ROBUST PARAMETER ESTIMATION IN SIMPLE LINEAR REGRESSION WITH GUMBEL-DISTRIBUTED ERRORS AND OUTLIER CONTAMINATION

DEMET AYDIN, OMER BAYRAKTAR
Department of Data Science and Analytics
Sinop University
Faculty of Science and Arts, 57000, Sinop
TURKEY

ABSTRACT: - This article focuses on parameter estimation in a simple linear regression model, assuming that the error terms follow the one of the skewed distributions is Gumbel and dataset contains outliers. The unknown model parameters are estimated using the traditional least squares (LS) method as well as robust estimation methodologies, including the least absolute deviation (LAD), weighted least absolute deviation (WLAD), and least squares median (LMS). The performance of these estimators is evaluated against the LS estimator using BIAS and mean square error (MSE) measures as criteria. A Monte Carlo simulation study was conducted to assess their effectiveness on datasets with outliers in explanatory variable (leverage point). The simulation results reveal that the LMS estimator outperforms the other methods in all considered outlier scenarios, demonstrating greater efficiency and robustness regardless of the presence of outliers. Furthermore, the practical applicability of the proposed estimators is illustrated through an analysis of the data set, which was intentionally modified by introducing varying percentages of outliers. This real-world example underscores the effectiveness of the robust estimators, particularly the LMS, in handling data contaminated by anomalies.

KEY-WORDS: - Gumbel distribution, regression model, robust estimation methods, x- direction outliers, Monte-Carlo simulation

1 INTRODUCTION

In this article, we will consider the robust parameter estimation in a simple linear regression model, which is defined as follows. Let x_i be the explanatory variable and y_i be the response variable for $i = 1,2,\cdots,n$. Suppose that the response variable y_i depends on the explanatory variable x_i in a linear way as:

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i,\ \ i = 1,2,\cdots,n$$

where θ_0 is the intercept, θ_1 is the slope or the regression coefficient, n is the number of observation and ε_i are the independently and identically distributed (iid) error terms.

In the context of regression models, it is a traditional assumption that error terms are normally distributed. However, in practical applications, non-normal distributions are encountered much more frequently [1-8]. The efficiency of the least squares (LS) estimators declines rapidly when the distribution of error terms deviates from normality. In this case, the standard errors of the parameter estimates are also overestimated, which leads to tests with a significance level larger than the nominal one and confidence intervals that are unnecessarily wide [9-10]. Unfortunately, the impact of violating the normality assumption on the efficiency of these estimators is often neglected. One way to deal with non-normal data is the Box-Cox transformation in [11], which aims to normalize the distribution of the data or at least approximate it. However, not all non-normal data may be amenable to this transformation. Furthermore, since the transformed parameters are often difficult to interpret, this study will consider the original data rather than the transformed data [12].

Many studies have explored the problem of estimating model parameters in regression analysis under the assumption of non-normality. In [6], the authors developed robust versions of the change point estimation methods to estimate the model parameters in a two-phase linear regression model with long-tailed symmetric distributed errors. The linear regression model with respect to Gumbel distribution was investigated in [13], and maximum likelihood estimates of the model parameters were obtained. In [14], authors used a modified version of the maximum likelihood method to estimate parameters in a linear model with asymmetric error distributions, i.e., Weibull and generalized logistic. For the parameters in the multiple linear regression model with non-normal symmetric error distributions, the modified maximum likelihood (ML) estimators, least squares (LS) estimators, and the M estimators were obtained in [7], and their efficiencies were compared for plausible alternatives with an assumed distribution.

On the other hand, in practice, it is often observed that datasets containing outliers are quite common. It is widely known that traditional estimation techniques, such as the maximum likelihood (ML) and least squares (LS) are non-robust in the presence of outliers and contaminated data. In other words, the performances of non-robust estimators significantly change in the presence of anomalous data points [15]. To address this issue, numerous robust estimation methods have been developed and proposed in the literature [3, 16-20]. These methods aim to reduce the influence of outliers, ensuring more reliable and accurate parameter estimates even when the data contains extreme observations.

