Journal of AL-Rafidain University College for Sciences (2023); Issue 54; 454 - 461

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Choosing Quantile Regression Model Via Lasso Optimality for the Effect of the Salinity of Soil in Al-Shamia City-Iraq

Fadel H. Hadi	Asaad N. Hussein Mzedawee
fadel.alhusiny@qu.edu.iq	asaad.nasir@qu.edu.iq

College of Administration and Economics - University of Al-Qadisiyah, Al-Qadisiyah, Iraq.

Article Information

Article History:
Received: December, 16, 2022
Accepted: March, 3, 2023
Available Online: December, 31, 2023

Keywords:
Quantile regression model, Lasso Technique, Model selection, Salinity of Soil

Correspondence:
Fadel H. Hadi
fadel.alhusiny@qu.edu.iq

Abstract

The variable selection is special case of model selection, choosing optimal model is considered hard mater, to overcome this problem lasso technique have been used. The quantile regression is a good tool for achieving evaluated the relationship between response variable and a set of explanatory variables because, it is effectiveness without assumptions. When built-in between the quantile regression and lasso technique give us a good method to achieving coefficient estimation and variable selection in different type regression models. To a choosing optimal model that show the variables more effected on earth salinity of AL-shamia city the lasso model selection in quantile regression model have been used .

https://doi.org/10.55562/jrucs.v54i1.613

1. Introduction

Since seminal work of many researchers, model selection is became very popular in many science such us genetics , economic and medicinal, Etc. Analysis of the high dimensional data is become very difficult problem, to overcome this problem model selection have been used. It have a good feature for variable selection and coefficients estimation. The model selection procedure give us big area in interpretation our phenomenon under study, because of it exclude unimportant variables from our model and focus on important variables. There are many model selection methods such us backward elimination, stepwise selection, and forward selection. These methods much using with regression models. But it classical and need long time. Recently, more effective methods have been used in selecting variables and identifying important models. Such us, Lasso adaptive lasso and elastic-net etc. also these method are used with regressio . But these methods have a good feature ,it active for achieving variable selection and coefficient estimation in the same time. When mix between these methods with regression model give us optimal model. This model is consider best for representing studied phenomenon. Quantile regression model is consider important model ,because ,it have a good feature compared with other regression model. But, when the quantile regression overlapping with variable selection method give us best models. In this paper, we used the lasso method with quantile regression model to choosing optimal model salinity of earth. This paper organized as following . Some concept about quantile regression model have been used in first section. In the second section showed lasso quantile regression. In third section showed analysis of real data . The conclusion and recommendation have been shown in forth section

2. Some concept about quantile regression model

The classical linear regression (CLReg) model is a procedure for measuring the relationship between dependent variables (y) and a set of independent variables (Nelder, J. A., & Baker, R. J. (1972)).

The CLReg model is focus on the mean of response variable (E(Y|X)). It is represented by the conditional mean of the response variable (Y) given X, E(Y|X). This mean E(Y|X) give us only one regression line has been estimated (Horowitz, J. L., & Spokoiny, V. G. (2001)).

