A New Method for Detection and Estimation of Outliers in Multiple Linear Regression Model

[1]Dr. Nabeel George Nacy, [2]Dr.Ghazi. I .Raho, [3]Zrean Salam Ahmed
[1, 3] *Salahaddin University,* [2] *Amman Arab University*

Abstract:

This paper aims to suggest a new method to detect outliers in multiple linear regression model and suggest three new methods to estimating this outliers. The suggested new method to detect outliers depending on the statistic DFSTAT proposed by Beasley et al. (1980) and modified the ellipse equation which proposed by Nacy (2001) and it also suggest three new methods to estimating these outliers according to the methods proposed by Nacy (2001) after modifying.

Keywords: Belsely Nacy , multiple linear regression models , detecting outliers , estimating outliers .

Introduction: Regression analysis is a statistical technique, which helps us to investigate and to fit an unknown model for quantifying relations among observed variables.

Let us have a set of n observations (y_i , x_{1i} , x_{2i} , … , x_{pi}) for $i=1,2,\dots,n$ of $(p+1)$ dimensional random vector (y , x_1 , x_2 , … , x_p). To serve the purpose in regression analysis , the classical model assumes a relation of the scalar type $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$ where y_i is the dependent (response) variable , x_{1i} , x_{2i} , … , x_{pi} are p independent (predictor) variables ,β_0 is the intercept on the Y-axis ,$\beta_1, \beta_2, \dots, \beta_p$ are the regression coefficients for each of the independent (predictor) variables and ε_i is the error or residual .

We can rewrite the system of equations by matrix notation as : $Y = X\beta + \varepsilon$, where Y is the $(nX1)$ response vector and X is $nX(p+1)$ design matrix , β is the $(p+1)X1$ parameters vector (regression coefficients), and ε is the $(nx1)$ vector . To fit a regression model , we estimate the parameters using least squares errors method , which minimizes the sum of squared deviations of the observed and fitted response , which is commonly referred to as sum square of residuals:

$$\sum_{i=1}^{n} e_i^2 = e^T e = (Y - X\beta)^T (Y - X\beta) \qquad \dots\dots(1)$$

Minimization of (1) results into the least squares estimate of β which is :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$ [1]

An *outlier* is defined as an observation that does not conform to the pattern (model) suggested by the homogeneous majority of the observations in a data set [3]. That does not conform to the linear regression line well. These observations have unusually high residual errors. [5]

Influential observations are those observations that, individually or collectively, excessively influence the fitted regression equation as compared to other observations in the data set. [4]

Masking Effect is a phenomenon that the effect of one observation on the fitting of the model is masking with the effects of another outlier (outliers). That means the effect of all observations appear together; while the effect of each observation does not appear individually. To get rid of this phenomenon most methods deliberately detect many outliers together instead of detecting one outlier. [6]

Suggested method to detect outliers in multiple linear regression: According to the statistics *DFSTAT* (W_{ij}) which suggested by *Belsely , Kuh&Welsch (1980)*[2], as :

$$W_{ij} = \frac{\hat{\beta}_j}{S\sqrt{(X^T X)^{-1}_{jj}}} - \frac{\hat{\beta}_{j(i)}}{S_{(i)}\sqrt{\left(X^T_{(i)} X_{(i)}\right)^{-1}_{jj}}} \qquad \dots\dots\dots(2)$$