In this study, we consider the simple linear regression model with iid error terms ε_i having a Gumbel distribution, also known as Extreme value type I distribution, with scale parameter $\sigma > 0$. The corresponding probability density function and cumulative density function for $\varepsilon_i \in \mathbb{R}$, are given respectively by [15, 21-23]

$$f(\varepsilon_i\,;\sigma\,) = \frac{1}{\sigma} exp\left(-\frac{\varepsilon_i}{\sigma} - exp\left(-\frac{\varepsilon_i}{\sigma}\right)\right),$$

and

$$F(\varepsilon_i\,;\sigma\,) = exp\left(-exp\left(-\frac{\varepsilon_i}{\sigma}\right)\right).$$

Mean, variance, skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) values of the Gumbel distribution are given as

$$E(\varepsilon_i) = \sigma\gamma,$$
$$Var(\varepsilon_i) = \frac{\pi^2}{6}\sigma^2,$$
$$\sqrt{\beta_1} = \frac{12\sqrt{6}}{\pi^3}\varsigma(3)$$

and

$$\beta_2 = \frac{27}{5},$$

where γ is Euler's constant defined by

$$\gamma = -\int_0^\infty \ln(\varepsilon_i)\, exp(-\varepsilon_i)\, d\varepsilon_i\,,$$

and $\varsigma(3)$ is Apéry's constant [24].

The graphs of the density function of the Gumbel distribution, Figure 1, for various values of the scale parameter σ, indicate that the Gumbel distribution is unimodal and skewed to the right.

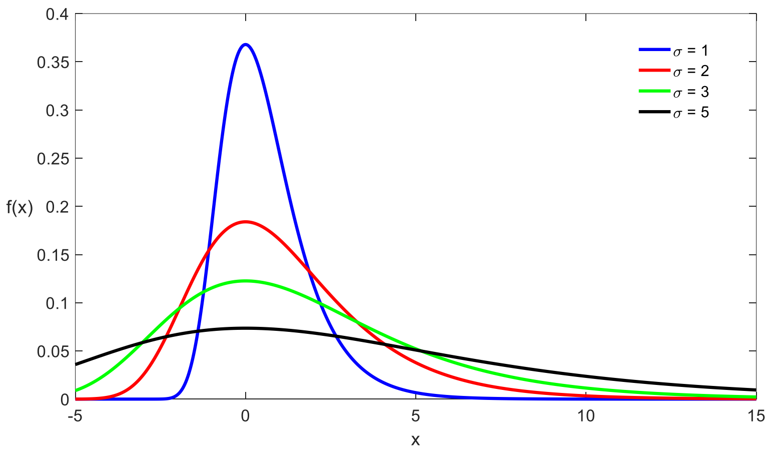

Figure 1: Different density functions of the Gumbel distribution for some selected values of σ.

This paper provides robust parameter estimations for the simple linear regression model with Gumbel error distribution under the assumption that data contains outliers in the x- direction (the leverage points). The least median of squares (LMS) method, is introduced by [20] to address the problem of low breakdown point, can resistant the effect of 50% contamination in the data, unlike the LS and least absolute deviation (LAD) estimators that have zero distortion point [25]. The LMS method is applied for the estimation of the regression model parameters, and the results are compared with those of LAD, weighted least absolute deviation (WLAD), and LS estimation.

The paper is organized as follows: Section 1 introduces the simple linear regression model and the Gumbel distribution. In Section 2, the estimation methods are briefly discussed. In Section 3, the performance of the estimation methods mentioned in Section 2 is evaluated by extensive simulation study on data sets with and without outliers based on various comparison criteria. In Section 4, a dataset is analyzed to illustrate the application of the methods. Finally, conclusions are provided in Section 5.

2 ESTIMATION METHODS

In this section, the methodologies used for estimating the regression model parameters are described.

2.1 LEAST SQUARES ESTIMATORS-LS

The least squares (LS) estimator is highly sensitive to outliers, lacking robustness even when confronted with a single anomalous data point. This sensitivity can significantly undermine the reliability and accuracy of the LS estimator [15, 26].

To obtain the estimators of the regression parameters, we first order $z_i = y_i - \theta_0 - \theta_1 x_i$ so that $z_{(1)} < z_{(2)} < \cdots < z_{(n)}$ with $z_{(i)} = y_{[i]} - \theta_0 - \theta_1 x_{[i]}$ $(i = 1,2,\cdots,n)$, where $\left(y_{[i]}, x_{[i]}\right)$ may be referred to as the concomitants of $z_{(i)}$, representing the pair of $(y_i\,, x_i)$ values that determine $z_{(i)}$ [3]. Since complete sums are invariant to ordering, the LS estimators of θ_0 and θ_1 are obtained by minimizing the following function

$$\phi_{LS}(\theta_0\,,\theta_1) = \sum_{i=1}^{n}\left(z_{(i)}\right)^2.$$

2.2 LEAST ABSOLUTE DEVIATION ESTIMATORS-LAD

The least absolute deviation (LAD) estimators are less sensitive to outliers compared to the ordinary LS estimators, particularly in the presence of heavy-tailed distributions or outliers [16]. However, they are sensitive to leverage points [27]. Additionally, the LAD estimators are asymptotically normally distributed without requiring any assumptions on the distribution of errors, as discussed in [28-30].

