Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A new approach using the genetic algorithm for parameter estimation in multiple linear regression with long-tailed symmetric distributed error terms: An application to the Covid-19 data

Abdullah Yağışkaş̄a,⁎, İklim Gedik Balayb, Birdal Şenoğlua

a Department of Statistics, Ankara University, 06100, Ankara, Turkey
b Business School, Ankara Yıldırım Beyazıt University, 06760, Ankara, Turkey

A R T I C L E  I N F O

Keywords:
Multiple linear regression
Long-tailed symmetric distribution
Maximum likelihood
Modified maximum likelihood
Genetic algorithm

A B S T R A C T

Maximum likelihood (ML) estimators of the model parameters in multiple linear regression are obtained using genetic algorithm (GA) when the distribution of the error terms is long-tailed symmetric. The efficiency of the ML estimators obtained using GA with the corresponding ML estimators obtained using other iterative techniques via an extensive Monte Carlo simulation study. Robust confidence intervals based on modified ML estimators are used as the search space in GA. Our simulation study shows that GA outperforms traditional algorithms in most cases. Therefore, we suggest using GA to obtain the ML estimates of the multiple linear regression model parameters when the distribution of the error terms is LTS. Finally, real data of the Covid-19 pandemic, a global health crisis in early 2020, is presented for illustrative purposes.

1. Introduction

Consider the following multiple linear regression model

\[ Y = X\theta + \varepsilon \]

where \( Y = [y_1, y_2, \ldots, y_n]^T \) is a \( n \times 1 \) vector of responses, \( X = [x_{11}, x_{12}, \ldots, x_{1q}]^T \) is a \( n \times q \) non-stochastic design matrix, \( \theta = [\theta_1, \theta_2, \ldots, \theta_q]^T \) is a \( q \times 1 \) vector of unknown coefficients and \( \varepsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T \) is a vector of independent and identically distributed error terms.

Linear regression models are very popular and therefore widely used in many different areas of science, see, for example, Ghosal et al. [11], Jommonkaw et al. [19], Liao et al. [23], Mc Garry et al. [25].

The distribution of the error terms is assumed to be normal in the context of the linear regression model in many cases encountered in real-life problems. However, the distribution of the error terms can depart from normality in practice because of the presence of heavy tails, see, for example related to non-normality, Geary [9], Huber [14], Pearson [27], Tiku et al. [34]. Therefore, in this paper, we strategically choose long-tailed symmetric (LTS) distribution as the distribution of the error terms in the multiple linear regression model since it is more flexible than normal distribution with its heavy tails. It is also used for modeling the outliers existing in the direction of the tails. Thus, LTS distribution is utilized to obtain more robust solutions to the statistical inference problems as an alternative to the normal distribution, see, for example, Tiku et al. [31] and Tiku and Suresh [33].

There exist plenty of studies about the regression models with non-normal error terms in the literature. Islam et al. [17] and Tiku et al. [31] studied parameter estimation and hypothesis testing for the simple linear regression under non-normal error distributions. Then, Islam and Tiku [16] discussed the multiple linear regression model with non-normal error terms. Furthermore, multivariate linear regression models under the assumption of non-normal error terms are considered by Islam [15] and Islam et al. [18].

It is very crucial to estimate the parameters in a multiple linear regression model efficiently. Least squares (LS) method is convenient to estimate the parameters of a multiple linear regression model when the error terms are normally distributed. However, LS estimators lose their efficiencies when the normality assumption is not satisfied [35].

To estimate multiple linear regression model parameters under a non-normal error structure, ML estimation method is widely used because ML estimators are unbiased and fully efficient under certain regularity conditions. However, ML equations cannot be solved analytically when the distribution of the error terms is non-normal. For this reason, numerical methods are generally used to derive ML estimates of the unknown model parameters in literature. It should be noted that some problems such as
non-convergence of iterations, convergence to the wrong root, slow convergence may arise in numerical methods, see Refs. [1,2,28,36,37]. Different than the earlier studies genetic algorithm (GA) which is a population-based heuristic optimization method is utilized to obtain the ML estimates of model parameters in multiple linear regression. It is a powerful method besides being easy for solving an optimization problem and can comfortably be used for large-scale and complex nonlinear optimization problems, see Xia et al. [39]. Further, as Lu et al. [24] stated that satisfactory solutions can easily be obtained according to the optimization objectives, and the shortcomings of numerical optimization methods are overcome by using GA due to its characteristics which make it have outstanding advantages in iterative optimization. Garcia et al. [8] also reported that GA is one of the most robust heuristics automated methods to solve optimization problems. See also Yalcinkaya et al. [40,41] for the details and advantages of GA. The main reason for choosing GA in obtaining the estimates of the parameters is that the use of GA warrants the convergence to the global optimal solution in optimization problems that are extremely complex and suspected to be multi-modal, see Goldberg [12]. We then compare the efficiencies of the ML estimators using GA with corresponding estimators using traditional iterative techniques, such as Newton-Raphson (NR), Nelder-Mead (NM), and iteratively re-weighting algorithm (IRA).