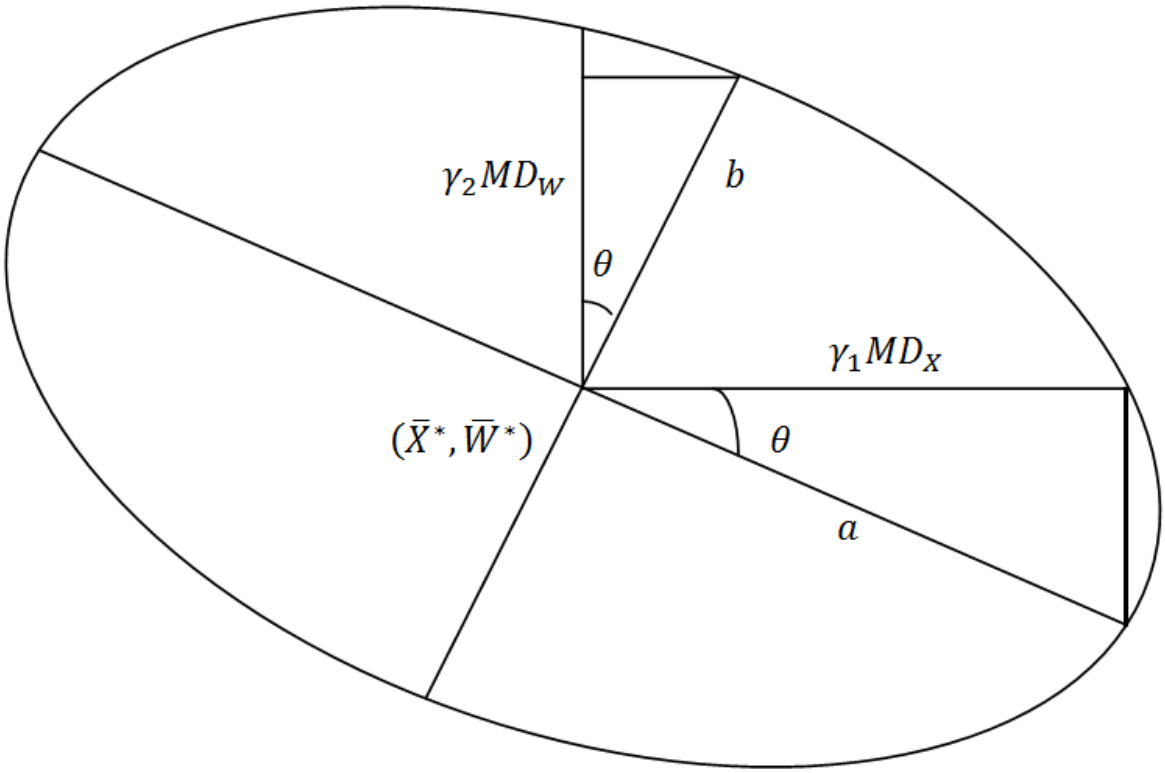
Where $\hat{\beta}_{j(i)}$ is the estimate of β_j when i^{th} row of X and Y have been deleted.

$S_{(i)}$ is the estimate of standard error when i^{th} row of X and Y have been deleted.

$X_{(i)}$ is X matrix with i^{th} observation (row) deleted.

By using ellipse equation which suggested by *Nacy (2001)*[7[as:

$$Z_{ij} = \frac{\left((X_{ij} - \bar{X}_j^*)\cos\theta_j + (W_{ij} - \bar{W}_j^*)\sin\theta_j\right)^2}{{a_j}^2} + \frac{\left((W_{ij} - \bar{W}_j^*)\cos\theta_j - (X_{ij} - \bar{X}_j^*)\sin\theta_j\right)^2}{{b_j}^2} \quad \ldots.(3)$$



Where $\bar{X}_j^*$ and $\bar{W}_j^*$ are the median for X_{ij} and W_{ij} respectively.

$$a_j = \frac{\gamma_{1j} MD_{X_j}}{\cos\theta_j} \qquad and \qquad b_j = \frac{\gamma_{2j} MD_{W_j}}{\cos\theta_j} \quad ; \quad \theta_j = \tan^{-1}\hat{\beta}_j$$

In this proposed method, the median is used instead of mean (in *Nacy's method*) because it is not influence with outliers and also the mean deviation is used instead of standard deviation.

Where: $2a_j$ is the major axis of the ellipse for the j^{th} variable.

$2b_j$ is the minor axis of the ellipse for the j^{th} variable.

$\theta_j$is the angle between major axis and X axis for the jth variable.