Using the LAD method, estimators of θ_0 and θ_1 are found by minimizing the following function

$$\phi_{LAD}(\theta_0\,,\theta_1) = \sum_{i=1}^{n}\left|z_{(i)}\right|,$$

with respect to unknown parameters.

2.3 WEIGHTED LEAST ABSOLUTE DEVIATION ESTIMATORS-WLAD

The weighted least absolute deviation (WLAD) method is recognized for its robustness, especially in dealing with outliers and non-normal error distributions. Furthermore, it exhibits an asymptotically normal distribution, provided that the density function and its derivative are uniformly bounded. This characteristic ensures that the WLAD method remains reliable and effective under a wide range of conditions, making it a valuable tool for statistical analysis in practical applications [31-32]. Using the WLAD method, estimators of θ_0 and θ_1 are found by minimizing the following function

$$\phi_{WLAD}(\theta_0\,,\theta_1) = \sum_{i=1}^{n} w_i \left|z_{(i)}\right|,$$

with respect to the unknown parameters. One of the primary challenges in implementing the WLAD method is determination of the appropriate weights w_i in the function $\phi_{WLAD}(\theta_0\,,\theta_1)$. According to the study in [33], the WLAD method, where the weight function is derived using the delta method, demonstrates performance that is nearly on par with that of maximum likelihood estimation. This finding highlights the effectiveness of the delta method in weight determination, making it a viable approach for enhancing the accuracy and robustness of the WLAD method in various statistical applications [15]. Moreover, the commonly used weight function in the literature is

$$w_i = \frac{i(n-i+1)}{(n+2)(n+1)^2}.$$

In this study, we use the weight function w_i. Since, under the assumption that the error terms follow a Gumbel distribution and data includes outliers in the x-direction, the WLAD estimators obtained with the weight function w_i, shows superior performance with the smaller MSE values compared to the WLAD estimators obtained with the weight function proposed in [33] .

2.4 LEAST MEDIAN OF SQUARES ESTIMATORS-LMS

The least median of squares (LMS) method proposed in [17] is notable for its robustness, indicating a high level of insensitivity to outliers and deviations from model assumptions. This characteristic makes the LMS method particularly valuable in scenarios where data integrity is compromised by anomalies [34]. The LMS estimates for θ_0 and θ_1 are obtained by minimizing the following function

$$\phi_{LMS}(\theta_0\,,\theta_1) = Med_i\big(z_{(i)}\big)^2,$$

with respect to the regression parameters θ_0 and θ_1.

3 SIMULATION STUDY

This section evaluates and compares the performance of various methods for estimating the parameters of a simple linear regression model, assuming that the error terms have a Gumbel distribution. The assessment is conducted through a Monte-Carlo simulation study, where the consistency of the estimators is measured using the BIAS and mean square error (MSE) as evaluation criteria.

The simulation process involves the following steps:

Step 1: Set the population values of model parameters as

$$\boldsymbol{\theta} = (\theta_0\,,\theta_1)^T = (0,1)^T.$$

Step 2: Set the sample size as $n = 15$, 30, 60 and 100.

Step 3: Set the outliers rate in the data as $\xi = 0$, 0.05, 0.10 or 0.15.

Step 4: Generate the x_i design variable from the standard normal distribution.

Perform Steps from 4 to 10 for $M = 5000$ replications.

Step 5: Generate the values of the Gumbel distribution using the inverse transformation method

$$\varepsilon_i = F_{\varepsilon_i}^{-1}(u_i) = -\,ln(-\,ln\,u_i),$$

where $u_i \sim U(0,1)$, $i = 1,2,\cdots,n$

Step 6: Substituting the $\boldsymbol{\theta}$ vector in Step 1, x_i in Step 4 and ε_i in Step 5 into the model, obtain the values of dependent variable y_i.

Step 7: Determine the number of outliers $r = n\xi$ in the data.

Step 8: Replace r random selected values with r outliers generated by using the uniform distribution (see, Figure 2).

Step 9: Compute the estimates of model parameters θ_0 and θ_1 using the LS, LAD, WLAD and LMS estimators, respectively.

