مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

A comparison between general linear regression and quantitative robust regression (data for kidney patients as an example)

Osama abdulazeez kadhim Al-Quraishi
Ministry of Education Iraq
ousamastat@gmail.com
Orcid:000-0002-1701-2679

Sama saadi Ali Alhashimi
Rusafa Management Institute, Baghdad
samaalhashimi@mtu.edu.iq

Abstract:

Regression analysis is considered one of the statistical methods that are widely used several papers because it describes the relationship between variables in the form of an equation and is defined as a statistical tool used to find out the relationship between the dependent variable and one or more independent variables. The research problem was the presence of outliers in the data of the dependent variable and this leads to the inefficiency of the linear regression model estimated using general linear regression due to the failure to meet the conditions for using Ordinary least squares method (OLS), and thus a lack of confidence in its estimated and predictive accuracy, The aim of the research is to use quantitative robust regression on a sample divided into three quartiles (0.25, 0.50, 0.75) represented by (74) male kidney patients in Al-Kadhimiya Hospital for the period (1/2024-3/2024) and compare it with general linear regression and choose the best model using Comparison standards (Hannan & Quinn, Schwartz Bayesian, Akaike Information), to determine the extent of the effect of independent variables (Diabetes, Hypertension, triglycerides, age) on the increase in blood urea, and the researcher reached the advantage of quantitative robust regression (first quartile and second quartile) compared to general linear regression, While general linear regression appeared better than quantitative robust regression (third quartile), All independent variables have a significant effect on the increase in blood urea levels in all three models.

keywords: General linear regression , Quantile regression, Robust

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

مقارنة بين الانحدار الخطي العام والانحدار الكمي القوي (بيانات مرضى الكلى كمثال)

أسامة عبد العزيز كاظم القريشي
وزارة التربية العراقية
ousamastat@gmail.com
Orcid:000-0002-1701-2679

سما سعدي علي الهاشمي
معهد إدارة الرصافة، بغداد
samaalhashimi@mtu.edu.iq

المستخلص

يعتبر تحليل الانحدار أحد الأساليب الإحصائية التي تستخدم على نطاق واسع في عدة أوراق لأنه يصف العلاقة بين المتغيرات في شكل معادلة ويعرف بأنه أداة إحصائية تستخدم لمعرفة العلاقة بين المتغير التابع وواحد أو أكثر مستقل المتغيرات. تمثلت مشكلة البحث في وجود القيم المتطرفة في بيانات المتغير التابع وهذا يؤدي إلى عدم كفاءة نموذج الانحدار الخطي المقدر باستخدام الانحدار الخطي العام بسبب عدم استيفاء شروط استخدام طريقة المربعات الصغرى العادية (OLS)، و وبالتالي عدم الثقة في دقتها التقديرية والتنبؤية، هدف البحث هو استخدام الانحدار الكمي القوي على عينة مقسمة إلى ثلاثة أرباع (0.25، 0.50، 0.75) ممثلة بـ (74) من مرضى الكلى الذكور في مستشفى الكاظمية. للفترة (2024/1-2024/3) ومقارنتها مع الانحدار الخطي العام واختيار النموذج الأفضل باستخدام معايير المقارنة (Hannan & Quinn, Schwartz Bayesian, Akaike Information)، لتحديد مدى تأثير المتغيرات المستقلة ((السكري، ارتفاع ضغط الدم، الدهون الثلاثية، العمر) على زيادة يوريا الدم، وتوصلت الباحثة إلى ميزة الانحدار الكمي القوي (الربع الأول والربع الثاني) مقارنة بالانحدار الخطي العام، في حين ظهر الانحدار الخطي العام أفضل من الانحدار القوي الكمي (الثالث الربعية)، جميع المتغيرات المستقلة لها تأثير كبير على زيادة مستويات اليوريا في الدم في جميع النماذج الثلاثة.