MD_{X_j}andMD_{W_j} are the mean deviation of X_{ij} and W_{ij} respectively.

$$MD_{T_j} = \frac{\sum_{i=1}^{n}|T_{ij} - \bar{T}_j|}{n}$$

γ_{1j}andγ_{2j} are two positive real numbers which determine the length of major and minor axis which can be determined approximate:

$$\gamma_{1j} = \frac{\cos\theta_j}{MD_{Xj}\left({\hat{\beta}_j}^2 + 1\right)} * \sqrt{\left(\hat{\beta}_j(W_{1j} - \bar{W}_j^*) + (X_{1j} - \bar{X}_j^*)\right)^2 + \left({\hat{\beta}_j}^2(W_{1j} - \bar{W}_j^*) + \hat{\beta}_j(X_{1j} - \bar{X}_j^*)\right)^2} \quad \ldots(4)$$

$$\gamma_{2j} = \frac{\cos\theta_j}{MD_{Wj}\left({\hat{\beta}_j}^2 + 1\right)} * \sqrt{\left(\hat{\beta}_j(W_{2j} - \bar{W}_j^*) - {\hat{\beta}_j}^2(X_{2j} - \bar{X}_j^*)\right)^2 + \left((W_{2j} - \bar{W}_j^*) - \hat{\beta}_j(X_{2j} - \bar{X}_j^*)\right)^2} \quad \ldots(5)$$

Any values for γ_{1j} and γ_{2j}can be to taken rounded the calculated values from the equations above. [7]

If $Z_{ij} = 1$ then the point (X_{ij}, W_{ij}) locate on the ellipse perimeter.

$Z_{ij} < 1$then the point (X_{ij}, W_{ij}) locate inside the ellipse.

$Z_{ij} > 1$then the point (X_{ij}, W_{ij}) locate outside the ellipse, so this point can be recognized as an outlier.

The procedure of outliers estimation:

After detecting outliers and testing for the appropriate solution procedures on this observations to get rid of its abnormality, because the remaining of these observations in the statistical analysis process, leads to deformation of prediction and estimation process by being remote from reality due to its certain effect on abnormality of this observations. The estimated solution is deleted and distanced like those observations from the set of data and this also affects the accuracy of analysis and statistical estimation, especially when the number of the observations is small, or when the number of outliers is large and it becomes approximately half all of the observation. So the researchers has thought of another solution which depends on estimation of these observations in a way that has ended up with its abnormality after the next test and repeating it in the same group of data. Despite the few numbers of resources to conduct such a research on this solution because it is new and not wide spread.

We suggested three methods to estimate the outliers in multiple linear regression according to *Nacy's* methods [7]:

The first method (ZMR-1):

The estimated value of the outliers (X_{oj}, W_{oj}) is the intersection point between the ellipse equation (3) and the straight line from the outlier point to the center of ellipse:

$$\frac{\left((X_{ij}-\bar{X}_j^*)\cos\theta_j+(W_{ij}-\bar{W}_j^*)\sin\theta_j\right)^2}{{a_j}^2}+\frac{\left((W_{ij}-\bar{W}_j^*)\cos\theta_j-(X_{ij}-\bar{X}_j^*)\sin\theta_j\right)^2}{{b_j}^2}=1 \quad \ldots..(6)$$

$$\frac{W_{ij}-\overline{W}_j^{\,*}}{X_{ij}-\overline{X}_j^{\,*}}=\frac{W_{oj}-\overline{W}_j^{\,*}}{X_{oj}-\overline{X}_j^{\,*}} \ldots\ldots(7)$$

After solving the two above equations simultaneously , we obtained :

$$\hat{X}_{ij}=\overline{X}_j^{\,*}+\frac{(X_{oj}-\overline{X}_j^{\,*})}{\sqrt{\left[\frac{(X_{oj}-\overline{X}_j^{\,*})\cos\theta_j+(W_{oj}-\overline{W}_j^{\,*})\sin\theta_j}{a_j}\right]^2+\left[\frac{(W_{oj}-\overline{W}_j^{\,*})\cos\theta_j-(X_{oj}-\overline{X}_j^{\,*})\sin\theta_j}{b_j}\right]^2}} \ldots.(8)$$

$$\hat{W}_{ij}=\overline{W}_j^{\,*}+\frac{(W_{oj}-\overline{W}_j^{\,*})}{\sqrt{\left[\frac{(X_{oj}-\overline{X}_j^{\,*})\cos\theta_j+(W_{oj}-\overline{W}_j^{\,*})\sin\theta_j}{a_j}\right]^2+\left[\frac{(W_{oj}-\overline{W}_j^{\,*})\cos\theta_j-(X_{oj}-\overline{X}_j^{\,*})\sin\theta_j}{b_j}\right]^2}} \ldots.(9)$$

Hence the point which is closest to the outlier point , suppose as an estimator to the outlier point .