Step 10: Calculate the BIAS and MSE measures for model parameters which are given by

$$\mathrm{BIAS}(\hat{\theta}_0) = \textstyle\sum_{i=1}^{\mathrm{M}}(\theta_0 - \hat{\theta}_{0_i}),$$
$$\mathrm{BIAS}(\hat{\theta}_1) = \textstyle\sum_{i=1}^{\mathrm{M}}(\theta_1 - \hat{\theta}_{1_i}),$$
$$\mathrm{MSE}(\hat{\theta}_0) = \frac{1}{\mathrm{M}}\textstyle\sum_{i=1}^{\mathrm{M}}(\theta_0 - \hat{\theta}_{0_i})^2,$$
$$\mathrm{MSE}(\hat{\theta}_1) = \frac{1}{\mathrm{M}}\textstyle\sum_{i=1}^{\mathrm{M}}(\theta_1 - \hat{\theta}_{1_i})^2,$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the estimates of model parameters θ_0 and θ_1 in the i-th simulation for $i = 1, 2, \ldots, \mathrm{M}$. Here, smaller BIAS and MSE measures mean better performance as they represent the cases where the estimates are closer to the corresponding population values of model parameters. Additionally, we also present the relative efficiency of the LS estimators to make it easier to compare the efficiency of estimators based on their MSE values. This approach provides a clearer assessment of how each estimator performs compared to the LS method. The RE is calculated by the formula:

$$\mathrm{RE} = \frac{\mathrm{MSE}}{\mathrm{MSE_{LS}}} \times 100.$$

MATLAB software is used for all computations. Tables 1-2 provide the BIAS, MSE and RE measures obtained by each estimation method in estimating the parameters of the regression model for different values of sample size, corresponding to data with 0%, 5%, 10% and 15% outliers in x- direction.

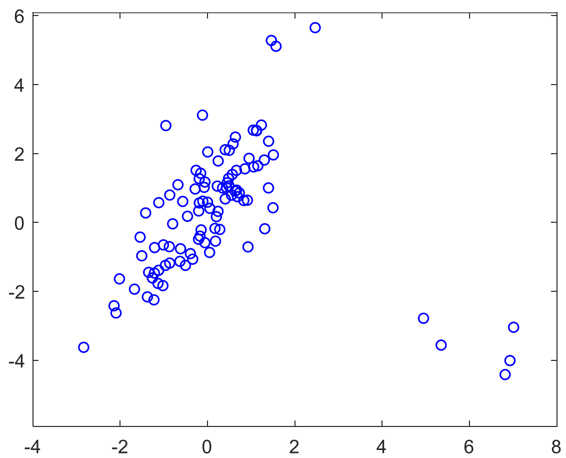

Figure 2: Scatter plot for simulated data with $\mathrm{M} = 1$, $n = 100$ and $\xi = 0.05$.

The Results of Simulation Study:

The simulation results are presented in Tables 1–2. When data do not contain outliers, the LMS estimator shows the best performance among the others with all sample sizes in terms of the BIAS and MSE measures for parameter θ_0 and the BIAS measure for parameters θ_1. However, it should be noted that the LAD and LS estimators of θ_1 outperforms the other estimation methods with the smallest BIAS values for all sample sizes. This suggests that, in case of no outliers, traditional LS methods can still provide reliable estimates, particularly for the slope parameter θ_1, despite the overall superior performance of the LMS estimator.

Table 1. The LS, LAD, WLAD and LMS estimates of model parameters for different sample sizes and corresponding BIASs and MSEs with varying sample sizes in the case that data contain 0% and 5% percentages of outliers.