الكلمات المفتاحية : الانحدار الخطي العام، الانحدار الكمي، قوي

Introduction:

Regression analysis is a statistical tool used to determine the relationship between the dependent variable and one or more Independent variables , Regression analysis is one of the statistical methods that is widely used in most papers because it describes the relationship between variables in the form of an equation, Regression analysis is used for several important purposes, namely finding the regression equation that describes the data available to the researcher, estimating its coefficients to know the

strength and direction of the relationship between variables, and estimating the values of the dependent variable and predicting them in the future, which is useful in planning and decision-making. Regression analysis is divided into two main parts: linear regression and Non- linear regression, In this paper we will discuss linear regression comparing with robust quantile regression, which is considered one of the good methods, The research problem was the presence of outliers in the data of the dependent variable, which leads to the inefficiency of the estimated general linear regression model due to not meeting the conditions for using (OLS) and thus the lack of confidence in its estimation and predictive accuracy. The research aimed to use quantitative robust regression (quartiles) on a sample consisting of (74) only male kidney patients in Al-Kadhimiya Hospital for the period of time (1/2024-3/2024), comparing it with general linear regression and choosing the best model using comparison standards (HOC, BIC, AIC). The research also aimed to determine the extent to which independent variables (Diabetes, Hypertension, triglycerides, age) affect the increase affect the increase in the ratio of urea in the blood, It is necessary to present some literature that dealt with the subject of study. In 2009, Kazim[7] compared the estimates of the OLS method and the linear goals programming method for the multiple linear regression model, In 2023,Shahut and Shatwan[13] used multiple linear regression to predict the sales volume of the building materials investment complex in Misrata, using the OLS method to estimate the model parameters, Also in 2023 Bassiouni[3] presented a comparative study between methods for treating double linearity using internal migration data in Egypt, In 2014 Long Yunk[10] discussed robust regression, which deals with the distortion of the data of the dependent variable due to it containing outliers using the method of least absolute deviation, In 2018, researcher Elisabeth[6] made a comparison between general linear regression (OLS) and quantitative

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

robust regression, explaining the disadvantages and advantages of robust regression,

1- **the theoretical side**:

1-1: General linear regression:[9][11][13]

As it is known, the least squares (OLS) method estimates the conditional mean of the response variable (Y) across the values of the explanatory variables (Xi), so that it is possible for one variable to be predicted by the other. The general linear regression model is written according to the following formula:

$$Y = X\beta + \varepsilon \qquad (1)$$

Where :

Y: $Dependent\ variable\ vector\ of\ \ (nx1)$,X : Matrix of explanatory variables (nxp), β : Vector of parameters to be estimated (Px1), ε : Random error vector$(nx1)$,

let : $\varepsilon \sim N(0, \sigma^2 I_n)$

The least squares estimator is according to the following formula:

$$\hat{\beta}_{LS} = (\acute{X}X)^{-1}\acute{X}Y \qquad (2)$$

The variance and covariance matrix of the estimator ($\hat{\beta}_{LS}$) are given by the following formula:

$$Var - Cov(\hat{\beta}_{LS}) = \sigma^2(\acute{X}X)^{-1} \qquad (3)$$

The vector of parameters of the linear regression model estimated by the least squares (OLS) method has the Best Linear Unbiased Estimator, One of the conditions for applying ordinary least squares is that there be no multicollinearity between the explanatory variables, and the number of observations must be greater than the required number of parameters.

1-2 : robust quantile linear regression :[1][12][14]

Robust quantile regression is used to estimate the parameters of the regression model when there are outlier values in the dependent variable,

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

which are values that differ significantly from the rest of the set of values, robust regression aims to reduce the impact of outliers on the estimation of regression parameters, By minimizing the robust loss function, which is less sensitive to outliers than the squared error function used in ols , Quantitative regression is one of the regression methods used in statistics and econometrics, introduced by Koenker and Basset in 1978, and quantile regression estimates the conditional mean or (other quantities) of the response variable Basically, quantile regression is an extension of linear regression and is used when the conditions for general linear regression are not met, Quantile regression is more desirable if conditional partition functions are important. One of the advantages of quantile regression compared to general linear regression is that quantile regression estimates are more robust against outliers or anomalies in response measurements.

