MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 3, pp. 950–956 (2025)

Weighted mixture regression model based on t-distribution in the presence of a leverage point

Hassan S. U.

Department of Statistics, University of Al-Qadisiyah, Aljamea'a Str., Diwanyiah, Iraq

(Received 12 July 2024; Accepted 18 July 2025)

The Expectation-Maximization (EM) algorithm is an efficient method for estimating the parameters of a mixture regression model in the presence of outliers in the Y-direction. Unfortunately, this method breaks down when leverage points are present in the dataset. The most common procedure used in the literature involves removing leverage points after identifying them with single detection methods. However, some authors have pointed out that single detection methods can be inaccurate and have therefore proposed multiple diagnostic approaches. This manuscript proposes the Weighted EM (WEM) method to address the problem of leverage points without requiring data deletion. Moreover, it builds upon the DRGP (RMVN) framework, which is one of the multiple diagnostic methods. Real data and simulation studies were conducted to evaluate the efficiency of the proposed method compared to existing approaches. The results show that the WEM method is more robust and reliable than other methods, particularly when sample sizes are small.

Keywords: *robust; mixture regression; leverage point; DRGP.*

2010 MSC: 62P20, 62-08 **DOI:** 10.23939/mmc2025.03.950

1. Introduction

The Mixture Regression model has been used in many scientific fields, such as econometrics, engineering, biology, and others, due to its ability to model the relationship between independent and dependent variables. Quandt [1] and Quandt and Ramsay [2] note that there may be more than one pattern of data across a single dataset, and variables are likely to cluster according to these patterns. The structure of such data is constructed via unknown latent groups of variables would lead to unknown regression models. In other words, the regression probabilistic model in terms of latent class variable Z such that $Z = i$, the linear regression model can be written as follows,

$$Y = X\beta + \varepsilon, \tag{1}$$

where X is the design matrix of $(p + 1)$ independent variables with constant, β_i is the regression parameters of g sub-populations (groups), ε_i is the error term which has to be independent of x with density $f_i(\cdot)$ and mean zero, and Y is the dependent variable which has the same distribution of ε_i with different parameters. The maximum Likelihood Estimation method (MLE) which is one of the best methods to estimate the regression model parameters when the distribution of errors is normal, is hard to derive when errors follow such a distribution. Moreover, it is not resistant to the presence of outliers. Consequently, alternative methods should take into account two aspects: easy computation and robustness. Dempster [3] proposed the Expectation–Maximization (EM) algorithm which is robust to outliers and easy to calculate but it is not resistant to leverage points. This issue of robustness in mixture regression models has been given the attention of the researchers in statistics literature. Markatou [4] and Shen et al. [5] tried to reduce the impact of outliers in mixture regression models by allocate down weights for each data point. Peel and McLachlan [6] suggested robust version of mixture regression relying on the t-distribution. That is by replacing the normal density function with the t-distribution in the mixture regression model as a robust procedure to overcome the problem of outliers. This procedure is reasonable because the normal density is considered a special case of t-distribution when the number of degrees of freedom tends to infinity. Song et al. [7] and Yao et al. [8]

considered error distributions such as Laplace and t-distribution, respectively. Both distributions can be expressed by scale mixture of a normal distribution. Yao et al. [8] proposed robust mixture regression by incorporating the approach of mixture t-distributions of Peel and McLachlan [6]. This paper assumes there may be a problem with accurately identifying all leverage points in robust mixture regression model, particularly when at least one high leverage point is present. This is because these methods use individual diagnostic methods. Thus, the target of this paper is to incorporate multiple diagnostic measures, DRGP (RMVN) with EM to improve the performance of robust mixture regression model when the random errors follow t-distribution. The rest of this paper is organized to present the Mixture t-distribution in Section 2. Section 3 describes Robust Mixture Regression with t-distribution by incorporating Multiple Diagnostic Measure. Section 4 includes simulation study; Section 5 presents a real data example and the conclusion appears in Section 6.