One of the main contributions of this paper is to propose confidence intervals based on Tiku’s [29,30] modified maximum likelihood (MML) estimators for the corresponding multiple linear regression model parameters as the search space in GA. Yalcinkaya et al. [40,41] and Acitas et al. [3] use GA and particle swarm optimization to obtain the ML estimators of the parameters of skew normal distribution and Weibull distribution, respectively. They both strategically use the confidence intervals based on the MML estimators of parameters of interest as search space and show that using the proposed approach provides a narrower search space improving the GA’s performance for convergence.

The usage of heuristic methods is very limited in the context of regression. Therefore, we aim to implement these methods in solving optimization problems in multiple linear regression with non-normal error terms instead of classical approaches. To the best of our knowledge, this is the first study to obtain ML estimates of the parameters of a multiple linear regression model with LTS distributed error terms using GA.

It should be noted that there are many long-tailed symmetric distributions, see, for example, Lange et al. [21] and Lange and Sinhasheimer [22]. Our results can easily be extended to these distributions so they may be the topic of our future studies.

A virus called as Covid-19, which appeared in Wuhan province of China in late 2019, spread rapidly worldwide in early 2020. On July 7, 2020, World Health Organization (WHO) reported that Covid-19 infected over 11 million 500 thousand people worldwide and killed more than 500 thousand of them. In the Covid-19 pandemic, which turned into a global health crisis, the growth rate of cases and the number of deaths per million varies from country to country because of the different characteristics of their governments and people. It is important to model the mortality rate on fighting Covid-19. Therefore, in this study, we analyze the Covid-19 data which includes characteristics of governments and people, both to make a scientific contribution to the fight with Covid-19 and to demonstrate an implementation of our proposed methodology.

The rest of the paper is organized as follows. Section 2 presents the LTS distribution. Section 3 includes the ML estimation for multiple linear regression model parameters when the distribution of error terms is LTS. The details of GA, the procedure of identifying the search space in GA, and the details of IRA are also given in Section 3. The simulation study and its results are presented in Section 4. Real data of the Covid-19 pandemic is examined in Section 5 to show the implementation of the proposed methodology. In the final section, the concluding remarks are given.

### 2. Long-tailed symmetric distribution

Probability density function (pdf) of the LTS distribution is given by

\[
f(\varepsilon) = \frac{1}{\pi \sqrt{k \sigma^2}} \left( \frac{1}{\sqrt{k \sigma^2}} + \frac{1}{\varepsilon^2} \right)^{-\gamma}, \quad -\infty < \varepsilon < \infty
\]

where \( k = 2p - 3, \ p \geq 2, \ \sigma > 0, \) and \( B (\cdot, \cdot) \) is the beta function. The expected value and variance of the LTS distribution are given by \( E(\varepsilon) = 0 \) and \( \text{Var}(\varepsilon) = \sigma^2 \), respectively. Also, the kurtosis value \( \gamma \) is evaluated by \( \gamma = 3(p - 3)/2(p - 5/2) \). Table 1 shows kurtosis values for some representative values of shape parameter \( p \).

| \( p \) | 2.5 | 3.0 | 3.5 | 5.0 | 10 | \( \infty \) |
|-------|-----|-----|-----|-----|----|--------|
| \( \gamma \) | 9.0 | 6.0 | 4.2 | 3.4 | 3.0 |
don’t give the NR and NM algorithms because they are very common methods to obtain the ML estimators in the literature.

3.1. Genetic algorithm

Genetic Algorithm (GA), an iterative population-based search technique proposed by Holland [13], is a very popular heuristic algorithm to find the optimum of the objective function. The idea underlying GA is to imitate the behavior of heredity characteristics of chromosomes on the passing from one generation to another based on the evolutionary mechanisms. Each solution and a set of solutions in every generation are called as chromosome and population, respectively. The following procedure given stage by stage is used during GA.

Generating Initial Population: GA starts with an initial population of N chromosomes generated from the search space via an initialization strategy. Assume that the initial population is denoted by $W^0 = [w_1^0, w_2^0, \ldots, w_N^0]$, where $w = [\theta_0, \theta_1, \ldots, \theta_q, \sigma]$ is a vector of unknown parameters. Here, q and N are the number of the independent variables in the multiple linear regression model and the population size, respectively. Also, the vector of $w^r$, $r = 1, 2, \ldots, N, m = 0, 1, 2, \ldots$ represents the values of the rth chromosome in the population at mth iteration.