1-3 : The quantile function:[5][8][14]

For a random variable X with probability distribution function

$F(X) = p\ (X \leq x)$

The q th quantile of X^* is defined as the inverse function

$QR_x(q) = F_x^{-1}(q) = \text{Inf}\ \{x : F(x) \geq q\ \}$

where $0 < q < 1$. In particular, the median is QR (1/2)

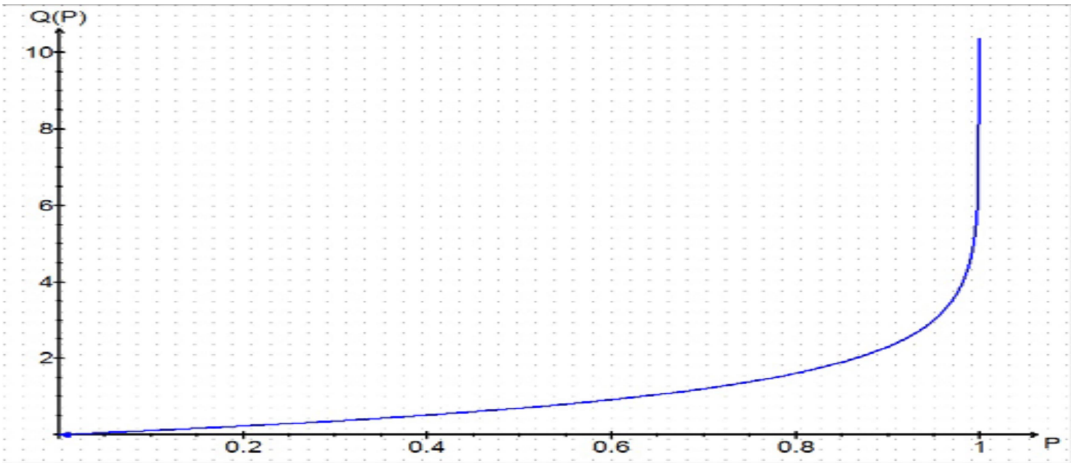

Figure (1) Quantitative function

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

If we assume that we have a random sample (y_1, x_1).... (y_n, x_n) then the quantile regression model takes the following form:

$$QR_q(Y) = \beta_0(q) + \beta_1(q)X_1 + \beta_2(q)X_{21} + \cdots + \beta_K(q)X_{K-1} + \varepsilon \quad \ldots(4)$$

Whereas:

QR: Quantile regression , q: A specific quantity (rabiyat, hundreds, tens...etc.) , Y: Dependent variable , β_K: Model parameters , X_K: Independent variables

ε: Random error vector(nx1):

Quantile regression works to minimize the absolute deviation between random error and a specific quantity (quartiles, decimals,) under study as follows:

Median regression deviation (LAD)=Min$\sum_{i=0}^{n}|e_i|$ ... (5)

following form:The process Least Absolute Deviation in the partition regression model takes the

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\beta}^{N} q|e_i| + \sum_{i:y_i < x_i\beta}^{N} (1-q)|e_i|$$

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\beta}^{N} q|y_i - x_i\beta_q| + \sum_{i:y_i < x_i\beta}^{N} (1-q)|y_i - x_i\beta_q| \quad \ldots(6)$$

In the first part of the above equation, q works to reduce the overestimation, and in the second part of the above equation, (1-q) works to raise the underestimation. QR uses the linear programming method or the interior point method to estimate the model parameters.