			θ_0			θ_1		
Outliers Rate	Sample size	Method	BIAS	MSE	RE	BIAS	MSE	RE
0%	n=15	LS	0.5739	0.4467	100	0.0031	0.1376	100
		LAD	0.3784	0.3019	68	**0.0015**	0.1654	120
		WLAD	0.3634	0.2970	66	0.3108	0.2545	185
		LMS	**0.0187**	**0.0393**	**9**	0.0096	**0.0390**	**28**
	n=30	LS	0.5759	0.3898	100	0.0019	0.0618	100
		LAD	0.3799	0.2190	**56**	**0.0012**	0.0779	126
		WLAD	0.3841	0.2330	60	0.3470	0.1960	317
		LMS	**0.0145**	**0.0144**	**4**	0.0215	**0.0146**	**24**
	n=60	LS	0.5776	0.3616	100	0.0041	0.0288	100
		LAD	0.3796	0.1787	49	**0.0029**	0.0370	128
		WLAD	0.3913	0.1936	54	0.3765	0.1778	617
		LMS	**0.0068**	**0.0041**	**1**	0.0198	**0.0045**	**16**
	n=100	LS	0.5739	0.3474	100	0.0033	0.0166	100
		LAD	0.3737	0.1610	46	**0.0019**	0.0210	126
		WLAD	0.3870	0.1742	50	0.3852	0.1697	1022
		LMS	**0.0043**	**0.0017**	**0.5**	0.0167	**0.0019**	**11**
5%	n=15	LS	0.4002	0.3438	100	0.8916	0.8411	100
		LAD	0.2964	0.2819	82	0.5222	0.5107	61
		WLAD	0.2779	0.2612	76	0.5357	0.4784	57
		LMS	**0.0274**	**0.0434**	**13**	**0.0114**	**0.0428**	**5**
	n=30	LS	0.3919	0.2516	100	0.9941	1.0059	100
		LAD	0.3010	0.1901	76	0.6174	0.5187	52
		WLAD	0.2871	0.1724	69	0.5322	0.3809	38
		LMS	**0.0238**	**0.0192**	**8**	**0.0172**	**0.0189**	**2**
	n=60	LS	0.4084	0.2139	100	0.9734	0.9587	100
		LAD	0.3185	0.1462	68	0.4882	0.3025	32
		WLAD	0.3119	0.1407	66	0.4542	0.2480	26
		LMS	**0.0083**	**0.0050**	**2**	**0.0160**	**0.0052**	**0.5**
	n=100	LS	0.4143	0.2012	100	1.0282	1.0641	100
		LAD	0.3212	0.1307	65	0.5307	0.3261	31
		WLAD	0.3124	0.1239	62	0.4432	0.2201	21
		LMS	**0.0057**	**0.0018**	**0.9**	**0.0158**	**0.0020**	**0.2**

Bold indicates the best method in estimating model parameters with respect to either BIAS, MSE, or RE measure with varying values of sample size.

Table 2. The LS, LAD, WLAD and LMS estimates of model parameters for different sample sizes and corresponding BIASs and MSEs with varying sample sizes in the case that data contain 10% and 15% percentages of outliers.

			θ_0			θ_1		
Outliers Rate	Sample size	Method	BIAS	MSE	RE	BIAS	MSE	RE
10%	n=15	LS	0.3359	0.3533	100	1.1366	1.3191	100
		LAD	0.1670	0.2785	79	1.1198	1.4295	108
		WLAD	0.1444	0.2130	60	0.8840	0.9588	73
		LMS	**0.0417**	**0.0520**	**15**	**0.0036**	**0.0503**	**4**
	n=30	LS	0.3548	0.2412	100	1.1270	1.2834	100
		LAD	0.2303	0.1891	78	1.0121	1.1653	91
		WLAD	0.2223	0.1531	63	0.6987	0.6025	47
		LMS	**0.0236**	**0.0207**	**9**	**0.0119**	**0.0203**	**2**
	n=60	LS	0.4084	0.2139	100	0.9734	0.9587	100
		LAD	0.3185	0.1462	68	0.4882	0.3025	32
		WLAD	0.3119	0.1407	66	0.4542	0.2480	26
		LMS	**0.0083**	**0.0050**	**2**	**0.0160**	**0.0052**	**0.5**
	n=100	LS	0.3726	0.1751	100	1.2325	1.5243	100
		LAD	0.2477	0.1148	66	1.2718	1.6544	109
		WLAD	0.2321	0.0834	48	0.6474	0.4569	30
		LMS	**0.0086**	**0.0024**	**1.4**	**0.0204**	**0.0028**	**0.2**
15%	n=15	LS	0.3315	0.3514	100	1.1285	1.3009	100
		LAD	0.1520	0.2625	75	1.1072	1.4023	108
		WLAD	0.1347	0.2081	59	0.8827	0.9615	74
		LMS	**0.0410**	**0.0545**	**16**	**0.0047**	**0.0528**	**4.1**
	n=30	LS	0.3219	0.2506	100	1.2725	1.6311	100
		LAD	0.1588	0.2017	80	1.3781	1.9326	118
		WLAD	0.0681	0.1221	49	1.1199	1.3329	82
		LMS	**0.0354**	**0.0278**	**11**	**0.0126**	**0.0267**	**2**
	n=60	LS	0.3441	0.1890	100	1.2950	1.6842	100
		LAD	0.1921	0.1336	71	1.3959	1.9645	117
		WLAD	0.1039	0.0709	38	1.0564	1.1813	70
		LMS	**0.0173**	**0.0079**	**4**	**0.0183**	**0.0079**	**0.5**
	n=100	LS	0.3540	0.1686	100	1.3237	1.7585	100
		LAD	0.2210	0.1150	68	1.4162	2.0175	115
		WLAD	0.1024	0.0527	31	1.0787	1.2130	69
		LMS	**0.0089**	**0.0032**	**2**	**0.0199**	**0.0036**	**0.2**

Bold indicates the best method in estimating model parameters with respect to either BIAS, MSE, or RE measure with varying values of sample size..