2. Mixture t-distribution

Let the latent class variable Z is independent of X, the probability of $(Z = i)$ equals π_i where $i = 1, 2, \ldots, g$, then the conditional density function without observing Z is as follows,

$$f(y_j, X_j, \Phi) = \sum_{i=1}^{g} \pi_i f(y_j - X_i^{\intercal}\beta; \sigma_i, v_i), \quad j = 1, 2, \ldots, n, \tag{2}$$

where $f(\cdot)$ is the error density of t-distribution with v degree of freedom, scale parameter σ, $\Phi = (\pi_1, \pi_2, \ldots, \pi_g, \beta_1, \beta_2, \ldots, \beta_g, \sigma_1, \sigma_2, \ldots, \sigma_g)'$ is the unknown parameter vector,

$$f\left(e_i; \sigma_i, v_i\right) = \frac{\Gamma\left(\frac{v_i+p}{2}\right)|\sigma|^{-0.5}}{\left(\sqrt{\pi_i v_i}\right)^p \Gamma\left(\frac{v_i}{2}\right)\left\{1 + \frac{e_j^2}{v_i \sigma_i}\right\}^{0.5(v_i+p)}}. \tag{3}$$

In general, the Maximum Likelihood Estimate method (MLE) is used to find $\widehat{\Phi}$

$$\log f\left(y_j, x_j, \Phi\right) = \sum_{j=1}^{n} \log\left(\sum_{i}^{g} \pi_i \; f_i\left(y_j; x_j'\beta_i, \sigma_i\right)\right) \tag{4}$$

and then $\widehat{\Phi} = \arg\max_{\Phi} f\left(y_j, x_j, \Phi\right)$.

Since Eq. (4) does not have a solution, Peel and McLachlan [6] introduced the EM algorithm that is sensitive to high leverage points. Yao et al. [8] modified the EM algorithm by applying their suggested method after adaptively trimming leverage points.

3. Robust mixture regression with t-distribution

Assume that $Z_{ij} = 1$ if the jth observation is from the ith component and 0 otherwise, such that $Z_i = (Z_{j1}, Z_{j2}, \ldots, Z_{jg})'$. Then, the complete log-likelihood function for the complete data set,

$$\ell_n^c(\Phi) = \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij} \log\left\{\pi_i \; f_i(y_j - x_j'\beta_i; \sigma_i, v_i)\right\}, \tag{5}$$

Peel and McLachlan [6] pointed out that this maximizer has no explicit solutions for β_i, σ and v_i, respectively. Yao et al. [8] found that the above problem can be solved when the t-distribution expressed as a scale mixture of normal distributions. They assumed ζ is another latent variable such that $(y|\zeta)$ is normally distributed as $N(\mu, \sigma^2|\zeta)$ and has density function,

$$f(y; \mu, \sigma|\zeta) = \frac{1}{(2\pi)^{0.5}}|\sigma/\zeta|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(y-\mu)'(\sigma|\zeta)^{-1}(y-\mu)\right), \tag{6}$$

where ζ follows gamma distribution with shape and scale parameters both equal to $\frac{1}{2}v$, respectively, and the $\mathrm{cor}(\zeta, Z) \approx 0$, then the density function of ζ is

$$f\left(\zeta; \tfrac{1}{2}v, \tfrac{1}{2}v\right) = \frac{1}{\Gamma\left(\frac{1}{2}v\right)}\left(\tfrac{1}{2}v\right)^{-\left(\frac{1}{2}v\right)} y^{\left(\frac{1}{2}v-1\right)} e^{-\frac{2\zeta}{v}}. \tag{7}$$

However, previous authors found that the complete likelihood function of mixture t-distribution model for y, Z and ζ can be written as follows,

$$\begin{aligned}\ell_n^c(\Phi; y, \zeta, Z) = &\sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\log(\pi_i) + \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\log\left\{f\left(\zeta_j; \tfrac{1}{2}v_i, \tfrac{1}{2}v_i\right)\right\} \\ &+ \sum_{j=1}^{n}\sum_{i=1}^{g} Z_{ij}\left\{-\tfrac{1}{2}\log\left(2\pi\sigma^2\right) + \tfrac{1}{2}\log(\zeta_j) - \tfrac{\zeta_j e_j^2}{2\sigma^2}\right\}.\end{aligned} \tag{8}$$

3.1. The expectation (E-Step)

If the degrees of freedom in the second term of Eq. (8) are known, the E-step involves finding the expectation of the complete data log-likelihood, $E\left[\ell_n^c(\Phi)|y, x, \Phi^{(k)}\right]$ as follows,

$$(1) \qquad E\left[Z_{ij}|x, y, \Phi^{(k)}\right] = \Phi_{ij}^{(k+1)} = \frac{\pi_i^{(k)}\left(\dfrac{\Gamma\left(\frac{v_i+p}{2}\right)\frac{1}{\left|\sigma_i^{(k)}\right|}}{\left(\sqrt{\pi_i^{(k)}v_i}\right)^p\Gamma\left(\frac{v_i}{2}\right)\left\{1+\frac{e_j^{2(k)}}{v_i\sigma_i^{(k)}}\right\}^{0.5(v_i+p)}}\right)}{\sum_i^g \pi_\iota^{(k)} f\left(y_j - x_j\beta_\iota^{(k)}; \sigma_i^{(k)}, v_\iota\right)};$$

