ISSN: 2664-9640

Algae prediction model based on QR

Wen Xue

School of Mathematics and Statistics, Southwest University, Chongqing 400715, China

Abstract

This article mainly quantile regression(QR) and uses QR to realize the prediction oP seaweed data, judge the number oP the seventh type oP seaweed data to control the amount oP seaweed in the river, and Pinally predict the number oP seaweed by QR, The mean square error MSE oP the test is 6.513, which is 20.008 less than the mean square error MSE (27.400) predicted by the least squares linear regression. In order to further improve the accuracy, we choose to Pill in the k-nearest neighbor method in the image preprocessing part. Missing values, and using logarithmic linear regression Por comparison, Pound that the MSE oP logarithmic algae logarithm was 9.691, which was still not accurate compared with QR. Por the improvement oP the model, the BP neural network is adopted, and the MSE Pinally predicted by the BP neural network is 1.661, and the accuracy is significantly improved. But in the end, because the BP neural network is a black box algorithm, it cannot accurately control the independent variables to achieve the purpose oP controlling the amount algae in the river, so in the end, this paper still chooses QR. Continue to improve can combine principal component analysis (PCA) and QR, and finally get an MSE 0.566.

Keywords

Algae quantity; QR; Prediction; Mean Square Error(MSE).

1. Introduction

1.1. Research background

Multiple linear regression model is widely used in the process of interaction between dependent variables and independent variables. Among them, the least square method is most widely used to estimate the regression coefficient. When the random error terms of a linear model are independent and identically distributed in a distribution with the same mean value and variance, the least squares estimators of regression coefficients are unbiased estimators (blue); if the random error terms are independent and identically distributed in a normal distribution, the least squares estimators of regression coefficients are minimum variance unbiased estimators (mvul) and equal to their maximum likelihood estimators. The estimation of the above regression coefficients satisfies the properties of unbiasedness and validity.

In real life, the assumption of random error term cannot be satisfied. When the data distribution is obviously skewed, the estimation of least square method will no longer have the above properties. In order to make up for the defects of OLS(Ordinary least squares) in regression analysis, Laplace [1] proposed median regression (minimum absolute deviation estimation) in 1818. On this basis, in 1978, Koenker and Bassett [2] extended the median regression to the general quantile regression(QR), regressed the independent variables according to the conditional quantiles of the dependent variables, and obtained the regression model under all quantiles. Compared with OLS regression, QR can more accurately describe the change range of independent variable to dependent variable.

ISSN: 2664-9640

1.2. Research issues

This paper tries to solve the following problems

1.The trade-off between too many independent variables in seaweed data;

2.How to make the dependent variable become approximately normal distribution to meet the model hypothesis in linear regression;

3.How to choose the appropriate quantile in QR;

4.How to use QR based on principal component analysis.

1.3. Journals reviewed

Because the traditional linear regression requires the random variable ε be normal distribution and the variance is equal, in real life, these two hypotheses are often not satisfied. If we continue to use the OLR at this time, either the F test of the model or the coefficient T test of the model may not meet the P value less than 0.05 and fall short. In contrast, QR model does not need to make specific assumptions about distribution, nor does it use connection function to describe the relationship between dependent variables and independent variables, and is not affected by outliers, so the model is more stable [3]. In this paper, we try to determine the use of 0.6 QR, the so-called 0.6 QR, that is, the regression curve can contain 60% of the data points (y), the concept of quantile into the OLR, we get the QR. QR is a kind of regression model, we can apply it to linear regression, polynomial regression, kernel regression, nonlinear regression and so on. In practical operation, we change the loss function from least square method to weighted least square method, get different results through different quantiles, then analyze and choose according to the results.

2. Theoretical framework

2.1. Basic model

(1) Ordinary least squares（OLS）

If the model after linear regression passes the F test, it proves that the linear relationship of the model is significant; if the t test of the regression coefficient also passes, it indicates that the estimation of the regression coefficient can be considered as the unbiased estimation of the regression coefficient, and then studies the correlation between the dependent variable and the independent variable.Multiple linear regression studies dependent variable y and the linnear relationship of controllable variableX_1、X_2、X_3、……、X_n:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + s,$$

From the knowledge of linear algebra, it can be transformed into:

the least squares of estimation β: $Y = \beta X + \varepsilon$, $\hat{\beta} = (X^T X)^{-1} X^T Y$

Principal component analysis(PCA)

When the number of independent variables is large and there is a complex relationship, multicollinearity is easy to occur, which affects the unbiased and effective estimation of regression coefficients. PCA can reduce the dimension of multiple variables into a few variables, and can make the principal components represent the vast majority of information of the original variables and be uncorrelated with each other.