The second method (ZMR-2):

It can estimated the outlier values from the two relationships:

$$\hat{X}_{ij}=X_{ij}\mp\sqrt{Z_{ij}\,.\,MD_{X_j}} \qquad \ldots\ldots\ldots\ldots..(10)$$

$$\hat{W}_{ij}=W_{ij}\mp\sqrt{Z_{ij}\,.\,MD_{W_j}} \qquad \ldots\ldots\ldots\ldots(11)$$

The Positive sign is used in the above equations, if the observation value is smaller in the median ($X_{ij}<\overline{X}_j^{\,*}$ *or* $W_{ij}<\overline{W}_j^{\,*}$ respectively) where as the negative sign is used for the otherwise.

The third method (ZMR-3):

This method is used as the same previous method by using the following two relationships.

$$\hat{X}_{ij}=X_{ij}\mp\sqrt{\gamma_{1j}\,.\,Z_{ij}\,.\,MD_{X_j}} \qquad \ldots\ldots. (12)$$

$$\hat{W}_{ij}=W_{ij}\mp\sqrt{\gamma_{2j}\,.\,Z_{ij}\,.\,MD_{W_j}} \qquad \ldots\ldots\ldots(13)$$

The Positive sign is used in the above equations, if the observation value is smaller in the median ($X_{ij}<\overline{X}_j^{\,*}$ *or* $W_{ij}<\overline{W}_j^{\,*}$ respectively) where as the negative sign is used for the otherwise.

Practical aspect:

The data in table (*1*) contains (*50*) observations (patients) are taken from Hawler education hospital in Erbil. The data consists of the factors affecting on the blood pressure such as ((creatinin, urea, sodium, potassium, tryglicerides, cholesterol, LDL- cholesterol, age)).

Table (1): Data of Hawler education hospital

i	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	150	1.65	70	141	5.3	118.8	134.1	73	67
2	130	1.06	46	144	3.8	44	139	66.9	61
3	140	2.14	53.3	145	3.6	132.3	125.5	50.6	46
4	180	1.22	46.5	147	4.4	122.6	186.5	124.6	62
5	160	1.13	43.2	148	4.1	79.5	138.2	69.8	72
6	120	1.28	69.2	144	4.4	95	155.1	85.4	70
7	140	0.76	50.9	145	4.1	102.8	206.4	142.1	80
8	120	5.61	93	140	4.3	160.9	75	10.4	39
9	125	1.37	46.5	147	4	101.5	95.2	45	85
10	125	2	132	148	3.7	204.7	104.6	29.4	85
11	120	0.65	50.4	152	3.9	148.3	110.1	70.9	70
12	110	0.67	21.8	144	4.5	84.4	145.8	63.6	14
13	120	0.69	29.3	145	4.4	378.3	187.3	115	31
14	150	1.68	55.6	147	4.5	123.3	116.3	49.7	65
15	105	0.71	27.6	141	3.8	263.4	83.6	14.5	44
16	170	1	53.2	147	4	132.3	135.9	55.4	68
17	120	0.7	50	138	4.1	188.5	149.1	68.5	70
18	140	7.82	193.8	139	4.9	46.8	92.8	41.5	57
19	150	0.6	29.5	138	3.7	136.7	216.5	127.1	48
20	130	0.65	35.8	136	4.5	69.1	94.2	43.3	80
21	135	1	24.6	148	4.3	161.3	151.8	70.4	45
22	160	0.56	28	146	4.5	391.6	235.1	111.4	67
23	125	0.98	38.8	145	4.2	80	113.6	60.7	65
24	130	0.59	12.9	141	3.2	64.1	164.5	96.6	80
25	140	1.29	96.2	138	3.8	162.6	175	89	85
26	140	5.35	229.4	164	5.8	123	94	130	85
27	110	0.93	28.5	145	4.2	109	111	85	38
28	120	0.9	39.2	146	3.9	113	179	85	35
29	140	1.21	38.8	145	3.2	208	151	125	49
30	120	0.7	50	146	4.6	188.5	149.1	68.5	70
31	160	1.39	45.1	150	4.3	167.9	160.6	87.6	73
32	115	2.21	123.1	144	4	92	116	85.7	75
33	160	0.84	24.5	137	3.5	126	234	116	87
34	120	1.1	38.3	146	4.1	51	139	78.4	70
35	140	1.21	51	144	4.7	96	131	67.8	55
36	120	0.93	37.7	141	4	84	182	48	71
37	120	0.7	50	138	4.1	188.5	149.1	68.5	70
38	135	1.12	42.6	135	3.5	77.1	140.1	64.5	61
39	120	2.6	187.1	139	2.7	103.1	137	71.6	79
40	120	1.74	111	137	5.6	52.2	106	42.6	90
41	90	1.35	111.4	132	4.7	279	77	40	71
42	110	1.58	182.7	146	4	266	126	54	55
43	130	1.18	85	146	3.9	155	168.3	135.9	60
44	100	1.12	49.5	116	3.7	85.7	148.9	83.9	73
45	115	2.21	123.1	144	4	92	116	29.4	75
46	150	0.99	74.2	159	4	117.1	105	51.3	72
47	130	1.12	24.6	148	5	161.3	151.8	70.4	47
48	140	1.53	74.8	144	3.8	90	118	95.5	82
49	140	1.1	38.3	146	4.1	51	139	78	70
50	130	1.78	184.2	148	4.8	118	198	78	85