$$(2) \qquad E\left[\zeta_j|y, Z_{ij}=1, \Phi^{(k)}\right] = \zeta_{ij}^{(k+1)} = \frac{v+1}{v + \left\{\frac{\left(y_j - x_j\beta_i^{(k)}\right)}{\sigma_i^{(k)}}\right\}^2}.$$

3.2. The maximization (M-Step)

The M-step is to update *parameters* by maximizing $\left[\ell_n^c(\Phi)|y, x, \Phi^{(k)}\right]$, where $\Phi^{(k)}$ is the updates Φ at the k iteration as follows,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n}\frac{\Phi_{ij}^{(k+1)}}{n}.$$

The updating regression parameter $\beta_i^{(k+1)}$ can be computed using weighted least squares method, when the weights rely on $\zeta_{ij}^{(k+1)}$,

$$\beta_i^{(k+1)} = \left(\sum_{j=1}^{n} x_j x_i' \Phi_{ij}^{(k+1)}\zeta_{ij}^{(k+1)}\right)^{-1}\left(\sum_{j=1}^{n}\Phi_{ij}^{(k+1)}\zeta_{ij}^{(k+1)}x_{\_}y_j\right). \tag{9}$$

It is obvious that $\zeta_{ij}^{(k+1)}$ is calculated from (2) of E-step in which increasing the standardized residuals result in decreasing $\zeta_{ij}^{(k+1)}$. Consequently, $\beta_i^{(k+1)}$ will be robust against outliers in y space and $\sigma^{(k+1)}$ is a robust scale estimate of the mixture regression,

$$\sigma^{(k+1)} = \left\{\frac{\sum_{j=1}^{n}\sum_{i=1}^{g}\Phi_{ij}^{(k+1)}\zeta_{ij}^{(k+1)}\left(y_j - x_j\beta_i^{(k+1)}\right)^2}{n}\right\}^{1/2}. \tag{10}$$

However, the Eq. (9) of mixture regression model estimate is not robust to leverage points. Peel and McLachlan [6] mentioned that when v_i is unknown there is no explicit solution for v_i in M-step. Yao [8] introduced profile likelihood to overcome such a problem,

$$L(v) = \max_{\Phi}\sum_{j=1}^{n}\log\left\{\sum_{i=1}^{g}\pi_i f_i\left(y_j; x_j'\beta_i, \sigma_i, v\right)\right\}. \tag{11}$$

When $v_i = v_1 = v_2 = \ldots = v_g$, the $\widehat{v}$ can be computed using EM algorithm, $\widehat{v} = \arg\max_v L(v)$, and and it was found that $L(v)$ can be computed from a set of grid points and noted that when the maximum value of v is between 15–20, the t-distribution approximates the normal distribution. Yao et al. [8] pointed out that when the v is large enough, the t-distribution becomes close to the normal distribution, therefore the adaptively choosing v makes the EM algorithm consider the t-distribution

as a scale mixture of normal distribution. However, $L(v)$ estimate is not resistant to the presence of leverage points. Yao et al. [8] tackle this problem by identifying the leverage points based on Minimum Covariance Determinant (MCD) estimators introduced by Rousseeuw [9], and then trimmed them before fitting the mixture model.

4. Modified robust mixture regression in term of t-distribution

It is obvious that accurate identification of leverage points is a crucial procedure due to the mixture regression based on t-distribution tackles only the problem of outliers. It is well-known that Hat matrix, $H = X(X'X)^{-1}X'$ can be used to identify leverage points, when jth diagonal element of H, $h_{jj} > \frac{2p}{n}$ the jth predictor x_j is a high leverage point. Sometimes, the Mahalanobis Distance $MD = (x_j - \overline{x})\mathrm{Cov}(X)^{-1}(x_j - \overline{x})'$ has been used to detect the high leverage points too. The MD can be written in terms of diagonal elements of H as follows,