1-4: Comparison standards:[2][3][9]

Some statistical standards can be used to compare the estimated regression models as follows:

1- Hannan&Quinn (HOC)

It is calculated according to the following formula

HOC=-2 Log-likelihood + 2k ln[ln(n)]

2- Schwartz Bayesian(Bic)

It is calculated according to the following formula

Bic=-2 Log-likelihood+ k ln(n)

3- Akaike Information(Aic)

It is calculated according to the following formula

Aic=-2 Log-likelihood+2k

Whereas:

K : Model rank , n : Number of sample observation

$$\text{Log} - \text{likelihood} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum e_i{}^2 \qquad \ldots(7)$$

2-The application side:

2-1: Data collection:

Data were collected from a random sample of (74) men with kidney failure (kidney dysfunction) hospitalized in Al-Kadhimiya Teaching Hospital for the period (2023:10-2024:1). The research variables can be explained as follows::

Y: The dependent variable represents the level of urea in the blood

The independent variables are:

X_1:Diabetes , X_2:Hypertension X_3: Triglyceride , X_4: Age of patients

The data was presented in the table below after arranging the values of the adopted variable in ascending order:

Table (1) represents the values of the research variables

i	Y	X_1	X_2	X_3	X_4	i	Y	X_1	X_2	X_3	X_4
1	31	80	11	150	47	38	93	232	14	295	61
2	32	88	12	140	50	39	96	206	16	302	60
3	39	73	14	167	51	40	96	204	16	290	61
4	42	133	14	166	49	41	97	190	18	302	61
5	44	83	13	166	48	42	97	191	16	297	62
6	47	70	14	163	53	4	97	199	15	295	60

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

7	49	125	14	164	53	44	98	218	15	299	63
8	49	115	13	166	54	45	99	199	14	302	61
9	52	111	14	166	52	46	99	193	15	302	63
10	53	113	13	166	56	47	99	203	15	319	64
11	54	112	14	169	55	48	101	213	15	328	64
12	54	114	12	169	55	49	101	226	16	323	63
13	62	111	16	165	54	50	107	231	15	317	64
14	66	113	14	167	61	51	107	230	17	317	62
15	66	113	15	168	60	52	107	219	17	305	65
16	68	114	18	169	57	53	101	222	15	297	65
17	69	122	15	168	56	54	106	206	15	302	64
18	75	129	17	186	56	55	121	226	18	289	65
19	75	128	17	185	57	56	125	225	16	279	59
20	78	122	16	200	63	57	125	198	18	278	66
21	80	133	19	208	58	58	126	200	18	255	67
22	81	137	18	221	57	59	126	206	16	256	66
23	84	146	14	235	56	60	126	202	16	274	67
24	84	171	16	243	58	61	150	191	17	255	67
25	85	190	16	254	57	62	153	212	20	270	65
26	85	212	17	279	58	63	168	207	20	282	67
27	85	232	15	291	57	64	173	201	19	282	66
28	86	228	18	275	59	65	201	184	16	283	67
29	89	252	16	292	57	66	217	216	18	274	68
39	89	252	15	300	60	67	218	240	18	277	68
31	89	269	16	298	59	68	218	242	17	281	68
32	90	277	17	307	58	69	231	240	15	276	69
33	90	265	16	305	60	70	234	248	16	271	69
34	90	249	16	297	60	71	234	278	18	271	71
35	90	256	17	295	59	72	270	258	16	251	69
36	92	236	17	295	60	73	274	261	16	256	70
37	93	228	15	284	59	74	274	279	20	256	72

Using the statistical program (spss 26) and the program (GRTEL), the statistical analysis was carried out according to the following paragraphs

2-2:Test for outliers:

The box-plot chart was used to detect the presence of anomalous values in the values of the dependent variable. By looking at Figure (1), it was revealed that there are values outside the second wall of the box, and therefore these values are considered anomalous from the other values.