Several principal components with a total variance contribution rate of more than 85% are selected. According to the independent variables with different coefficients of each principal component, we can approximately judge which aspect this principal component represents, such as economic principal component, health principal component, etc. In general, each

ISSN: 2664-9640

independent variable in the first principal component will account for a certain proportion, because the first principal component represents a random error term, so it can not be analyzed.

Quantile regression(QR)

The estimators of regression coefficients under different quantiles are often different, that is, the dependent variable is greatly influenced by the independent variable at different levels.

Compared with the least square method, the QR method is more robust to outliers. Moreover, the QR does not require strong assumptions on the error term, so for the data with non normal random error term, the QR coefficient estimator is more robust.

2.2. Basic assumptions of the model

The random error terms are independent of each other;

There is no correlation between random error term and independent variable;

The random error term is independent and identically distributed in the normal distribution of zero mean, homo variance and zero covariance;

The independent variable is not a random variable.

3. Model

Firstly, there are missing values in the data. In order to preserve the integrity of the data, the knimputation function in the dmwr package is used to fill in the missing values using the k-nearest neighbor method. It finds the nearest K neighbors of any case according to KNN algorithm, and fills the missing values in the nearest neighbor case by setting function values (generally selecting mean, median, mode, etc.).

Next, consider the control group using least square regression as QR, so the dependent variable (A7) should meet the assumption of normal distribution and homogeneity of variance. From the QQ chart of the dependent variable, we can see that the dependent variable does not follow the normal distribution, so we do the logarithmic transformation and open transformation. It is worth noting that some data in the dependent variable is 0, and after the logarithmic transformation, it will become negative infinity. At this time, we directly replace the negative infinity with 0. Looking at Figure 1, this paper finally chooses logarithmic transformation, and the corresponding linear regression is changed to logarithmic linear regression.

Log the dependent variable A7, and then check the correlation coefficient for the dependent variable A7. It can be seen that there are few independent variables with strong correlation with A7, while there are some independent variables with strong correlation. For example, the correlation coefficient between PO2 and opo4 is 0.91, and the correlation coefficient between No3 and NH4 is 0.72. At this time, we need to use some methods to reduce multicollinearity and auto-correlation of explanatory variables. Through literature review, a ridge parameter selection method based on the principle of minimum mean square error unbiased estimation is adopted in ridge regression [4], so as to reduce multicollinearity. Similarly, Lasso regression [5]. This paper uses lasso regression (the input function should be in the form of matrix). Finally, according to the CP value, this paper takes the seventh step to minimize the CP value, that is, to select the first seven variables. It can be seen that the variables of lasso are MnO2, PO4, mxph, NO3, NH4, Cl, opo4, Chla.