$$MD_j = \frac{h_{jj} - n^{-1}}{(n-1)^{-1}}.$$

When $MD_j > \chi^2_{(p-1,0.975)}$ the jth MD is considered a high leverage point. Rousseeuw and van Zomeren [10] declared that due to $\overline{x}$ and $\mathrm{Cov}(X)$ are sensitive to the presence of outliers might create the masking effect. The masking effect is defined as some leverage points probably not being detected due to the effect of other high leverage points. The natural modification is to use a robust MD by replacing the mean and covariance with robust estimates of location and scale for X, excluding the constant term. Imon [11] considered the previous methods as individual diagnostic measures that are affected by the masking and swamping phenomena and then proposed the multiple diagnostics measure which is so called Generalized Potentials (GP). In spite of GP is better than the individual diagnostic measure, Midi et al. (2009) found that GP reduced the number of swamping cases but could not eliminate them completely, therefore, they introduced the Diagnostic Robust Generalized Potential (DRGP) based on MVE which is the sibling method of MCD to improve the performance of GP. Uraibi and Haraj [12] noted that constructing the comprehensive diagnostics depends primarily on identifying an efficient matrix of location and dispersion that has to employ to robust the Mahalanobis distance through fast concentration algorithms. One of these matrices that has proven its efficiency in previous studies, such as (Uraibi et al. 2015; Talib et al. 2022), is the Reweighted Multivariate Normal location and dispersion (RMVN) matrix, therefore, they proposed DRGP (RMVN) that showed that it is more efficient and performs better than others. As mentioned above, most robust mixture regression methods remove any high leverage points detected by robust identification methods. This paper suggests using DRGP (RMVN) to detect high leverage points and then proposes two procedures, remove them with the EM method in terms of the t-distribution or assign robust weights to reduce the effect of leverage points, as follows,

$$W_i = \frac{\chi^2_{(p,0.05)}}{MD_j}.$$

The removed method is called REM and the weighted one is called (WEM).

5. Simulation

Consider the mixture linear regression model as follows:

$$Y = \begin{cases} \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \varepsilon_1, & \text{if} \quad Z = 1, \\ \theta_0 - \theta_1 X_1 - \theta_2 X_2 + \varepsilon_2, & \text{if} \quad Z = 2, \end{cases}$$

where Z is the latent variable and when $P(Z = 1)$ is $\pi_1 = 0.25$ of the components of Y and $P(Z = 2)$ is constructed $\pi_2 = 1 - \pi_1$. X_1 and X_2 are sampled independently and identically from the $N(0, 1)$ distribution with $n = \{45, 75, 100, 150, 300, 500\}$. Let the initial values θ_i be

$$\theta_i = \begin{Bmatrix} \beta_1 = (0, \quad 1, \quad 2) & \sigma = 3 & \pi_1 = 0.25 \\ \beta_2 = (0, -1, -2) & \sigma = 3 & \pi_2 = 0.75 \end{Bmatrix}$$

and the error terms ε_1 and ε_2 have the same distribution (standard normal). Finding Y as a mixture distribution, requires generating n data points from uniform distribution $u \sim \text{UNIF}(0,1)$ such that

$$\varepsilon = \begin{Bmatrix} u \times N(0,1) & \text{if} & u \leqslant 0.75 \\ u \times t(v) & \text{if} & u > 0.75 \end{Bmatrix},$$

v is degree of freedom, when $t(v=1)$ corresponds to the Cauchy distribution, and $t(v=2)$ is t-distribution with two degrees of freedom, and then finding

$$Y = \begin{Bmatrix} 0 + X_1 + 2X_2 + \varepsilon & \text{if} & u \leqslant \pi \\ 0 - X_1 - 2X_2 + \varepsilon & \text{if} & u > \pi \end{Bmatrix},$$

where $\pi = 0.25$ of the components of Y. Two cases have been considered for contaminated data points as follows,

I. $\varepsilon \sim t_{(1)}$ with 10% of leverage points being $X_1 = X_1 \times 10,\ X_2 = X_2 \times 10$ and $Y = Y \times 200$;

II. $\varepsilon \sim t_{(2)}$ with 10% of leverage points being $X_1 = X_1 \times 10,\ X_2 = X_2 \times 10$ and $Y = Y \times 200$.