The measured variables are:
Y: the blood pressure units (*mm.Hg*).
X_1: creatinin units (*mg/100 ml*).
X_2: urea units (*mg/100 ml*).
X_3: Sodium units (*Meq/L*).
X_4: potassium units (*Meq/L*).
X_5: triglycerides units (*mg/100 ml*).
X_6: cholesterol units (*mg/100 ml*).
X_7: LDL- cholesterol units (*mg/100 ml*).
X_8: age.
The multiple linear regression model is:

The testing of normality problem :
Using *Shaprio-Wilk*test, the appropriate hypothesis is given by:
H_0: "The distribution of the random error is very close to normal distribution" H_1: "The distribution of the random error is different to normal distribution"

Table (2): The testing of normality problem

Shapiro -Wilk		
Statistic	*df*	*Sig.*
0.967	*50*	*0.171*

The P-value is *0.171* for random error using *Shapiro-Wilk* test conform the approximate normality of standardized residuals of model (*14*)
The testing of Homogeneous problem (for random error):
The appropriate hypothesis is given by:
H_0: "Homogeneous of variance random error"
H_1: "Not Homogeneous of variance random error"

Table (*3*): The testing of Homogeneous problem

Levene Statistic	*df1*	*df1*	*Sig.*
3.072	*1*	*38*	*0.09*

Since the value of (*p-value = 0.09*) is greater than the value of the level of significant (*0.05*), this means cannot reject the null hypothesis of any variation random error homogeneous and there is no problem of heterogeneity of variance random error for the model (*14*).

The testing of multicolinearity problem:-
For testing the multicolinearity problem for each predictor variable on all other predictors for the model (*14*) by using the variance inflation factor (*VIF*).

Table (4): The variance inflation factor for predictor variable

Prediction	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
VIF	*2.619*	*2.711*	*1.175*	*1.181*	*1.232*	*2.452*	*2.122*	*1.455*

From table (*4*) it is show that all the values of Variance Inflation Factor (*VIF*) are less than (*5*). Thus then it is concluded that there is no multicolinearity problem between the predictor variables.

Test of Auto correlation problem:
For the testing of Auto correlation problem between the error of the model (*11*) used the *Durbin – Watson* (*DW*) Test. The appropriate hypothesis is given by:
$H_0: \rho = 0$ "They have no Auto correlation between the residuals"
$H_1: \rho \neq 0$ "They have Auto correlation between the residuals"
The nullhypothesis is accepted if ($2 \leq DW < 4 - du$ *or* $du < DW \leq 2$).
The result of (*Durbin – Watson*) for the model (14) equals to (*1.986*), and The result of the tabulated values of upper and lower at degrees of freedom (*50,9*) and level of significant $\alpha = 0.05$ are $du = 1.930$, $dl = 1.201$,

means that $(1.930 < 1.986 < 2.07)$. We accept the null hypothesis (H_0) i.e. there is no Auto correlation between the residuals for the model (14).