Fitness Evaluation: Each chromosome has a fitness value evaluated according to the objective function. In this study, we use $lnl$ as an objective function. $lnl(w^m)$, $m = 0, 1, 2, \ldots$ represents the fitness value of the rth chromosome at iteration m. Firstly, the chromosomes in mth population are sorted and also classified by their fitness values from the best to the worst to obtain new $(m + 1)$ith population. It should be noted that the classifying a chromosome as better or namely the having better fitness value demonstrates the having bigger/smaller fitness value for maximization/minimization problems.

Elitism: A predetermined elite number (E) of chromosomes having the best fitness values are accepted as elite chromosomes and transferred directly to the new population without any modifications.

Selection: At a predetermined selection rate (SR), the worst chromosomes are replaced by new chromosomes generated randomly from the search space.

Mutation and Crossover: Finally, the mutation and crossover operators are applied according to mutation probability (MP) and crossover probability (CP) to the candidate chromosomes selected from the chromosomes except elites via a selection strategy. Here, we prefer the roulette wheel selection strategy. The basic principle of this strategy is that the chromosomes having the better fitness value have a greater chance of being selected.

Convergence Check: So, the new population $[w_1^{m+1}, w_2^{m+1}, \ldots, w_N^{m+1}]$ is obtained. If the convergence criteria is not hold, this process is continued by setting $m = m + 1$. When the process stops, the values of the best chromosome at the last population are called as the estimates of the parameters.

Identifying the search space of GA: We use the confidence intervals based on Tiku’s MML estimators of the parameters $\theta_0, \theta_1, \ldots, \theta_q$ and $\sigma$ as the search space in GA. Then asymptotic 100 (1 $-$ $\alpha$)% confidence intervals for the parameters $\theta_0, \theta_1, \ldots, \theta_q$ and $\sigma$ are given as follows:

\[
CI(\theta_0) = \left( \hat{\theta}_0 - z_{\alpha/2}se(\hat{\theta}_0), \hat{\theta}_0 + z_{\alpha/2}se(\hat{\theta}_0) \right),
\]

\[
CI(\theta_i) = \left( \hat{\theta}_i - z_{\alpha/2}se(\hat{\theta}_i), \hat{\theta}_i + z_{\alpha/2}se(\hat{\theta}_i) \right),
\]

\[
(l = 1, \ldots, q), and
\]

\[
CI(\sigma) = \left( \hat{\sigma} - z_{\alpha/2}se(\hat{\sigma}), \hat{\sigma} + z_{\alpha/2}se(\hat{\sigma}) \right)
\]

Here, $se(\cdot)$ is the standard error of the estimator of interest. In this study, we evaluate the standard errors by using the asymptotic variance -covariance matrix of Tiku’s MML estimators. The asymptotic variance -covariance matrix of the MML estimators is obtained by using the inverse of the expected Fisher information matrix given in Islam and Tiku [16].

MML estimators are explicit estimators of the model parameters and derived by using non-iterative method, see Ref. [29]. To obtain the MML estimators, firstly the likelihood equations given in (4)-(6) are expressed in terms of the ordered statistics (for given $\theta_0$ and $\theta_q, 1 \leq j \leq q$)

\[
e_i = y_i - \bar{y}_0 - \sum_{j=1}^{q} \theta_j x_{ij}, \quad 1 \leq i \leq n
\]

where $(y_{10}, x_{10}, \ldots, x_{10q})$ is called as the concomitant vector of observations corresponding to ith ordered $e_i$. Then the MML estimators of multiple linear regression model parameters when the errors have LTS distribution are formulated as

\[
\hat{\theta}_0 = \bar{y}_0 - \sum_{j=1}^{q} \hat{\theta}_j x_{0j},
\]

\[
\hat{\sigma} = K + D\hat{\sigma}
\]

\[
\Delta = B + \sqrt{(B^2 + 4nC)} / 2\sqrt{n(n-q-1)}
\]

where

\[
\bar{y}_1 = \sum_{i=1}^{n} \beta_i y_i / \delta, \quad \bar{x}_1 = \sum_{i=1}^{n} \beta_i x_{ij} / \delta, \quad \delta = \sum_{i=1}^{n} \beta_i
\]

\[
K = (X'X)^{-1}(X'y) = (K_1), \quad \beta = diag(\beta_i),
\]

\[
D = (X'X)^{-1}(X'1) = (D_1), \quad a = diag(a_i),
\]

\[
B = 2p \sum_{i=1}^{n} a_i \left( y_i - \bar{y}_1 - \sum_{j=1}^{q} K_1 x_{ij} - \bar{x}_1 \right)^2
\]

\[
C = 2p \sum_{i=1}^{n} a_i \left( y_i - \bar{y}_1 - \sum_{j=1}^{q} K_1 x_{ij} - \bar{x}_1 \right)^2
\]