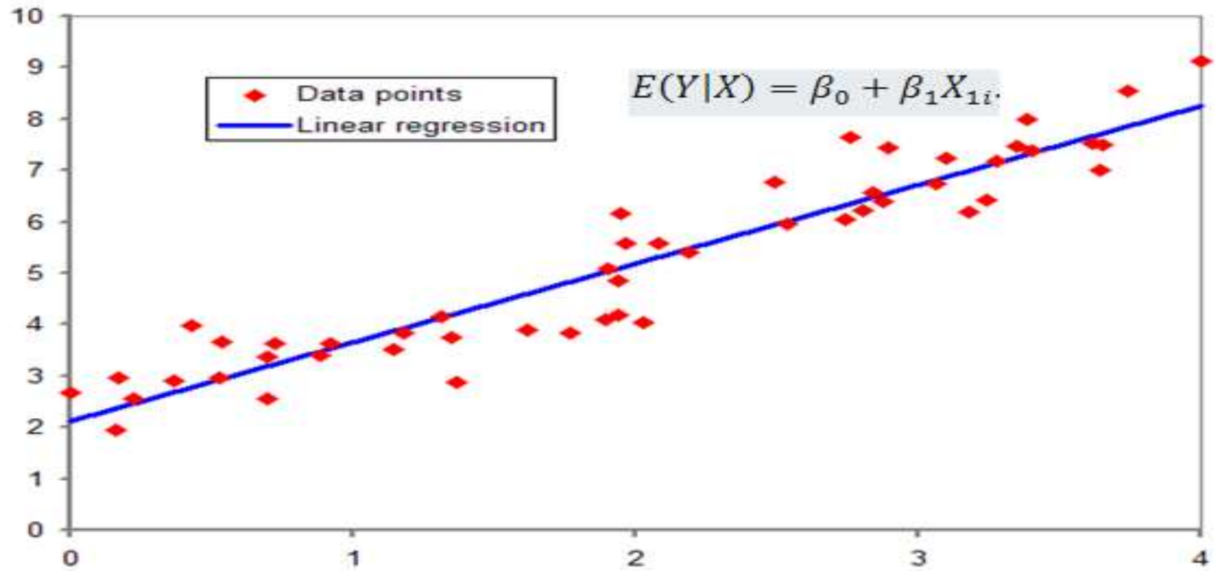

Figure (1): Coefficients estimation of the classical regression model

The CLReg model requires a set of suppositions about the independent variables, ,response variable and the random error (Myers, R. H. (2000)]. When achieved all these assumptions , CLReg model becomes very active for describing the relationship between independent variables and dependent variable. CLReg censored model with real limitations. Firstly, CLReg model cannot accommodate the non-central locations Secondly, sometimes the normal assumptions of the CLReg model are not achieving with applications, such us the random error have heavy tailed, via outliers values in the dataset. In this case, CLReg model becomes inactive because it is very sensitive to outliers values in the dataset. Also, the CLReg model requires linearity, independence and lack of multicollinearity etc. CLReg models have some limitations , any violation of these limitations leads to fake regression model . The quantile regression model give us the solution , because it very active until Even when violated normal assumptions. Also quantile regression model is very robust with outliers values (Cade, B. S., & Noon, B. R. (2003)). Therefore quantile regression model is active with heavy tails error distribution. Finally, the quantile regression have high ability for estimating the relationship between the independent variables and a dependent variable in any area of data spread (Levin, J. (2002)). Important part with quantile regression is called quantile levels ,it is within to the interval (0,1). The quantile regression model focus on conditional quantiles function($Q_{y_{i|x_i}}(\tau)$) .This amount $Q_{y_{i|x_i}}(\tau)$) is definition the conditional quantile function at the specific quantile level (τ_{th}).

Where the amount $Q_{y_{i|x_i}}(\tau) = F^{-1}_{y_{i|x_i}}(\tau)$. CLReg model is focused on the amount conditional mean. While the Q Reg model is focused this amount $Q_{y_{i|x_i}}(\tau)$, where $0 < \tau < 1$ (conditional quantile function), The quantile regression model very flexibility to study the phenomena at various data segments, because it focus on Infinity of regression lines. Also, the quantile regression model provide us a complete relationship between a dependent variable and independent variables Infinity of regression lines (see figure 2). In other words, the Q Reg model can assess the relationship between a set of covariates and a response variable, at many positions in the distribution of the response variable (Koenker, R., & Xiao, Z. (2002)).