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

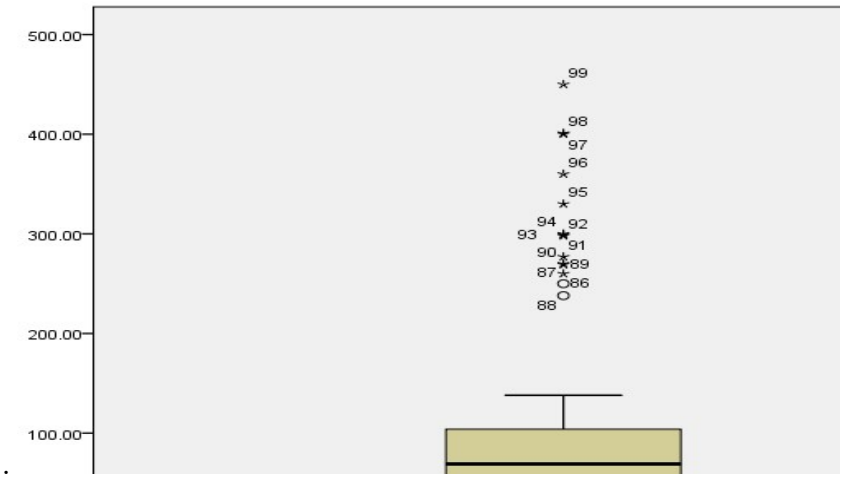

Figure (2) Box-plot for the dependent variable

2-3 : *Estimating a general linear regression model :*

Table (2) OLS estimation results

Parameter	Parameter estimate	Std. Error	t-value	Sig.	F	Sig	Adjusted R^2
const	-366.522	38.329	-9.563	2.85e-014	83.2145	1.18e-025	0.828298
X_1	.603	.115	5.241	1.66e-06			
X_2	.052	1.984	.026	0.9793			
X_3	0.643	.019	-5.930	1.08e-07			
X_4	8.614	.796	10.38	1.58e-016			

Looking at Table (1), we find that (p-value=1.18e-025) for the calculated (F) value is less than the significance level (0.05), which indicates the importance of the estimated general linear regression model. It turns out that the independent variables (x_1,x_3,x_4) (83%) exert a significant influence on the dependent variable because the marginal regression coefficients (b4, b3, b1) have probability values smaller than 05 and the rest of the influences (17%) are from external factors. .

2-4: Estimating a quantile robust regression model

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

The regression equation was estimated using quartiles as follows:

1- When (q=0.25)

Table (3) Estimation results($QR_{0.25}$)

Parameter	Parameter estimate	Std. Error	t-value	Sig.	F	Sig	PSEUDOR^2
const	-238.539	19.1004	−12.49	2.24e-019	122.872	1.28717e-030	0.57
X_1	.136	0.0573756	2.363	0.0209			
X_2	2.155	0.988688	2.179	0.0327			
X_3	-0.082	0.0540443	-1.518	0.1335			
X_4	4.769	0.396421	12.03	1.32e-018			

Looking at Table (3), the estimated first quartile regression model appeared significant because the calculated value (p-value=1.28717e-030) is less than the significance level (0.05), and it turns out that the independent variables (X_1,x_2,x_3) by (57%) have a significant effect. on the dependent variable because the marginal slope parameter (b4, b2, b1) has probability values smaller than 0.05 and the rest of the influences (43%) are from external factors.

2- When(q=0.05)

Table (4) Estimation results ($QR_{0.5}$)

Parameter	Parameter estimate	Std. Error	t-value	Sig.	F	Sig	PSEUDOR^2
const	−369.389	25.1290	-14.70	6.07e-023	162.339	2.53657e-034	0.52
X_1	.419825	0.0754848	5.562	4.71e-07			
X_2	3.71360	1.30074	2.855	0.0057			
X_3	0.45575	0.0711021	-6.410	1.53e-08			
X_4	7.41797	0.521542	14.22	3.38e-022			