Here,

\[
\alpha_i = \left( \frac{2/k} {1 + (1/k)t_{0}^2} \right)^{1/2},
\]

\[
\beta_i = \left( \frac{1 - (1/k)t_{0}^2} {1 + (1/k)t_{0}^2} \right)^{1/2}
\]

where $t_{0}$ is the expected value of the ordered statistics $Z_{0i} = e_i / \sigma$, i. e., $t_{0} = E(Z_{0i})$. If $C$ in (18) is a negative value, then we use the following $\alpha_i'$ and $\beta_i'$,

\[
\alpha_i' = \left( \frac{1/k} {1 + (1/k)t_{0}^2} \right)^{1/2},
\]

\[
\beta_i' = \left( \frac{1 - (1/k)t_{0}^2} {1 + (1/k)t_{0}^2} \right)^{1/2}
\]
They are as efficient. Also a remarkable property of MML estimators is that

\[ \theta^p \]

...and \( \theta \)

\[ \gamma \]

\[ = \frac{1}{1 + (1/k)^{\ell(\ell) / 2}} \]  

(1 ≤ \ell ≤ n)

(22)

\[ \theta_p^{(n+1)} = \left( \frac{\sum_{i=1}^{n} \phi_{i}^{(n)} x_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)}}{\sum_{i=1}^{n} x_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)} \theta_{i}^{(n)}} \right) \]  

instead of \( \alpha \) and \( \beta \), respectively for mathematical convenience; see Ref. [16] for further details. MML estimators are asymptotically equivalent to the ML estimators, therefore, under regularity conditions, they are fully efficient. Also a remarkable property of MML estimators is that they are as efficient as the ML estimators even for small sample sizes, see Refs. [5,37,38]. The high performance of MML estimators is shown in plenty of studies in the literature; see, for example, [6,17,20].

### 3.2. Iterative re-weighting algorithm

In this study, we provide IRA to compute the ML estimates of the parameters of interest. It should be noted that the IRA is EM type algorithm and its convergence is guaranteed, see Ref. [7]. The steps of IRA used in this study is given below.

**Step 1:** Identify the initial values \( \theta_{0} \), \( \theta_{1} \), ..., \( \theta_{p} \) and \( \sigma \).

**Step 2:** Compute the following equations

\[ \theta^{(n+1)} = \frac{2p}{\sigma^{2}} \sum_{i=1}^{n} \frac{y_{i} - \theta_{0}^{(n)} - \sum_{j=1}^{d} \theta_{j}^{(n)} x_{ij}^{(n)}}{w_{i}^{(n)}} \]  

(23)

where

\[ w_{i}^{(n)} = 1 + \frac{1}{k \sigma^{2}} \]  

(24)

and \( y_{i}^{(n)} \) is the number of iterations and \( l = 1, \ldots, q \) is the number of regression coefficients.
Step 3: Stop the iterations when the conditions $|\theta_0^{(m+1)} - \theta_0^{(m)}| < \epsilon$, $|\theta_i^{(m+1)} - \theta_i^{(m)}| < \epsilon$ $(i = 1, \ldots, q)$ and $|\sigma^{(m+1)} - \sigma^{(m)}| < \epsilon$ are hold. Here, $\epsilon > 0$ is a predetermined small constant.

4. Simulation study

In this section, we compare the efficiencies of the proposed parameter estimators for the multiple linear regression model with LTS distributed error terms via an extensive Monte Carlo simulation study. We simulate for $q = 1, 2, 3, 4$ separately and see that similar results are obtained for all $q$ values, therefore, we preferably give the results of $q = 3$ for the sake of brevity. The ML estimates of the parameters $\theta_0, \theta_1, \theta_2, \theta_3$, and $\sigma$ are obtained by using the GA, IRA, NR, and NM algorithms. It should be noted that the value of parameter $p$ is assumed to be known throughout this section. All computations are conducted using the R statistical software environment.

For the computations of the GA algorithm, we use ‘ga’ function with the following settings; that are the population size $N = 500$, the mutation probability $MP = 0.2$, the crossover probability $CP = 0.8$, and the elite number $E = 4$. The maximum number of iterations is taken to be 1000. The other conditions are taken to be default in the ‘ga’ function.

Without loss of generality, we take $\theta_0 = 0, \theta_1 = 1, \theta_2 = 1, \theta_3 = 1$, and $\sigma = 1$. The values of $x_j$ $(i = 1, \ldots, n; j = 1, 2, 3)$ in the design matrix are randomly generated from Uniform $(0, 1)$ distribution. The sample sizes are taken to be $n = 20, 50, 100, 200$, and 500 for small, moderate, large, very large, and extremely large sample size scenarios, respectively. We strategically choose $p = 3$ and $p = 5$ in LTS distribution to compare the effects of kurtosis levels $\gamma = 9$ and $\gamma = 4.2$, respectively. It is known that LTS distribution with $p = 3$ and $p = 5$ are equivalent to the Student’s $t$ distribution with 5 and 9 degrees of freedom, respectively. Each Monte Carlo simulation run is independently replicated $[100,000/n]$ times where $\lfloor \cdot \rfloor$ denotes the greatest integer function.