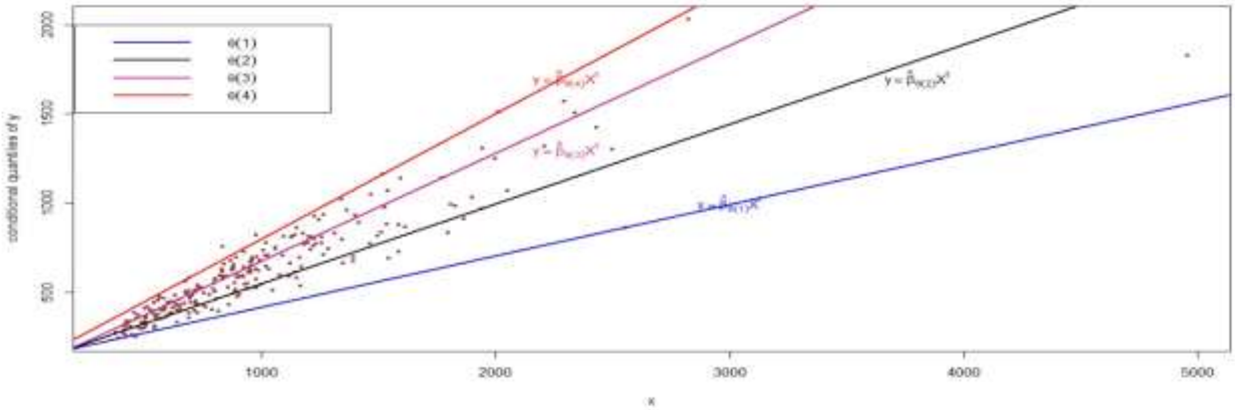

Figure (2): quantile regression lines via four quantile levels ($\tau_1, \tau_2, \tau_3, \tau_4$)

In figure (2) , there are four quantile regression lines, each one line belong to quantile regression model at specific quantile level. All these features makes the quantile regression common in empirical applications. it used in many different applications such us, agricultural economics (Kostov and Davidova,(2013)) ,Microarray study (Wang and He, (2007)), growth chart (Wei et al., (2006)) ecological studies (Cade and Noon,(2003)), etc. The quantile regression model can be written as:

$$y_i = x_i^T \beta_\tau + \epsilon_i, \qquad \tau \in (0,1), \tag{1}$$

where x_i^T is a $1 \times p$, p is matrix of independent variables , β_θ is a $p \times 1$ of unknown parameters vector, and τ is the quantile level. Here, ϵ_i is the residual term . Koenker and Bassett, (1978) provided that the coefficients estimation of β_τ can be estimated by:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta_\tau) \tag{2}$$

where $\rho_\tau(u)$ is the loss function , the equation (2) is not differentiable at the (0). Koenker and D'Orey, (1987) provided the minimization of (2) via a linear programming algorithm.

The variables selection procedure give us the forecasting accuracy and informative interpretation (Alhamzawi et al., (2012)). The philosophy of variables selection summarized choose optimal coefficients and exclude unimportant coefficients. For example, Lasso Tibshirani, (1996)), the adaptive Lasso (Zou, (2006)),etc. When , the quantile regression linked with Least Absolute Shrinkage and Selection Operator (lasso) give us model with a high capacity for interpretation and generalization

3. lasso quantile regression model

Lasso Regression model is important method for variable selection and coefficients estimation at same time. The lasso method one important method in variable selection field, that it introduced by Tibshirani (1996)). The Lasso regression coefficients are given by:

$$\hat{\beta}^{lasso} = \text{minimize} \sum_{i=1}^{n} \left(y - X\hat{\beta}\right)^2 \quad s.t \sum_{j=1}^{p} |\beta_j| \le t \tag{3}$$

where $\sum_{j=1}^{p} |\beta_j| \le t$ $\quad L_{1-}$ $norm\ for\ regreesion\ coefficents.$

t is tuning parameter which is responsible about quantity of shrinkage

There are another formulation of Lasso Regression model

$$\hat{\beta}^{lasso} = \text{minimize} \sum_{i=1}^{n} \left(y - X\hat{\beta}\right)^2 + \sum_{j=1}^{p} |\beta_j| \tag{4}$$

where λ *is shrankage parameter* , $\lambda \geq 0$.
The quantile regression model linked with lasso technique as follows:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta_\tau) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

where the ($\lambda \sum_{j=1}^{p} |\beta_j|$) is the penalty lasso and λ is parameter shrankage .

The equation (5) is achieve variable selection and coefficient estimation in quantile regression model at same time. Unfortunately, also the equation (5) is not differentiable at the (0) point . we see that the loss function is not differentiable at the (0) via the following figure ..