Detection and estimation of outliers:

Applying equation (*2*) to find values of (W_{ij}) and using the equation of ellipse (*3*) . According to the original values (X_{ij}) and values of W_{ij} extracted, there will be eight ellipse equations, for each predictor variable to detect outlier observation according to the suggested method.

The estimate of those outliers using the suggested methods and a range of programs on *MATLAB* language for calculations, the final results of estimation are shown in table (*5*).

From table (*5*), It is noted that the suggested method of estimating has changed value of outlier with the ratio between *((-47.5)-82.9) %* related to the original value.

The suggested method of estimating change values of two observations is only at *ZMR1* method and four observations only at *ZMR3* methods and twenty-one observations at *ZMR2* method from *400* observations of the model.

From table (*6*) we note that the suggested method *ZMR2* only increased the adjusted *R square* and decreased *Mean Squares Error (MSE)* and gives relative efficiency *109%*.

Table (5):The estimate of all outlier observations for suggested methods

Observation	*Predictor Variables*	*Original Value*	*The value of estimation by methods*		
			ZMR1	*ZMR2*	*ZMR3*
1	X_3	*141*		*143.132*	
	X_4	*5.3*		*4.251*	
4	X_7	*124.6*		*118.990*	
8	X_1	*5.61*		*3.491*	
12	X_8	*14*		*18.527*	
13	X_5	*378.3*		*369.456*	
16	X_1	*1*		*1.829*	
	X_7	*55.4*		*61.490*	
18	X_1	*7.82*	*5.498*	*4.102*	*4.469*
22	X_5	*391.6*		*382.905*	
24	X_4	*3.2*		*3.911*	
26	X_1	*5.35*		*4.049*	*2.986*
	X_4	*5.8*		*4.970*	
	X_6	*94*		*101.133*	
	X_7	*130*		*123.454*	
36	X_6	*182*		*174.613*	
	X_7	*48*		*55.486*	

41	X_3	132		134.581	
44	X_3	116	120.884	122.477	123.519
46	X_3	159		155.323	152.792
50	X_2	184.2		175.808	

Table (6) the results of the efficiency for the suggested methods of estimating outliers

Methods	(	Iteration	R	MSE	Efficiency
Before Estimation	____	____	0.437	186.195	1
ZMR1	(8.9,8.8)	1	0.400	198.516	0.938
ZMR2	(6,4)	3	0.481	171.475	1.086
ZMR3	(7.7,6.7)	1	0.385	203.477	0.914

Test of significance for all models :

To test the significant of the model, we used the appropriate hypothesis as:
$H_O: \beta_1 = \cdots = \beta_8 = 0$ "Non-Significant"
$H_1: at\ least\ one\ of\ \beta_i's\ does\ not\ equal\ zero.$ "Significant"

Table (7): test of significance for the model.

methods	F-test	P-value
Before estimation	5.752	0.000
ZMR1	5.077	0.000
ZMR2	6.686	0.000
ZMR3	4.828	0.000

From table (7), this noted that p-value is less than 0.001.Thus it is concluded that the model is significant, which means that at least one of $\beta_i's$ does not equal to zero.

Test of significance for parameters of the model:

To test the significance of the parameters for the model, we use the proposed hypothesis for each parameter is used as:

$$H_O: \beta_i = 0\ (non - significant)$$

Table (*8*): the results of testing the values of parameters for the model

Method										
Before Estim-ation		*-94.205*	*8.021*	*-.213*	*1.185*	*-.221*	*-.007*	*.227*	*.015*	*.413*
	t - statistic	*-2.024*	*3.392*	*-3.414*	*3.828*	*-.059*	*-.241*	*2.851*	*.162*	*2.905*
	P - value	*.050*	*.002*	*.001*	*.000*	*.953*	*.811*	*.007*	*.872*	*.006*
ZMR-1		*-92.632*	*8.663*	*-.200*	*1.164*	*.308*	*-.013*	*.239*	*.002*	*.376*
	t - statistic	*-1.835*	*2.954*	*-3.079*	*3.434*	*.080*	*-.432*	*2.861*	*.026*	*2.600*
	P - value	*.074*	*.005*	*.004*	*.001*	*.936*	*.668*	*.007*	*.979*	*.013*
ZMR-2		*-97.721*	*18.240*	*-.287*	*1.154*	*-.990*	*.000*	*.261*	*-.006*	*.393*
	t - statistic	*-1.963*	*4.164*	*-4.146*	*3.411*	*-.248*	*.017*	*3.228*	*-.061*	*2.905*
	P - value	*.056*	*.000*	*.000*	*.001*	*.806*	*.986*	*.002*	*.951*	*.006*
ZMR-3		*-115.98*	*9.755*	*-.188*	*1.273*	*1.749*	*-.014*	*.211*	*.052*	*.385*
	t - statistic	*-2.081*	*2.928*	*-3.001*	*3.395*	*.456*	*-.459*	*2.540*	*.534*	*2.616*
	P - value	*.044*	*.006*	*.005*	*.002*	*.651*	*.649*	*.015*	*.596*	*.012*

From table (*8*), it is noted that the *p-value* is less than *0.01* for the values of $\beta_1, \beta_2, \beta_3, \beta_6$ *and* β_8 that means it is significant, but not significant to the value of β_4, β_5 *and* β_7.The situation remains as it's after estimating outliers , except β_5 that has changed it's value from *(-0.007*) to (*0.000*), by the suggested method *ZMR2* which correspond to variables X_5 (triglycerides units), that is consistent with the medical concept more than it was before the estimation.

Conclusions: From the practical aspect , we conclude :

1. The suggested method to detecting outliers depends on supposing of the values of γ_1 and γ_2, these two values are determined by the tolerance range for the spread of points inside the ellipse. This method gives good results in the suggested methods for the detecting and estimating outliers at maximize γ_1 and minimize γ_2 get a bigger R^2_{ad} and smaller *MSE*.
2. The suggested method *ZMR-2* revalued R^2_{ad} and devalued *MSE* with relative efficiency *109%* after it detected *21* observations of total *400*.
3. $\beta_5$have changed it's value from *(-0.007)* to (*0,000*), by the suggested method *ZMR2* which corresponds to variable X_5 (triglycerides units), that is consistent with the medical concept more than before the estimation.
4. From the practical aspect it is obtained the summary results it is noticed that the best method to detect and to estimate is method *ZMR-2*; it has achieved relative efficiency *109%* .
5. The mathematical model remained significant by *F-test* after detecting and estimating outliers in each method.

References

[1] Akter, S. & Khan, M. H. (2010) " Multiple-Case Outlier Detection in Multiple Linear Regression Model Using Quantum - Inspired Evolutionary Algorithm" COMPUTERS, Vol. 5, no. 12, p.p. 1779 – 1788.

[2] Belsley, D. A. ,Kuh, E. &Welsch, R. E. (1980) " Regression Diagnostics Identifying Influential Data and Sources of Collinearity " , John Wiley, New York.

[3] Billor, N. &Kiral, G. (2008) "A Comparison of Multiple Outlier Detection Methods for Regression Data"Communications in Statistics—Simulation and Computation, no. 37, p.p. 521-545.

[4] Chatterjee, s. &Hadi, A. S. (1988), " Sensitivity Analysis in linear Regression " John Wiley, New York, p. p. (129, 95, 96).

[5] Hu, Y. (2011) "linear regression"Journal of Validation technology, Spring 2011, p.p. 15–22.

[6] Rousseuw, P. J. &Leroy, A. M. (1987) " Robust Regression and Outliers Detection" John Wiley, Ney York.

[7] Raho,Ghazi &nor Burhan (2001) Management information system computerization

[8] ناسي، نبيل جورج. (2001)، " تقييم كفاءة طرق تقدير القيم الشاذة لنماذج الانحدار" إطروحة دكتوراه مقدمة الى مجلس كلية الإدارة و الاقتصاد في جامعة بغداد ،العراق .