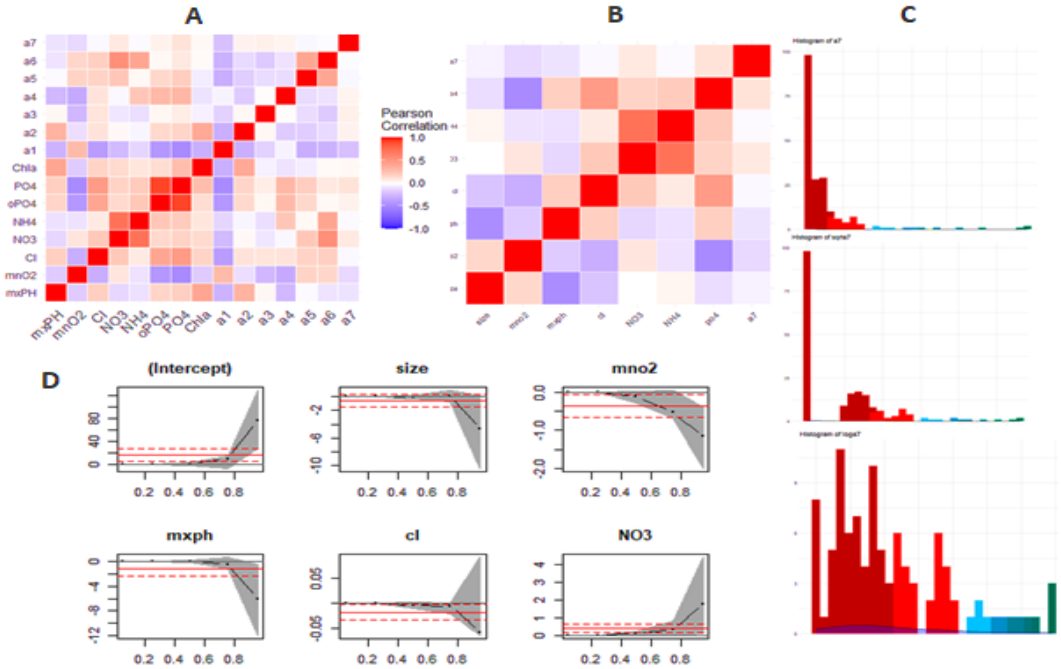
3.1. First model:OLS & QR

Compared with the least square linear regression and quantile linear regression, it is not difficult to find that when there are outliers in the graph, the fitting result of QR is better, and the result should also conform to the definition of quantile, that is to say, this paper uses tau = 0.6, then we minimize the loss function to find the parameters, and the regression curve P should have a certain value 60% of the data is below the curve (try tau = 0.25, 0.3, 0.45, 0.5, 0.6,

ISSN: 2664-9640

0.75, 0.9, and find that when tau = 0.6, the mean square error of quantile linear regression is the smallest). Therefore, QR is not a regression model, but a kind of regression model, or an improvement idea. We can apply it to linear regression, polynomial regression, kernel regression and so on. The most fundamental thing is to change the loss function from the least square method to the weighted least square method, and get different results through different quantiles, and then calculate according to the results analysis.

In the least square method, P test is 2.444×10^{-12}, which shows that the linear relationship of the model is established. From table 1, when No3 increases by one unit, the number of algae will increase by 0.47642; when NH4 increases by one unit, the number of algae will decrease by 0.0009. Therefore, to control the number of algae, a certain amount of algae can be added to the river NH4 and reduce No3 emission to the river.



Figur1: The correlation graph of each variable in seaweed dataset (a); the correlation graph of independent variable and A7 screened by lasso (b); from top to bottom: frequency histogram of A7, frequency histogram of A7 after exponential transformation, frequency histogram of A7 after logarithmic transformation (c); the result of QR when quantile is 0.2, 0.4, 0.6, 0.8 (d)

Table 1 : The results of least square linear regression and QR

Name	Coefficient	SD	t test	p value	Coefficient
SIZE	-0.9489	0.71535	-1.326	0.1869	-0.0929
MNO2	-0.4483	0.23219	-1.931	0.0555	-0.1386
MXPH	-1.5118	0.90683	-1.667	0.0977	0.19764
Cl	-0.0189	0.01152	-1.642	0.1029	-0.0062
NO3	0.47642	0.20666	2.305	0.0226*	0.16734
NH4	-0.0009	0.00040	-2.228	0.0275*	-0.0003
PO4	0.00625	0.00492	1.271	0.2059	0.00412
Constant	19.87516	8.36204	2.377	0.0188*	0.48105

ISSN: 2664-9640

Table 2 : The results of the linear regression of the logarithmic least square method

Name	MNO2	MXPH	cl	NO3	NH4	PO4
constant	-0.0625	-0.0899	-0.0030	0.05398	-0.0001	0.0012
SD	0.02980	0.1160	0.00013	0.02458	0.0000	0.0006
t test	-2.099	-0.776	-2.036	2.196	-2.264	2.046
P value	0.0372*	0.4388	0.0432*	0.0293*	0.0247*	0.0421*

In the least square method, P test is 0.019, which shows that the linear relationship of the model is established. From table 1, when MnO2 increases by one unit, the number of seaweed will decrease by $e^{0.0125}$; when CL increases by one unit, the number of seaweed will decrease by $e^{0.0030}$; when No3 increases by one unit, the number of seaweed will increase by $e^{0.05398}$, When NH4 is increased by one unit, the number of algae will decrease by $e^{0.0001}$, and when PO4 is increased by one unit, the number of algae will decrease by $e^{0.0421}$. Therefore, to control the number of algae, a certain concentration of NH4, MnO2 and CL can be increased to the river, the discharge of No3 and PO4 to the river can be reduced.

Finally, the mean square error predicted by least square method is 27.400, the mean square error predicted by logarithmic least square method is 9.691, and the mean square error predicted by QR is 6.513.

3.2. Second model:BP neural network

BP (back propagation) neural network is the most traditional neural network. That is the neural network using the back propagation algorithm. According to the possible problems of BP neural network for the data set with small amount of data, it is easy to over fit because of the small amount of data. The idea of BP neural network algorithm is to use the error back propagation to correct the weight, so as to minimize the mean square error between the deeds output and the expected output of the network. In the prediction of seaweed problem: A7 is the output unit, and each variable is the input unit K. suppose there is a hidden unit V, and the arrow from the input unit K to the hidden unit V represents the connection weight.