The simulated mean, mean square error (MSE), and deficiency (DEF) criteria are used in the comparisons. The mathematical expression of the DEF criterion which is used to compare joint efficiencies of the estimators is given below

$$DEF = MSE(\hat{\theta}_0) + MSE(\hat{\theta}_1) + MSE(\hat{\theta}_2) + MSE(\hat{\theta}_3) + MSE(\hat{\sigma}).$$  

We report the mean, MSE, and DEF values for the ML estimators of the parameters in Table 2. We present the discussions about the simulation results below.

It is seen from Table 2 that the NR algorithm gives non-meaningful values for the small sample sizes (i.e., $n = 20$) for all shape parameter values.

Considering the mean values in Table 2, it can be easily said that the biases for all estimators (except for $n = 20$) are negligible.

For $\theta_0$: GA has better performance than the NM and NR with the smallest MSE values for all shape parameter values and all sample sizes. On the other hand, IRA outperforms all methods in all of the cases according to the MSE criterion. Therefore, GA has the second-best performance in estimating $\theta_0$ since it has quite small MSE values.

For $\theta_1, \theta_2, \theta_3$, and $\sigma$: GA has the best performance among NR, NM, and IRA according to MSE criterion for all cases.

The simulation results show that GA outperforms NM and NR algorithms according to the DEF criterion because of providing the lowest deficiency in all of the cases. However, IRA has a little bit smaller deficiency than GA for a few cases, i.e., $n = 100, 200, 500$ when $p = 5$ and $n = 500$ when $p = 3$. On the other hand, GA is more efficient than IRA in the other cases (60% of cases). Therefore, GA outperforms IRA according to DEF criterion in most of the cases.

As a result, it can be said that GA is more preferable than the other algorithms to obtain the ML estimators of the parameters of the multiple
linear regression model when the error terms have LTS distribution because of the superior performance of the GA.

5. Application

In this section, a real-life data set about the Covid-19 pandemic is analyzed to show the implementation of the proposed methodology, see Ref. [10] for the detailed backgrounds of the data. The data set consists of measurements about some indicators and indices for 52 countries in the context of the Covid-19 pandemic. Therefore, it is called as Covid-19 data in the rest of this section.

Covid-19 data include \( n = 52 \) observations about mortality rates (deaths from Covid-19 per million people), cultural tightness, government efficiency, and growth rates of the virus. We use the following multiple linear regression model

\[
y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \epsilon_i \quad (1 \leq i \leq n)
\]

and obtain the estimated regression equation. Here, the dependent (response) variable \( Y \) represents the mortality rate, and the independent (explanatory) variables \( X_1, X_2, \) and \( X_3 \) represent the cultural tightness, the government efficiency, and the growth rate, respectively.

In regression analysis, it is an important concern to identify problems such as heteroscedasticity, multicollinearity and to take preventive action. Since the mortality rates for some countries (e.g., Italy, Spain) are far greater than those of the other countries, the mortality rates have a high skewness and variance. Such response variables in a regression model can lead to heteroscedasticity. To prevent violations of homoscedasticity, we apply log-transformation on the raw mortality rates similar to Ref. [10]. In the regression analysis, the log-transformed mortality rates are used as the response.

The multicollinearity has been checked by Ref. [10] via the variance inflation factors, and it is seen that there is no problematic multicollinearity among the explanatory variables. To see the numerical value of the relationship between explanatory variables, we give the correlations between them in Fig. 1. The entries on the main diagonal show the distribution of the variables, and the entries below and above the main diagonal display the scatter plots of each pair of variables and the Pearson correlation coefficients, respectively.

To find the appropriate regression equation which provides a better fit to the Covid-19 data, we obtain the ML estimates of the model parameters in multiple linear regression by using GA and IRA algorithms because of their superior performances in the simulation study.

We consider two different cases in identifying the estimate of the shape parameter \( p \) of LTS distribution. In Case 1, it is estimated by using profile likelihood (PL) method given in Ref. [40]. In Case 2, we obtain the estimates of all parameters including the shape parameter \( p \) by using GA, simultaneously.

To compare the fitting performances of the regression equations based on the ML estimates obtained by using GA and IRA algorithms, we use AIC (Akaike information criterion), and RMSE (The root mean square error) criteria.

AIC is a popular measure proposed by Akaike [4] to compare the fitting performances of the possible models according to the log-likelihood values and the number of estimated parameters. The mathematical expression of AIC is

\[
AIC = 2p - 2lnL
\]

where \( p \) is the number of estimated parameters. AIC is a useful measure when the number of estimated parameters is different for the compared models. It should be noted that the number of estimated parameters is different for Case 1 and Case 2 in our problem.