Although the LAD and WLAD estimators of the parameter θ_1 are robust, they show lower efficiency in terms of MSE compared to the LS estimator (see, RE values in Table 1). It implies that while LAD and WLAD methods are designed to handle outliers effectively, they may not always outperform the LS for certain parameters in specific cases.

Across all outlier scenarios, the LS, LAD, and WLAD estimators of the parameter θ_0 are biased. However, their BIAS values are smaller than those of observed values in dataset without outliers. Also, in terms of MSE measure, the LS estimator outperforms the LAD estimator in certain cases. The LMS estimator demonstrates its superiority as a robust estimation method compared to the other robust methods, LAD and WLAD methods, with the smallest BIAS and MSE values. Additionally, since it is slightly affected by outliers, the LMS estimator is robust against outliers under different outlier scenarios, and it is followed by the WLAD estimator. On the other hand, the LS estimator, while performing well when data do not contain outliers for the slope parameter θ_1 in terms of BIAS measure, is not resistant to outliers. When the percentage of outliers increases, the performance of LS estimator degrades faster, since its BIAS and MSE values get more larger compared to robust methods. In contrast, although the LAD and WLAD estimators are designed to be robust, they do not outperform the LMS estimator when a higher proportion of outliers is present.

4 AN EXAMPLE

In this section, a dataset taken from the literature is analzed to demonstrate the applicability of the estimation methods given in Section 2. For this purpose, we use the wind speed data modeled by using Gumbel distribution, which is originally reported in [15], see Figure 3. The dataset contains the daily mean wind speed for October in 2015 in Sinop, Turkey.

The example process for the data is consist of following steps:

Step 1: Take the daily mean wind speed data as residuals ε_i.

Step 2: Generate the x_i values from the standard normal distribution. x_i:

2.9491,	-0.6300,	-0.0469,	2.6830,
-1.1467,	0.5530,	-1.0765,	1.0306,
0.3275,	0.6521,	-0.2789,	0.2452,
1.4725,	-2.2751,	-1.6333,	0.4155,
-0.6548,	-0.2963,	-1.4969,	-0.9048,
-0.4042,	-0.7258,	-0.8665,	-0.4218,
-0.9427,	1.3419,	-0.9884,	1.8179,
-0.3744,	-1.4517,	-0.6187.	

Step 3: By using the residuals ε_i in Step 1 and x_i values in Step 2, obtain the values of y_i with the following formula introduced in [14]:

$$y_i = x_i + \varepsilon_{(i)},\ i = 1,2,\cdots,31.$$

Step 4: Determine the number of outliers $r = n\xi$ in the data, where $\xi = 0$, 0.05, 0.10 or 0.15.

Step 5: Replace r random selected values with r outliers generated by using the uniform distribution (see, Figure 4).

Step 6: Compute the estimates of model parameters θ_0 and θ_1 using the LS, LAD, WLAD and LMS estimators.

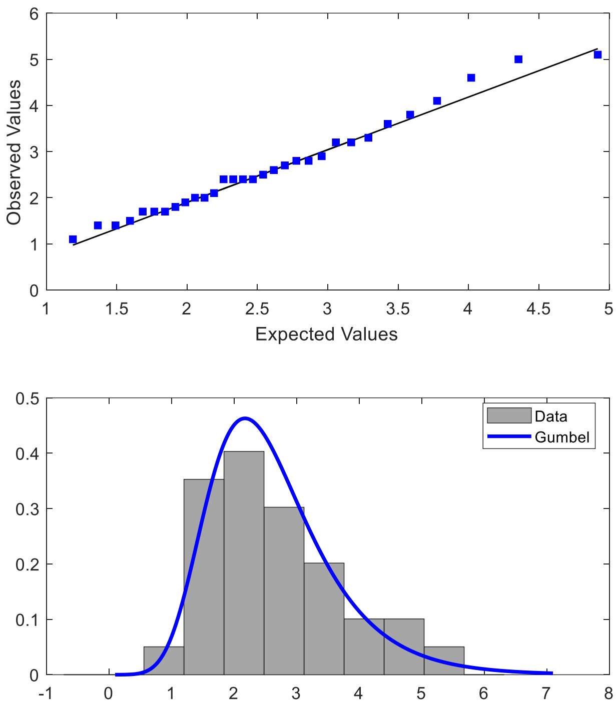

Figure 3: QQ plots showing fits of the data to Gumbel distribution and histogram with fitted densities for the wind speed data

Then, we obtain estimates of the parameters in linear regression model by using the methodologies of the LS, LAD, WLAD and LMS estimators for the data with and without outliers. The results are presented in Table 3.