These steps are repeated 5000 times and the mean squared errors of the estimated components and bias are computed as follows,

$$\text{Mse}(\widehat{\theta}_i) = \frac{\sum_{j=1}^{5000} \left(\widehat{\theta}_i - \theta_i\right)^2}{5000}, \quad i = 1, 2,$$

$$\text{Mse}(\widehat{\pi}_i) = \frac{\sum_{j=1}^{5000} \left(\widehat{\pi}_i - \widehat{\pi}_{0.25,i}\right)^2}{5000}, \quad i = 1, 2.$$

The EM, REM and WEM are compared and the best method is the one that has the lowest values of the above criteria. The results which are presented in Table 1, show when the degree of freedom is one and the dataset is having 10% leverage points, the $\text{Mse}(\widehat{\theta}_i)$ WEM has the lowest values than EM and REM methods, when $n = \{45, 75, 100, 150, 300\}$ except $n = 500$ is displayed that REM outperforms EM and WEM. On the other hand, $\text{Mse}(\widehat{\pi}_{0.25,i})$ values of REM and WEM are equivalent with different sample sizes and are better than the $\text{Mse}(\widehat{\pi}_{0.25,i})$ value from EM.

Table 1. The $\text{Mse}(\widehat{\theta}_i)$, $\text{Mse}(\widehat{\pi}_{0.25})$ of simulation study when $(v = 1)$.

n	Method	$\text{Mse}(\widehat{\theta_0})$	$\text{Mse}(\widehat{\theta_1})$	$\text{Mse}(\widehat{\theta_2})$	$\text{Mse}(\widehat{\theta}_{0.25,i})$
45	EM	1.625	1.049	2.692	0.024
	REM	0.587	1.273	1.686	0.015
	WEM	0.509	1.021	1.492	0.015
75	EM	1.014	0.843	2.757	0.025
	REM	0.361	0.503	0.590	0.005
	WEM	0.275	0.500	0.275	0.005
100	EM	0.802	0.742	2.641	0.025
	REM	0.258	0.371	0.431	0.004
	WEM	0.211	0.271	0.360	0.004
150	EM	0.566	0.856	3.20	0.028
	REM	0.145	0.204	0.346	0.004
	WEM	0.108	0.168	0.239	0.004
300	EM	0.287	0.931	3.569	0.04
	REM	0.047	0.057	0.056	0.002
	WEM	0.053	0.059	0.055	0.002
500	EM	0.083	0.972	3.845	0.044
	REM	0.029	0.038	0.029	0.002
	WEM	0.034	0.045	0.039	0.002

Table 2 shows the results of the simulation of case II when the distribution of random errors is t-distribution with $(v = 2)$ and the 10% of leverage points are present in the dataset. It is obvious that when $n = \{45, 75, 100, 150\}$ the values $\text{Mse}\left(\widehat{\theta_0}\right)$, $\text{Mse}\left(\widehat{\theta_1}\right)$ and $\text{Mse}\left(\widehat{\theta_2}\right)$ of WEM is much lower than their counterparts in EM and REM methods. The performances of REM and WEM become very

close when $n = \{300, 500\}$ due to the values of Mse $\left(\widehat{\theta}_0\right)$, Mse $\left(\widehat{\theta}_1\right)$ and Mse $\left(\widehat{\theta}_2\right)$ for both methods are approximately equal, and their Mse$(\widehat{\pi}_{0.25,i})$ values are equal. It is notable that the Mse $\left(\widehat{\pi}_{0.25,i}\right)$ values of WEM are lower than those of the others when the sample size is smaller than 300.

Table 2. The Mse $\left(\widehat{\theta}_i\right)$, Mse $\left(\widehat{\pi}_{0.25}\right)$ of simulation study when $(v = 2)$.

n	Method	Mse$(\widehat{\theta}_0)$	Mse$(\widehat{\theta}_1)$	Mse$(\widehat{\theta}_2)$	Mse$(\widehat{\pi}_{0.25,i})$
45	EM	1.154	1.121	3.33	0.033
	REM	0.801	1.388	2.312	0.015
	WEM	0.591	1.113	1.599	0.014
75	EM	0.526	1.006	3.916	0.036
	REM	0.312	0.727	1.274	0.007
	WEM	0.258	0.556	1.051	0.006
100	EM	0.638	0.953	3.378	0.036
	REM	0.168	0.286	0.406	0.004
	WEM	0.123	0.253	0.331	0.004
150	EM	0.446	0.999	3.873	0.038
	REM	0.122	0.133	0.180	0.004
	WEM	0.095	0.112	0.114	0.004
300	EM	0.295	1.034	3.994	0.041
	REM	0.047	0.049	0.05	0.002
	WEM	0.043	0.045	0.05	0.002
500	EM	0.167	1.071	4.24	0.04
	REM	0.021	0.024	0.025	0.002
	WEM	0.021	0.024	0.026	0.002

6. Market value data

To better explain the robustness performance of the WEM method compared with REM and EM methods, the market value data of Iraqi trade banks, as previously presented by Uraibi and Haraj [12] was chosen. The Trading Rate and earning per share (EPS) variables were chosen in this paper out of nine financial ratios in the original dataset that affect the market value for the period (2011–2015). First, the regression model is fitted using EM and the residuals have been plotted in Figure 1. It obvious that there is more than one pattern of residuals and each pattern may represent a random distribution. Consequently, this situation leads to occur the heterogeneous problem and making the random distribution of residuals is mixture. Figure 2 shows the normalized EPS and Trading Rate ratios, which have some outliers (leverage points) due to some points lying far from the bulk of the data and also appear to suffer from heterogeneity.