It is known that the model having the lowest AIC value and the lowest RMSE value (the closest to zero) among the possible models provides the best fit to the data.

To identify the distribution of the residuals, a Q-Q plot of the estimated residuals

\[
\hat{\epsilon}_i = y_i - \hat{\theta}_0 - \sum_{j=1}^{3} \hat{\theta}_j x_{ji} \quad (1 \leq i \leq 52)
\]

is given in Fig. 2. It is seen that there exist residuals which are grossly anomalous in Covid-19 data. Therefore, we set aside observations that correspond to these outliers (i.e., \(-7.82, -5.77, \) and \(-5.30\)), see also the box-plot in Fig. 3.

After excluding the outliers from regression analysis, we estimate the residuals obtained from the regression model for the remaining \( n = 49 \) observations. A Q-Q plot of these residuals

![Fig. 2. LTS Q-Q Plot for the residuals (n = 52).](image)

![Fig. 3. Box-Plot for the residuals (n = 52).](image)
To test the following null hypothesis

\[ H_0: \text{The error terms are distributed as LTS} \]

versus the alternative

\[ H_1: \text{The error terms are not distributed as LTS}, \]

K-S test statistic \(D_{ks}\) is obtained as follows:

\[ D_{ks} = \sup_{x} (|F_{\hat{\theta}}(x) - F_{\theta}(x)|) \]  

where \(F_{\hat{\theta}}(\cdot)\) is the cdf of a LTS distribution with a known shape parameter \(p\) and \(F_{\theta}(\cdot)\) is the empirical cdf for the error terms. At significance level \(\alpha\), if the calculated value \(D_{ks}\) is greater than the tabulated value \(d_{\alpha}\) given in Ref. \[26\] or equivalently if the corresponding \(p\)-value is less than \(\alpha\), then the null hypothesis \(H_0\) is rejected.

The estimates of the regression parameters, the bootstrap standard errors of the corresponding estimates (given in the parenthesis), AIC, RMSE, and the K-S test statistics (i.e. \(D_{ks}\)) values are given in Table 3.

According to the K-S results based on GA and IRA estimates of the model parameters, the null hypothesis is not rejected at the significance level \(\alpha = 0.05\) for both tests since the \(D_{ks}\) values are less than the corresponding table value \(d_{0.05} = 0.17128\), see Table 3. This result implies that the distribution of the error terms obtained from the regression equation based on GA and IRA estimates of the model parameters is LTS. The Q-Q plot given in Fig. 4 supports this result, visually.

It is clear from Table 3 that the proposed method using GA has lower AIC and RMSE values than those of IRA in Case 1. Furthermore, GA has much smaller bootstrap standard errors than IRA for all parameter estimates. Therefore, GA is more reliable and preferable than IRA. Additionally, it is obvious from Table 3 that using GA is superior than using PL in model fitting according to the AIC and RMSE results given in Case 1 and Case 2. These conclusions show the superiority of the GA. As a result, we advice to use GA to obtain the estimates of the parameters (including the shape parameter \(p\)) in multiple linear regression model with LTS distributed error terms.

Additionally, to investigate the aspects of the robustness of the ML estimates based on GA and IRA methods to outliers, we give the results of the regression analysis for the data including outliers (i.e., \(n = 52\)), see Table 4. It is seen from Table 3 and Table 4 that the proposed method using GA has lower AIC and RMSE values than those of IRA. Furthermore, GA has much smaller bootstrap standard errors than IRA for all parameter estimates. Therefore, GA is more reliable and preferable than IRA. Additionally, it is obvious from Table 3 that using GA is superior than using PL in model fitting according to the AIC and RMSE results given in Case 1 and Case 2. These conclusions show the superiority of the GA. As a result, we advice to use GA to obtain the estimates of the parameters (including the shape parameter \(p\)) in multiple linear regression model with LTS distributed error terms.

Table 3

| Method | Case 1: \(p = 10.00\) | Case 2: \(p = 9.68\) |
|--------|------------------|------------------|
| GA     | \(-1.7189, 0.1088\) | \(-3.1170, 0.1200\) |
| IRA    | \(-0.1620, 0.1382\) | \(-1.4264, 0.1200\) |
| AIC    | \(190.8778, 199.2834\) | \(186.2225, 187.5000\) |
| RMSE   | \(1.5395, 1.6552\) | \(1.4349, 1.4382\) |
| \(D_{ks}\) | \(0.0620, 0.0916\) | \(0.0567, 0.0572\) |

Table 4

| Method | Case 1: \(p = 2.24\) | Case 2: \(p = 2.52\) |
|--------|------------------|------------------|
| GA     | \(-3.3888, 0.1828\) | \(-3.6262, 0.1643\) |
| IRA    | \(-0.2295, 0.1643\) | \(-1.4299, 0.7277\) |
| AIC    | \(220.598, 237.293\) | \(221.6684, 222.1225\) |
| RMSE   | \(1.9515, 2.2591\) | \(1.9253, 1.9253\) |
| \(D_{ks}\) | \(0.0847, 0.0881\) | \(0.0997, 0.0997\) |
performance of GA is less sensitive to the outliers than that of IRA. It is clear that AIC and RMSE values decrease for the censored sample for both models based on GA and IRA, this is an indication of the negative effect of the outliers on the estimated regression equation. However, the reduction rate is much bigger for the model based on IRA.