For all outlier scenarios, the LMS estimator is robust to outliers since its estimates do not change for both the parameters θ_0 and θ_1. Furthermore, the LMS estimator is followed by the WLAD estimator, since it is slightly influenced in the presence of data contamination. In contrast, the LS estimator demonstrates the poorest performance with the greatest difference (see, the values in parentheses in

Table 2) among all considered estimation methods, as expected, under different percentages of outliers. The results show a parallelism with the simulation results presented in Tables 1–2.

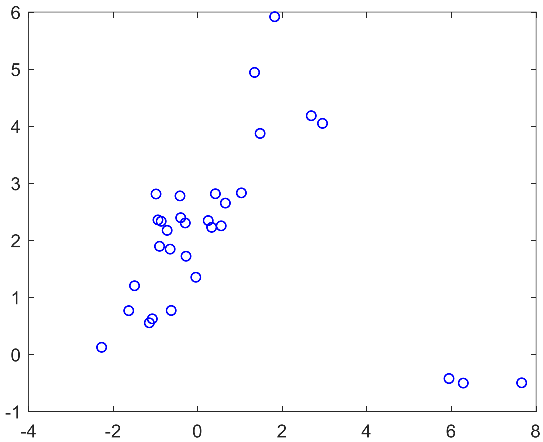

Figure 4: Scatter plot for the data with $\xi = 0.10$.

Table 3. The LS, LAD, WLAD and LMS estimates of model parameters for original data and modified data with different values of percentage of outliers

Outliers Rate	LS $\hat{\theta}_0$	LS $\hat{\theta}_1$	LAD $\hat{\theta}_0$	LAD $\hat{\theta}_1$	WLAD $\hat{\theta}_0$	WLAD $\hat{\theta}_1$	LMS $\hat{\theta}_0$	LMS $\hat{\theta}_1$
0%	2.6197	0.7625	2.3039	0.7004	2.4630	0.5378	2.3494	0.7997
5%	2.2550 (0.3647)*	0.0180 (0.7445)*	2.3325 (-0.0286)*	0.4833 (0.2171)*	2.3351 (0.1279)*	0.4862 (0.0516)*	2.3493 **(0.0001)***	0.7996 **(0.0001)***
10%	2.1744 (0.4453)*	-0.0184 (0.7809)*	2.2610 (0.0429)*	-0.1022 (0.8026)*	2.2190 (0.2440)*	0.0615 (0.4763)*	2.3493 **(0.0001)***	0.7996 **(0.0001)***
15%	2.0236 (0.5961)*	-0.1007 (0.8632)*	2.1453 (0.1586)*	-0.2172 (0.9176)*	2.0811 (0.3819)*	-0.1283 (0.6661)*	2.3493 **(0.0001)***	0.7996 **(0.0001)***

*The values in parentheses are the difference between the estimates obtained with and without outliers in the data.

5 CONCLUSIONS

In this study, we consider the performances of robust estimation methods used to obtain model parameters of a simple linear regression model with Gumbel error distribution for data including different percentages of outliers in x- direction. Estimates of model parameters are obtained by using the least absolute deviation (LAD), weighted least absolute deviation (WLAD), least median of squares (LMS), and the traditional least squares (LS) methodologies. The performance of the estimators is compared by using a Monte-Carlo simulation in terms of the BIAS and mean squared error (MSE) criteria for different data scenarios. Simulation results show that the LMS estimator is the most efficient in estimating both the model parameters θ_0 and θ_1, for almost all cases. The WLAD estimator also demonstrates the second-best performance. Hence, it can be the preferable choice after the LMS estimator.

In the application, the dataset modified by outliers in x- direction is analyzed to evaluate the efficiency of the robust estimators. The findings from the real-world data analysis indicate that the LMS estimators for the regression model parameters have superior performance. As can be seen in all of the simulation results and application, the LMS estimator, which has high breakdown points, is more reliable than the LS, LAD, and WLAD estimators in the presence of data contamination, because it can be resistant to the effects of high levels of contamination in data.

In conclusion, the robustness of the LMS estimator makes it a highly preferable robust alternative estimation method compared to the LAD and WLAD methods in the scenarios where data contamination is a concern.

ACKNOWLEDGMENTS

The authors thank the anonymous referees, whose comments and suggestions have improved the article.

REFERENCES

[1] Zeckhauser, R., & Thompson, M. (1970). Linear regression with non-normal error terms. *The Review of Economics and Statistics*, 280–286.

