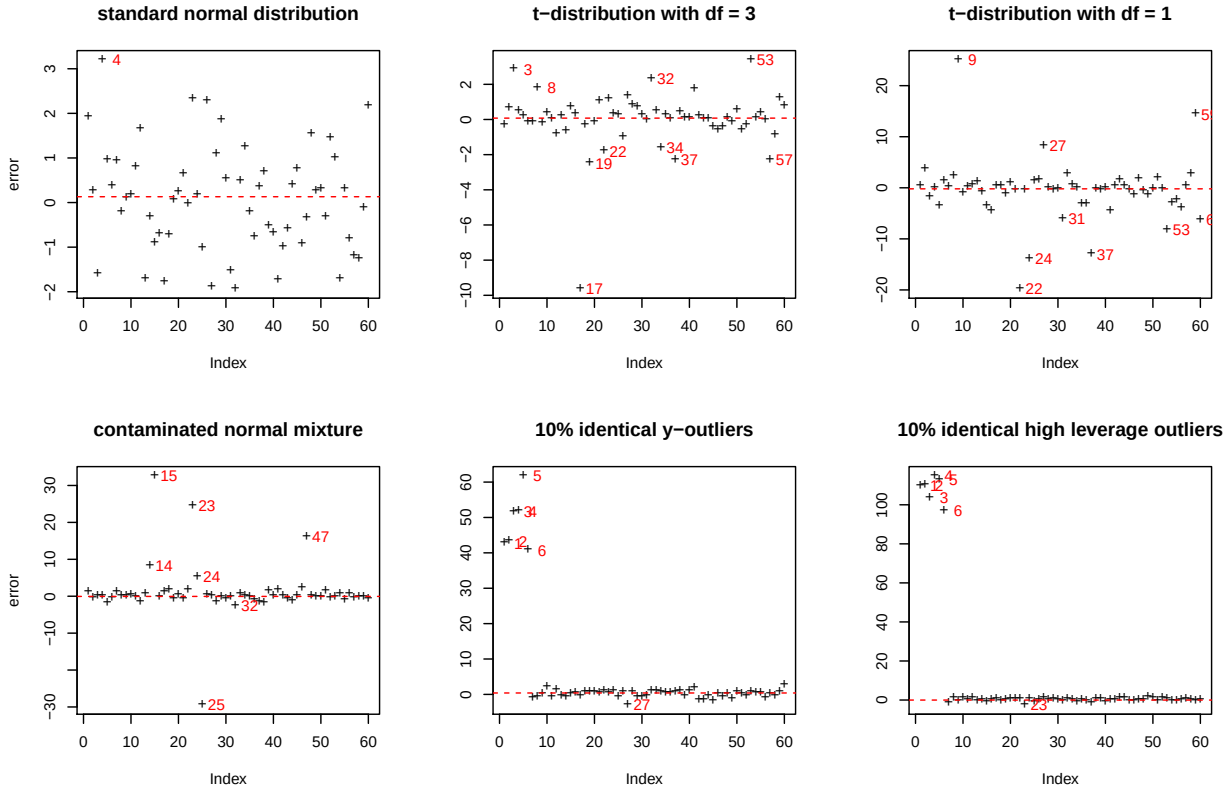

Figure 3: Simulated residual distributions of differnet cases. Upper left: case 1; upper middle: case 2; upper right: case 3; lower left: case 4; lower middle: case 5; lower right: case 6

Applying all the five methods to the cases, the fitted regression lines are displayed in *Figure 4* where the estimates of β_1 are also shown on the top left corners. According to this, the robustness of different methods are compared. For case 1, which is ideal that all the assumptions of OLS are met, all the regression lines overlap together and there is no significant difference among them. For heavy-tail cases and contamined normal-mixture case, the robust regression methods slightly outperform OLS .

In case 5 where only y-axis outliers exsit, OLS fitted line is pulled toward the outliers (in red) and the estiamte ($\hat{\beta}_{OLS} = 1.0614$) is biased from true value 3.00. However, robust regression method resisit the outliers in this case, and give reasonable estiamtes. In addition, when x-axis outliers exist, the results of M-estimations (including OLS) are distorted by the "bad" leverage point, but MM-estimation and LTS still work well in this case.