Looking at Table (4), we notice that the value (p-value=2.53657e-034) of the calculated (F) value is less than the significance level (0.05), which indicates the importance of the estimated second quartile regression model, and it turns out that (52%) of the independent variables Taken

together, they have a significant effect on the dependent variable due to all regression parameters (b1, b2, b3,b4,) with probability values smaller than (0.05) at a rate of (52%). The remaining influences (48%) are from external factors

3- When(q=0.75)

Table (5) Estimation results($QR_{0.75}$)

Parameter	Parameter estimate	Std. Error	t- value	Sig.	F	Sig	PSEUDOR^2
const	−386.675	63.9008	-6.051	6.63e-08	33.3653	1.7661e-015	0.642
X_1	0.505685	0.191951	2.634	0.0104			
X_2	-3.96481	3.30767	-1.199	. 0.2348			
X_3	-0.741336	0.180806	-4.100	0.0001			
X_4	11.0202	1.32623	8.309	5.43e-012			

Looking at Table (5), we notice that the value (p-value = 1.7661e-015) of the calculated (F) value is less than the significance level (0.05), which indicates the importance of the estimated third quartile regression model, and it was found that (64%) the percentage of influence of the independent variables(x_1,x_3 ,x4) on the dependent variable because the first marginal slope parameter (b1, b3, b4) has a probability value smaller than the significance level of 0.05) and the remaining effects of external factors.

2-4: Choosing the best estimated regression model :

A comparison was made between the general linear regression model and the quantitative Robust regression models through statistical comparison measures, as follows:

1- Comparison between the general regression model and the quantitative robust regression model (first quartile)

Table (6) comparison standards(R_{ols}) & ($QR_{0.25}$)

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

model	AIC	HOC	BIC	
R ols	692.2426	696.8382	703.7629	
$QR_{0.25}$	669.5764	674.1720	681.0968	BEST

By looking at the table above, it turns out that the first quartile regression (QR $_{0.25}$) is the best because it has the lowest comparison standards. The regression equation is written as follows:

$QR_{0.25}$(y)= -238.539+0 .136x_1 + 2.155 x_2 − 0.082x_3 + 4.769x_4

The regression equation above showed that every one-unit increase in blood sugar for patients ranked in the first quarter leads to an increase in their urea by (0 .136), and that a one-unit increase in blood pressure leads to an increase in urea by (2.155), while there appeared to be no significant effect of fat on urea, and finally When the age of patients increases by one unit, it leads to an increase in urea by (4.769)

2- Comparison between the general regression model and the quantitative robust regression model (second quartile).

Table (7) comparison standards (QR $_{0.50}$) & (R_{ols})

model	AIC	HOC	BIC	
R ols	692.2426	696.8382	703.7629	
$QR_{0.50}$	686.9160	691.5116	706.4363	BEST

By looking at the table above, it turns out that the second quartile regression (QR $_{0.50}$) is the best because it has the lowest comparison standards. The regression equation is written as follows:

$QR_{0.5}$(y)= −369.389+0.419825x_1 + 3.71360 x_2 + 0.455756x_3 + 7.41797 x_4

The regression equation showed that every one-unit increase in blood sugar for patients ranked in the second quarter leads to an increase in their urea by (0.419825), while a one-unit increase in blood pressure leads to an increase in urea by (3.71360), While a single unit increase in blood fat for patients leads to an increase in their urea by (0.455756)., and finally, when

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

the age of patients increases by one unit, it leads to an increase in urea by (7.41797).

3- Comparison between the general regression model and the quantitative robust regression model (third quartile).

Table (6) comparison standards ($QR_{0.75}$) & (R_{ols})

model	AIC	HOC	BIC	
R_{ols}	692.2426	696.8382	703.7629	BEST
$QR_{0.50}$	705.1416	709.7372	716.6619	

By looking at the table above, it turns out that the general regression is the best because it has the lowest comparison standards. The regression equation is written as follows:

$$y_t = -366.522 + 0.603x_1 + 0.052x_2 + 0.643x_3 + 8.614\,x_4$$

The regression equation showed that every one-unit increase in the blood sugar level for patients classified in the third quarter leads to an increase in their urea by (0.603), while no significant effect appeared on blood pressure. As for a one-unit increase in the blood fat percentage, it leads to an increase in their urea by (0.643). Finally, when the age of patients increases by one unit, it leads to an increase in their urea by (8.614).