[2] Tiku, M. L., Islam, M. Q., & Selçuk, A. S. (2001). Nonnormal regression. II. Symmetric distributions. *Communications in Statistics-Theory and Methods*, 30(6), 1021–1045.

[3] Tiku, M. L., & Akkaya, A. D. (2004). Robust estimation and hypothesis testing. New Age International.

[4] Li, J. C. H. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, 48, 1560–1574.

[5] Pek, J., Wong, O., & Wong, A. C. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology*, 9, 2104.

[6] Acitas, S., & Senoglu, B. (2020). Robust change point estimation in two-phase linear regression models: An application to metabolic pathway data. *Journal of Computational and Applied Mathematics*, 363, 337–349.

[7] Islam, M. Q., & Tiku, M. L. (2005). Multiple linear regression model under nonnormality. *Communications in Statistics-Theory and Methods*, 33(10), 2443–2467.

[8] Tiku, M. L., Islam, M. Q., & Selçuk, A. S. (2001). Nonnormal regression. II. Symmetric distributions. *Communications in Statistics-Theory and Methods*, 30(6), 1021–1045.

[9] Cox, D. R., & Hinkley, D. V. (1968). A note on the efficiency of least-squares estimates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(2), 284–289.

[10] Bartolucci, F., & Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, 48(4), 821–834.

[11] Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 26(2), 211–243.

[12] Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374), 296–311.

[13] Al-Adilee, A., & D Mohamed, K. (2014). Linear regression model related to Gumbel distribution. *Journal of Kerbala University*, 10(2), 70–78.

[14] Islam, M. Q., Tiku, M. L., & Yildirim, F. (2001). Nonnormal regression. I. Skew distributions. *Communications in Statistics-Theory and Methods*, 30(6), 993–1020.

[15] Aydin, D. (2018). Alternative robust estimation methods for parameters of Gumbel distribution: an application to wind speed data with outliers. Wind & structures, 26(6), 383–395.

[16] Huber, P. J. (1992). Robust estimation of a location parameter. In Breakthroughs in statistics: Methodology and distribution (pp. 492-518). New York, NY: Springer New York.

[17] Bai, J. (1995). Least absolute deviation estimation of a shift. *Econometric Theory*, 11(3), 403–436.

[18] Olive, D. (2006). Robust estimators for transformed location scale families. *Southern Illinois University. Mailcode*, 4408, 62901–4408.

[19] Chan, N. H., & Peng, L. (2005). Weighted least absolute deviations estimation for an AR (1) process with ARCH (1) errors. *Biometrika*, 92(2), 477–484.

[20] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871–880.

[21] Gumbel, E. J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics*, 12(2), 163–190.

[22] Aydin, D., & Şenoğlu, B. (2015). Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. *Journal of Modern Applied Statistical Methods*, 14(2), 123–140.

[23] Aydin, D. (2018). Estimation of the lower and upper quantiles of Gumbel distribution: an application to wind speed data. *Applied Ecology and Environmental Research*, 16(1), 1–15.

[24] Apéry, Irrationalit´e de ζ (2) et ζ (3), *Asterisque*, 61, 1979, 11–13.
[25] Sapra, S. K. (2003). High-breakdown point estimation of some regression models. *Applied Economics Letters*, 10(14), 875–878.
[26] Yavuz, A. A. (2013). Estimation of the Shape Parameter of the Weibull Distribution Using Linear Regression Methods: Non-Censored Samples. *Quality and Reliability Engineering International*, 29(8), 1207–1219.
[27] Croux, C., Filzmoser, P., Pison, G., & Rousseeuw, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13, 23–36.
[28] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2), 186–199.
[29] Dunsmuir, W. T., & Spencer, N. M. (1991). Strong Consistency and Asymptotic Normality of L1 Estimates of The Autoregressive Moving-Average Model. *Journal of Time Series Analysis*, 12(2), 95–104.
[30] Davis, R. A., & Dunsmuir, W. T. (1997). Least absolute deviation estimation for regression with ARMA errors. *Journal of Theoretical Probability*, 10, 481–497.
[31] Ling, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3), 381–393.
[32] Pan, J., Wang, H., & Yao, Q. (2007). Weighted least absolute deviations estimation for ARMA models with infinite variance. *Econometric Theory*, 23(5), 852–879.
[33] van Zyl, J. M., & Schall, R. (2012). Parameter estimation through weighted least-squares rank regression with specific reference to the Weibull and Gumbel distributions. *Communications in Statistics-Simulation and Computation*, 41(9), 1654–1666.
[34] Rousseeuw, P. J., & Leroy, A. M. (2003). Robust regression and outlier detection. John Wiley & Sons.