6. Conclusion

In this study, we focus the ML estimation of the parameters of the multiple linear regression model when the underlying distribution of error terms is LTS. It should be noted that the ML estimators of the parameters cannot be obtained analytically. Therefore, we resort to GA and traditional NR, NM and IRA algorithms. To improve the performance of GA, we use the robust confidence intervals based on the MML estimators of the regression parameters as the search space. We compare the efficiencies of the ML estimators obtained by using mentioned algorithms in terms of MSE and DEF criteria. Our simulation study shows that GA outperforms other traditional algorithms in most of the cases to obtain the ML estimates. Eventually, we strongly advise using GA to obtain the ML estimates of the parameters of the multiple linear regression model when the error terms have LTS distribution because of the superior performance of the GA.

CRediT authorship contribution statement

Abdullah Yalçıkaya: Conceptualization, Methodology, Software, Formal analysis, Validation, Data curation, Investigation, Writing – original draft, Review & editing. İklim Gedik Bayal: Conceptualization, Methodology, Software, Formal analysis, Resources, Validation, Writing – original draft, Review & editing. Birdal Şenoğlu: Conceptualization, Methodology, Supervision, Formal analysis, Validation, Writing original draft, Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Acitas, P. Kasap, B. Senoglu, O. Arslan, One-step M-estimators: Jones and Faddy's skewed t-distribution, J. Appl. Stat. 40 (7) (2013) 1545–1560.
[2] S. Acitas, P. Kasap, B. Senoglu, O. Arslan, Robust estimation with the skew t-distribution, Pakistan Journal of Statistics 29 (4) (2013) 409–430.
[3] S. Acitas, C.H. Adsadag, B. Senoglu, A new approach for estimating the parameters of Weibull distribution via particle swarm optimization: an application to the strengths of glass fibre data, Reliab. Eng. Syst. Saf. 183 (2019) 116–127.
[4] H. Akaike, Maximum likelihood identification of Gaussian autoregressive moving average models, Biometrika 60 (2) (1973) 255–265.
[5] G.K. Bhattacharyya, The asymptotics of maximum likelihood and related estimators based on type II censored data, J. Am. Stat. Assoc. 80 (390) (1985) 398–404.
[6] N. Celik, B. Senoglu, Robust estimation and testing in one-way ANOVA for Type II censored samples: skew normal error terms, J. Stat. Comput. Simulat. 88 (7) (2018) 1382–1392.
[7] N. Celik, B. Senoglu, O. Arslan, Estimation and testing in one-way ANOVA when the errors are skew-normal, Rev. Colomb. Estadística 38 (1) (2015) 75–91.
[8] A.M. Garcia, I. Sanet, M. Bouillon, R. Crecente, Calibration of an urban cellular automation model by using statistical techniques and a genetic algorithm. Application to a small urban settlement of NW Spain, Int. J. Geogr. Inf. Sci. 27 (8) (2013) 1593–1611.
[9] J.C. Geary, Testing for normality, Biometrika 34 (3/4) (1947) 295–252.
[10] M.J. Gelfand, J.C. Jackson, X. Pan, D. Nau, M.M. Dagher, P.V. Lange, C. Chiu, The importance of cultural tightness and government efficiency for understanding COVID-19 growth and death rates. Preprint on PsyArXiv. https://doi.org/10.31234/osf.io/m7f8a, 2020.
[11] S. Ghosal, S. Sengupta, M. Majumder, B. Sinha, Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases/March 14th 2020), Diabetes & Metabolic Syndrome: Clin. Res. Rev. 14 (4) (2020) 311–315.
[12] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, MA, 1989.
[13] J. Holland, Adaptation in Natural and Artificial System: an Introduction with Application to Biology, Control and Artificial Intelligence, University of Michigan Press, Ann Arbor, 1975.
[14] P.J. Huber, Robust Statistics, Springer Berlin Heidelberg, 2011, pp. 1248–1472.
[15] M.Q. Islam, Estimation and hypothesis testing in multivariate linear regression models under non normality, Commun. Stat. Theor. Methods 46 (17) (2017) 8521–8543.
[16] M.Q. Islam, M.L. Tiku, Multiple linear regression model under nonnormality, Commun. Stat. Theor. Methods 33 (10) (2005) 2443–2467.