The performance of the WEM method is the best among the EM and REM methods with the market value of Iraq stock market as shown in Table 3. The mean squared error (MSE) and mean absolute error (MAE) of WEM are 40.02 and 1.72, respectively, both lower than those of the other methods.

Table 3. The MSE and MAE of market value data.

Method	EM	REM	WEM
MSE	0.61	41.00	40.02
MAE	0.96	1.76	1.72

7. Conclusion

The main objective of this paper is to improve the performance of the EM method in the presence of leverage points in mixture regression data when the distribution of random errors is t-distribution. From the results presented in Tables 1–3, it can be concluded that trimming leverage points in small sample sizes is not an effective procedure as it reduces the degrees of freedom and consequently increases the mean squared error. Instead of removing some rows that have leverage points by giving zero weight, the WEM method assigns low weights to those rows, and in this case, the method preserves the degrees

of freedom. This is the main reason why the WEM method outperforms the others; moreover, it is resistant to the presence of outliers and leverage points together.

[1] Quandt R. E. A new approach to estimating switching regressions. Journal of the American Statistical Association. **67** (338), 306–310 (1972).

[2] Quandt R. E., Ramsey J. B. Estimating mixtures of normal distribution and switching regressions. Journal of the American Statistical Association. **73** (364), 730–738 (1978).

[3] Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodologica). **39** (1), 1–22 (1977).

[4] Markatou M. Mixture models, robustness, and the weighted likelihood methodology. Biometrics. **56** (2), 483–486 (2000).

[5] Shen R., Ghosh D., Chinnaiyan A. M. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics. **5**, 94 (2004).

[6] Peel D., McLachlan G. J. Robust mixture modelling using the t distribution. Statistics and Computing. **10**, 339–348 (2000).

[7] Song W., Yao W., Xing Y. Robust mixture regression model fitting by Laplace distribution. Computational Statistics & Data Analysis. **71**, 128–137 (2004).

[8] Yao W., Wei Y., Yu C. Robust mixture regression using the t-distribution. Computational Statistics & Data Analysis. **71**, (2014).

[9] Markatou M. Mixture models, robustness, and the weighted likelihood methodology. Biometrics. **56** (2), 483–486 (2000).

[10] Rousseeuw P. J., Van Zomeren B. C. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association. **85** (411), 633–639 (1990).

[11] Imon A. H. M. R. Sub-sample Methods in Regression Residual Prediction and Diagnostics. University of Birmingham (1996).

[12] Uraibi H. S., Haraj S. A. A. Group diagnostic measures of different types of outliers in multiple linear regression model. Malaysian Journal of Science. **41** (sp1), 23–33 (2022).

Модель регресії зваженої суміші на основі t-розподілу за наявності точки важеля

Хассан С. У.

Кафедра статистики, Університет Аль-Кадісія, вул. Альджамеа, Дівання, Ірак

Алгоритм очікування–максимізації (EM) є ефективним методом оцінки параметрів моделі регресії суміші за наявності викидів у Y-напрямку. На жаль, цей метод не працює, якщо точки важеля присутні в наборі даних. Найчастіше в статтях використовували процедуру видалення важелів після їх ідентифікації за допомогою певних методів виявлення. Деякі автори вказували на те, що окремі методи виявлення можуть бути неточними, тому запропонували кілька методів діагностики. Метод зваженої EM (WEM) був запропонований у цьому рукописі для подолання проблеми точок важеля без видалення. Крім того, він заснований на DRGP (RMVN), який є одним із багатьох методів діагностики. Для визначення ефективності запропонованого методу в порівнянні з попередніми методами були розглянуті реальні дані та симуляції. Результат показує, що метод оцінки методу WEM є надійнішим і надійнішим, ніж інші, особливо там, де розміри вибірки невеликі.

Ключові слова: *надійний; регресія суміші; точка важеля; DRGP.*