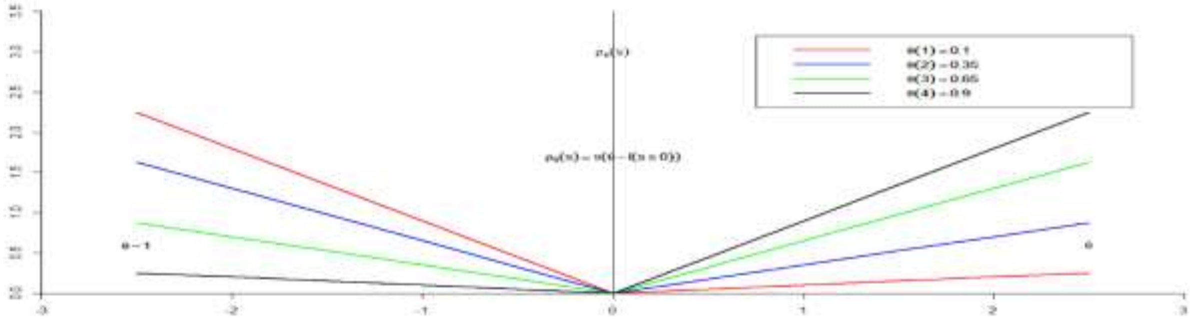

Figure (3): The check(loss) function at $\tau(1) = 0.10$ (red line), $\tau(2) = 0.35$ (blue line), $\tau(3) = 0.65$ (green line) and $\tau(4) = 0.65$ (black line)

For estimation the equation (5) , there are many package . in this paper we will used the rq.fit.lasso function with (quantreg) package.

4. Model selection of salinity of earth

It is the high level of salt in the soil. Soils are salty due to the accumulation of excess salts, which are usually more visible on the surface of the soil. Salts are transported to the surface of the soil by natural capillaries and are loaded from saline groundwater, then accumulate due to evaporation. Salinity can also build up in soils due to human activity. When soil salinity rises, the negative effects of salt rise, which can lead to degradation of soil and plants.

In this study, quantile regression model has been used to analyzing **earth salinity data**. The dataset was collected from Al-Shamia cultivation. The sample size of **earth salinity data** 300 zone. This data consists of the salinity amount referred to as (y: *response variable*). and eight independent variables are: The Soil type referred to as (x_1) , temperature referred to as (x_2), The number of once watering monthly referred to as (x_3), Groundwater level referred to as (x_4), The number of fish lakes in the zone referred to as (x_5), The amount of evaporation per day referred to as (x_6), and The amount of fertilizer used annually referred to as (x_7), number of Trocar used in one zone referred to as (x_8)., in this section we used four quantile levels $\tau \in \{0.25, 0.50, 0.75, 0.95\}$. as following:

4.1. Lasso quantile regression model at quantile level (0.25)

The following variable selection and coefficients estimation of earth salinity data by quantile regression at quantile level (0.25).

From the results listed in table (1), there are six independent variables effected in earth salinity. But the variables (Temperature and The amount of evaporation per day) are un effecting in earth salinity. The variable (amount of fertilizer used annually) has negative effect in earth salinity. But the rest independent variables have positive effect in earth salinity. This clear from the below figure

Table (1): show the parameters estimation and confidence interval (0.95%) at quantile level (0.25)

Variables name	Symbol of variables	parameters estimation	Upper limit (0.95%)	lower Limit (0.95%)
Soil type	x_1	0.9763	0.7442	1.311
Temperature	x_2	0.0000	-0.3421	0.9452
Number of once watering monthly	x_3	0.0276	0.0067	0.5342
Groundwater level	x_4	0.0885	0.00673	0.4532
number of fish lakes in the zone	x_5	-0.1118	-0.2341	0.4523
The amount of evaporation per day	x_6	0.0000	-0.3421	0.3412
amount of fertilizer used annually	x_7	-0.0655	-0.1796	0.3410
number of Trocar used in one zone	x_8	0.0986	0.5632	1.2031
BIC	0.1106376			

4.2. Lasso quantile regression model at quantile level (0.50)

The following variable selection and coefficients estimation of earth salinity data by quantile regression at quantile level (0.50).

Table (2): show the parameters estimation and confidence interval (0.95%) at quantile level (0. 50)

Variables name	Symbol of variables	parameters estimation	Upper limit (0.95%)	lower Limit (0.95%)
Soil type	x_1	0.9810	0.3530	1.0561
Temperature	x_2	-0.0336	-0.1562	0.2516
Number of once watering monthly	x_3	0.0795	0.0014	0.6442
Groundwater level	x_4	0.0000	-0.22751	0.37871
number of fish lakes in the zone	x_5	-0.0419	-0.1644	0.8459
The amount of evaporation per day	x_6	0.0214	0.0016	0.7342
amount of fertilizer used annually	x_7	0.0000	-0.28335	0.56623
number of Trocar used in one zone	x_8	0.2314	0.0542	0.8031
BIC	0.02816844			

From the results listed in table-2- ,there are six independent variables effected in earth salinity. But the variables (Groundwater level and amount of fertilizer used annually) are un effecting in earth salinity. There are two independent variable (Temperature, number of fish lakes in the zone) has negative effect in earth salinity. But the rest independent variables have positive effect in earth salinity. This clear from the below figure

4.3. Lasso quantile regression model at quantile level (0.75)

The following variable selection and coefficients estimation of earth salinity data by quantile regression at quantile level (0.50).