Conclusions:

1- All linear regression models appeared significant

2- The advantage of the quantitative hippocampal regression model (first quartile) compared to the general regression model

3- The advantage of the quantitative hippocampal regression model (second quartile) compared to the general regression model

4- The advantage of the general regression model compared to the quantitative hippocampal regression model (third quartile)

5- The significant effect of independent variables (sugar, high blood pressure, age) on the rise in blood urea levels within the limits of the first spring data.

6- The effect of all independent variables (sugar, high blood pressure, fat, age) on the high level of urea in the blood within the limits of the second spring data.

7- The effect of each of the independent variables (sugar, fat, and age) on the increase in the level of urea in the blood within the limits of the data of the third spring

References:

[1]: Badshah, W., & Bulut, M. (2020). Model selection procedures in bounds test of cointegration: Theoretical comparison and empirical evidence. Economies, 8(2), 49.http://dx.doi.org/10.3390/economies8020049*

[2]:Bassiouni, Abdel Rahim (2023). Journal of Trade and Finance. Faculty of Commerce

[3]: Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1460.. https://doi.org/10.1002/wics.1460*

[4]:de la Rubia ,José Moral& de la O (2022). Quantile regression, a little-known analysis option inpsychological research. International Journal of Psychology and Counselling. Vol. 14(3), pp. 26-35.

[5]: Waldmann, E. (2018). Quantile regression: a short story on how and why. Statistical Modelling, 18(3-4), 203-218.

[6]:Kazem, Safaa Karim(2009). Comparison between multiple linear regression model parameter estimates using (ols) and linear objective programming. Journal of the College of Administration and Economics. Issue 77 .

[7]:Koenker, Roger (2005). Quantile Regression. Cambridge University Press. pp. 146–7. ISBN 978-0-521-60827-5.

مجلة الغري للعلوم الاقتصادية والادارية

مجلد (20) عدد (3) 2024

[8]: Koenker, R. (2017). Quantile regression: 40 years on. Annual review of economics, 9(1), 155-176., https://doi.org/10.1920/wp.cem.2017.3617

[9]: Konstantopoulos, S., Li, W., Miller, S., & van der Ploeg, A. (2019). Using quantile regression to estimate intervention effects beyond the mean. Educational and psychological measurement, 79(5), 883-910.https://doi.org/10.1177/0013164419837321

[10]: Yong, L. (2014). Novel global harmony search algorithm for least absolute deviation. Journal of Applied Mathematics, 2014(1), 632975.http://dx.doi.org/10.1155/2014/632975.

[11]:O. O. John and E. C. Nduka (2009). QUANTILE REGRESSION ANALYSIS AS A ALTERNATIVE ROBUST TO ORDINARY LEAST SQUARES. Scientia Africana, Vol. 8 (No.2), December, 2009 pp 61-6. https://www.researchgate.net/publication/333403546

[12]: Graham, B. S., Hahn, J., Poirier, A., & Powell, J. L. (2015). Quantile regression with panel data (No. w21034). National Bureau of Economic Research.. http://www.nber.org/papers/w21034

[13]:Shahout, Muhammad Abu Bakr and Shatwan(2023). Predicting sales volume using the multiple linear regression method. Journal of Technical Research, Tanta University. Volume 1, Issue 1.

[14]:Wei, Y.; Pere, A.; Koenker, R.; He, X. (2006). "Quantile Regression Methods for Reference Growth Charts". *Statistics in Medicine*. **25** (8): 1369–1382. doi:10.1002/sim.2271. PMID 16143984. S2CID 7830193