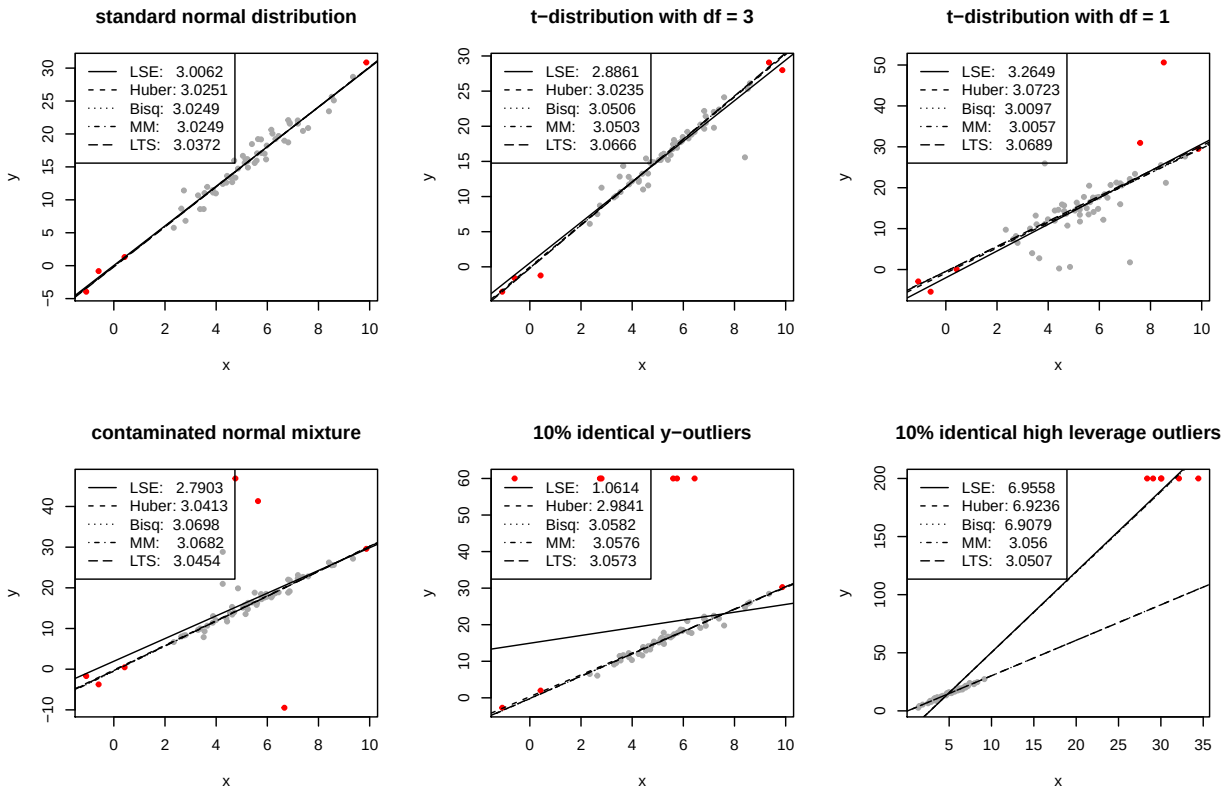

Figure 4: Fitted lines and the corresponding estimates (slope). Upper left: case 1; upper middle: case 2; upper right: case 3; lower left: case 4; lower middle: case 5; lower right: case 6

The efficiency of these estimation methods are evaluated by standard error of $\hat{\beta}_1$ at normal case (case 1), which are computed using boostrapping techniques. Table 1 reports those standard errors of β_1: OLS estiamte has smallest standard error, agreeing with the good property of "BLUE"; $\hat{\beta}_1$ of Huber, bisquare and MM regression have standard error slightly larger than that of OLS, confirming their high efficiency; LTS provides estimates with low efficiency, and its standard error is more than twice as OLS's.

Table 1: Standard error of $\hat{\beta}_1$ under normal case by Boostraping

	OLS	Huber	Bisquare	MM	LTS
SE(b1)	0.0736	0.076	0.0764	0.0773	0.1737

Conclusion

The OLS estiamte have multiple good statistical properties, while assumptions of linear regression model are rarely meet in real data. In addition, because OLS estiamte is volnerable to outliers in various cases, more robust methods are required to weaken outliers' impact on the model. In this report, M-estiamtion generalizes OLS estimtion into an equivalent of WLS, so as to lower the weight of observations with large residuals. This provides a high-efficient estimator, though class of M-estiamtor can still be easilty distored by leverage points. In comparison, LTS and LMS regression are robust to multiple types of outliers and enable to attain the upper bound of breakdown points (0.5, when sample size is large). However, their variance is large at normal case. Finally, MM-estimation implement a computing methods that inherit the good quality of high-breakdown value estimates and M-estimation, achieving high efficiency and breakdown point value simutaneously.

Reference

1. Maronna, Ricardo A, Victor J Yohai, and Douglas R Martin. Robust Statistics : Theory and Methods. Wiley Series in Probability and Statistics. Chichester, England: J. Wiley, 2006.
2. Rousseeuw P J, Leroy A M. Robust regression and outlier detection[M]. John wiley & sons, 2005.
3. Rousseeuw P J, Croux C. Alternatives to the median absolute deviation[J]. Journal of the American Statistical association, 1993, 88(424): 1273-1283.
4. Fox J, Weisberg S. Robust Regression in R An Appendix to An R Companion to Applied Regression[J]. 2010.
5. Yohai V J. High breakdown-point and high efficiency robust estimates for regression[J]. The Annals of Statistics, 1987: 642-656.
6. Yu C, Yao W. Robust linear regression: A review and comparison[J]. Communications in Statistics-Simulation and Computation, 2017, 46(8): 6261-6282.