Table (3): show the parameters estimation and confidence interval (0.95%) at quantile level (0. 75)

Variables name	Symbol of variables	parameters estimation	Upper limit (0.95%)	lower Limit (0.95%)
Soil type	x_1	0.9963	0.3530	1.0561
Temperature	x_2	0.0000	-0.1562	0.2516
Number of once watering monthly	x_3	-0.0390	0.0014	0.6442
Groundwater level	x_4	0.0000	-0.22751	0.37871
number of fish lakes in the zone	x_5	0.0000	-0.1644	0.8459
The amount of evaporation per day	x_6	0.0204	0.0016	0.7342
amount of fertilizer used annually	x_7	0.0000	-0.28335	0.56623
number of Trocar used in one zone	x_8	0.0099	0.0542	0.8031
BIC	0.01976			

From the results listed in table-3- ,there are three independent variables effected in earth salinity. But the rest independent variables are un effecting in earth salinity. (Number of once watering monthly) has negative effect in earth salinity. But the rest independent variables have positive effect in earth salinity.

4.4. Lasso quantile regression model at quantile level (0.95)

The following variable selection and coefficients estimation of earth salinity data by quantile regression at quantile level (0.95).

Table (4): show the parameters estimation and confidence interval (0.95%) at quantile level (0. 95)

Variables name	Symbol of variables	parameters estimation	Upper limit (0.95%)	lower Limit (0.95%)
Soil type	x_1	0.9506	0.2134	1.2721
Temperature	x_2	0.0000	-0.1211	0.4290
Number of once watering monthly	x_3	0.0000	-0.0254	0.2835
Groundwater level	x_4	0.00245	0.00051	0.37871
number of fish lakes in the zone	x_5	0.0000	-0.2122	0.3445
The amount of evaporation per day	x_6	0.0000	-0.2321	0.3521
amount of fertilizer used annually	x_7	0.11730	0.03435	0.56623
number of Trocar used in one zone	x_8	0.0000	-0.1573	0.6431
BIC	0.03581694			

From the results listed in table (3), there are three independent variables effected in earth salinity. But the rest independent variables are un effecting in earth salinity. All independent variables have positive effected in earth salinity.

5. Conclusion and recommendations

5.1. Conclusions

All our models under study have importance and Clarity in explain the relationship between the response variable (earth salinity) and the independent variables. The first model: quantile regression model at (0.25) is modeling as following

$$Q_{y_{i|x_i}}(0.25) = 0.9763x_{1i} + 0.0276x_{3i} + 0.0885x_{4i} - 0.1118x_{5i} + -0.0655x_{7i} + 0.0986x_{8i}$$

In conditional quantile regression ($Q_{y_{i|x_i}}(0.25)$) , there are two independent variable not effected . The second model : quantile regression model at (0.50) is modeling as following

$$Q_{y_{i|x_i}}(0.50) = 0.9810x_{1i} - 0.0336x_{2i} + 0.0795x_{3i} - 0.0419x_{5i} + +0.0214x_{6i} + 0.2314x_{8i}$$

In conditional quantile regression ($Q_{y_{i|x_i}}(0.50)$) , there are two independent variable not effected . The third model : quantile regression model at (0.75) is modeling as following

$Q_{y_{i|x_i}}(0.75) = 0.9963x_{1i} - 0.0390x_{3i} + 0.0204x_{6i} + 0.0099x_{8i}$In conditional quantile regression ($Q_{y_{i|x_i}}(0.75)$) , there are four independent variable not effected

The fourth model : quantile regression model at (0.95) is modeling as following:

$$Q_{y_{i|x_i}}(0.95) = 0.9963x_{1i} - 0.0390x_{4i} + 0.0204x_{7i}$$

In conditional quantile regression ($Q_{y_{i|x_i}}(0.95)$) , there are five independent variable not effected. The best model can representation the relationship between the earth salinity and independent variables with quantile regression model at quantile level (0.25)

5.2. Recommendations

Expanding this study with Penalized functions that are more flexible in selecting variables and estimating the parameters of the studied model such us using Elastic Net and adaptive lasso. We recommended used much quantile leaves in evaluating the relationship between earth salinity and independent variables.