[17] M.Q. Islam, M.L. Tiku, F. Yildirim, Nonnormal regression. I. Skew distributions, Commun. Stat. Theor. Methods 30 (6) (2001) 993–1020.
[18] M.Q. Islam, F. Yildirim, M. Yazici, Inference in multivariate linear regression models with elliptically distributed errors, J. Appl. Stat. 41 (8) (2014) 1746–1766.
[19] S. Jommonkavoo, S. Utra, V. Ratanavanzha, Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression, and path analysis models, Sustainability 12 (1) (2020) 295.
[20] Y.M. Kantar, B. Senoglu, A comparative study for the location and scale parameters of the Weibull distribution with given shape parameter, Comput. Geosci. 34 (12) (2008) 1900–1905.
[21] K.L. Lange, R.J. Little, J. Taylor, Robust statistical modelling using the t-distribution, J. Am. Stat. Assoc. 84 (1989) 881–896.
[22] K.L. Lange, J.S. Sinheimer, Normal/independent distributions and their applications in robust regression, J. Comput. Graph. Stat. 2 (1993) 175–198.
[23] K. Liao, E.S. Park, J. Zhang, L. Cheng, D. E. Qi, Q. Ying, J.Z. Yu, A multiple linear regression model with multiplicative log-normal error term for atmospheric concentration data, Sci. Total Environ. 767 (2021) 144282.
[24] X. Lu, Y. Wu, J. Lian, Y. Zhang, C. Chen, P. Wang, L. Meng, Energy management of hybrid electric vehicles: a review of energy optimization of fuel cell hybrid power system based on genetic algorithm, Energy Convers. Manag. 205 (2020) 112474.
[25] K. McGarry, S.A. Siedlecki, J. Salisbury, S.R. Allin, Multiple linear regression models for reconstructing and exploring processes controlling the carbonate system of the northeast US from basic hydrographic data, J. Geophys. Res.: Oceans 126 (2) (2021), e2020JC016480.
[26] L.H. Miller, Table of percentage points of Kolmogorov statistics, J. Am. Stat. Assoc. 51 (1956) 111–121.
[27] E.S. Pearson, The analysis of variance in cases of non-normal variation, Biometrika 23 (1/2) (1931) 114–133.
[28] S. Puthenpurapu, N.K. Sinha, Modified maximum likelihood method for the robust estimation of system parameters from noisy data, Automatica 22 (2) (1986) 231–235.
[29] M.L. Tiku, Estimating the mean and standard deviation from a censored normal sample, Biometrika 54 (1–2) (1967) 155–165.
[30] M.L. Tiku, Estimating the parameters of normal and logistic distribution from censored samples, Aust. J. Stat. 10 (2) (1968) 64–74.
[31] M.L. Tiku, M.Q. Islam, A.S. Selcuk, Nonnormal regression. II. Symmetric distributions, Commun. Stat. Theor. Methods 30 (6) (2001) 1021–1045.
[32] M.L. Tiku, S. Kumra, Expected values and variances and covariances of order statistics for a family of symmetric distributions (Student's t), selected tables in mathematical statistics 8 (1985) 141–270.
[33] M.L. Tiku, R.P. Suresh, A new method of estimation for location and scale parameters, J. Stat. Plans. Inferencce 30 (2) (1992) 281–292.
[34] M.L. Tiku, W.K. Wong, G. Bian, Estimating parameters in autoregressive models in non-normal situations: symmetric innovations, Commun. Stat. Theor. Methods 28 (2) (1999) 315–341.
[35] J.W. Tukey, A survey of sampling from contaminated distributions, Contributions to probability and statistics (1960) 448–485.
[36] D.C. Vaughan, On the Tiku-Suresh method of estimation, Commun. Stat. Theor. Methods 21 (2) (1992) 451–469.
[37] D.C. Vaughan, The generalized scant hyperbolic distribution and its properties, Commun. Stat. Theor. Methods 31 (2) (2002) 219–238.
[38] D.C. Vaughan, M.L. Tiku, Estimation and hypothesis testing for a nonnormal bivariate distribution with applications, Math. Comput. Model. 32 (1–2) (2000) 53–67.
[39] Z. Xia, K. Mao, S. Wei, X. Wang, Y. Fang, S. Yang, Application of genetic algorithm-support vector regression model to predict damping of cantilever beam with particle damper, J. LowFreq. Noise Vib. Act. Contr. 36 (2) (2017) 138–147.
[40] A. Yalcinkaya, B. Senoglu, U. Yolcu, Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm, Swarm and Evolutionary Computation 38 (2018) 127–138.
[41] A. Yalcinkaya, U. Yolcu, B. Senoglu, Maximum likelihood and maximum product of spacings estimations for the parameters of skew-normal distribution under doubly type II censoring using genetic algorithm, Expert Syst. Appl. 168 (2021) 114407.