References

[1] Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). "Bayesian adaptive Lasso quantile regression", Statistical Modelling, 12(3), 279-297.

[2] Cade, B. S. and Noon, B. R. (2003). "A gentle introduction to quantile regression for ecologists." Frontiers in Ecology and the Environment, 1(8): 412–420.

[3] Horowitz, J. L., & Spokoiny, V. G. (2001). "An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica, 69(3), 599-631.

[4] Koenker, R. and G. J. Bassett (1978). Regression quantiles. Econometrica Vol. 46, No. 1, 33–50.

[5] Koenker, R. and V. D'Orey (1987). "Algorithm AS 229: Computing regression quantiles", Journal of the Royal Statistical Society. Series C (Applied Statistics)Vol. 36, No. 3.

[6] Koenker, R., & Xiao, Z. (2002). "Inference on the quantile regression process". Econometrica, 70(4), 1583-1612.

[7] Kostov, P., & Davidova, S. (2013). "A quantile regression analysis of the effect of farmers' attitudes and perceptions on market participation". Journal of Agricultural Economics, 64(1), 112-132.

[8] Levin, J. (2002). "For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement". In Economic Applications of Quantile Regression (pp. 221-246). Physica-Verlag HD.

[9] Myers, R. H. (2000). Classical and modern regression with applications (Duxbury Classic). Duxbury Press, Pacific Grove.

[10] Nelder, J. A., & Baker, R. J. (1972). "Generalized linear models: Wiley Online Library. reference growth charts", Statistics in medicine, 25(8): 1369–1382. 1

[11] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1 (1996), pp. 267-288

[12] Wang, H., G. Li, and G. Jiang (2007). "Robust regression shrinkage and consistent variable selection through the LAD-Lasso", Journal of Business & Economic Statistics, Vol. 25, No. 3, pp. 347–355.

[13] Wei, Y., Pere, A., Koenker, R., and He, X. (2006). "Quantile regression methods for reference growth charts" Statistics in medicine, Vol. 25, No. 8.

[14] Zou, H. (2006). "The adaptive lasso and its oracle properties", Journal of the American statistical association, Vol. 101, No. 476, 1418-1429.

مجلة كلية الرافدين الجامعة للعلوم (2023)؛ العدد 54؛ 454 - 461

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)
مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS
Journal of AL-Rafidain University College for Sciences

اختيار نموذج الانحدار الكمي عبر أمثلية لاسو لتأثير ملوحة التربة في مدينة الشامية- العراق

د. فاضل حميد هادي	د. أسعد ناصر حسين
fadel.alhusiny@qu.edu.iq	asaad.nasir@qu.edu.iq
قسم الإحصاء - كلية الإدارة والاقتصاد - جامعة القادسية، القادسية، العراق	

المستخلص

يعد الاختيار المتغير حالة خاصة لاختيار النموذج، ويعتبر اختيار النموذج الأمثل أمرًا صعبًا، للتغلب على هذه المشكلة تم استخدام تقنية اللاسو. يعد الانحدار الكمي أداة جيدة لتحقيق تقييم العلاقة بين متغير الاستجابة ومجموعة من المتغيرات التفسيرية لأنه فعال بدون افتراضات. عند الدمج بين تقنية الانحدار الكمي وتقنية اللاسو، يمنحنا طريقة جيدة لتحقيق تقدير المعامل واختيار المتغير في نماذج الانحدار المختلفة. لاختيار النموذج الأمثل الذي يبين المتغيرات الأكثر تأثيراً على ملوحة الأرض لمدينة الشامية تم استخدام اختيار نموذج اللاسو في نموذج الانحدار الكمي .

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 2022/12/16

تاريخ قبول البحث: 2023/3/3

تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية:

نموذج الانحدار الكمي، تقنية اللاسو، اختيار النموذج، ملوحة التربة.

للمراسلة:

د. فاضل حميد هادي

fadel.alhusiny@qu.edu.iq

https://doi.org/10.55562/jrucs.v54i1.613