Finally, the mean square error of BP neural network training prediction value and the actual value is 1.6614, although it is 4.85126 lower than the best QR mean square error so far, because the neural network is a black box algorithm, it can not know the relationship between explanatory variables and explained variables, so it needs better white box algorithm to improve.

3.3. Third model:QR based on PCA

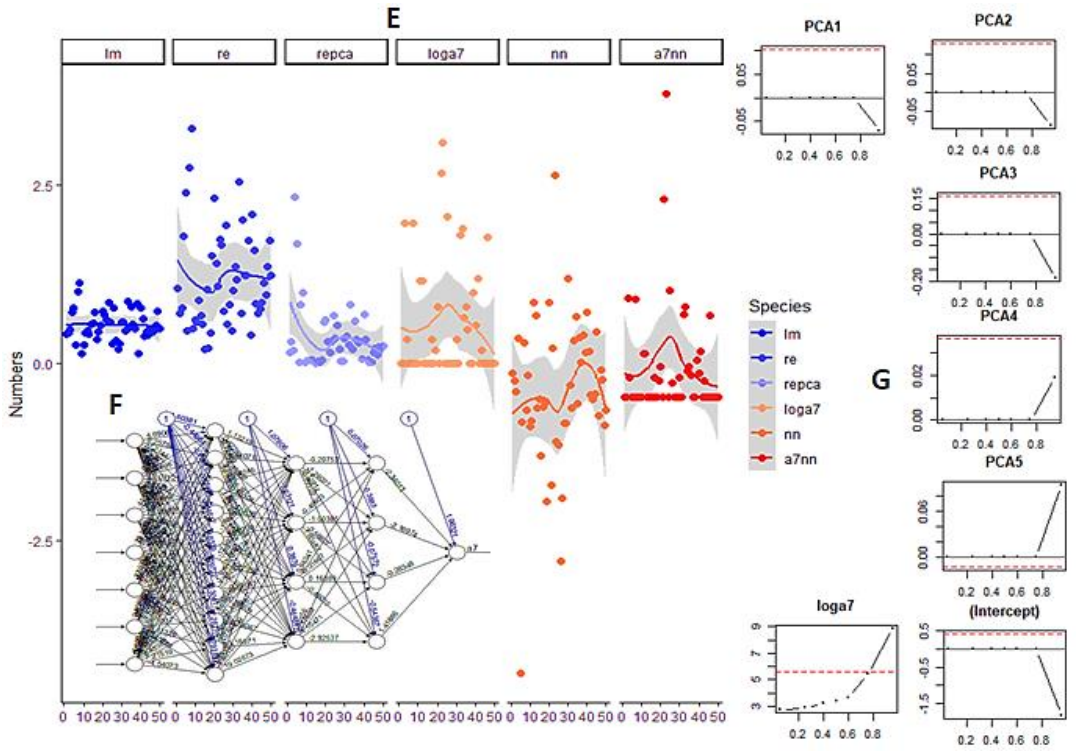
Through the first model, we have learned the advantages of QR, and the mean square error of QR is small, so this paper starts from how to improve QR. Because there are still some autocorrelation and multicollinearity among the existing explanatory variables, this paper first chooses a series of dimension reduction methods, such as stepwise regression, principal component analysis and so on. The results of stepwise regression are not ideal. After using principal component analysis, when the explanatory variables in the training set are replaced by the principal components which are not related to each other and have no redundant information, the results of QR are better than expected, and the mean square error of prediction fitting is 0.5658, even lower than that of neural network which has always been known for its accuracy.

First, the results of principal component analysis are shown in Table 3:

ISSN: 2664-9640

Table 3 : Results of principal component analysis

Name	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7
Size	0.282	0.333	0.587		0.602	0.299	
MNO2	0.414	0.199	-0.444	0.536	0.287	-0.403	-0.244
MXpH	-0.279	-0.422	-0.447	-0.118	0.673	0.266	
Cl	-0.478		0.132	0.744	-0.130	0.366	-0.219
NO3	-0.281	0.589	-0.290	0.102			0.692
NH4	-0.295	0.567	-0.192	-0.359		0.143	-0.633
PO4	-0.526		0.346		0.283	-0.723	
Var	0.3117	0.2562	0.1604	0.1041	0.0776	0.0592	0.3094
Total Var	0.3117	0.5679	0.7283	0.8323	0.9099	0.9691	1.0000



Figur2: The comparison of OLS, QR, PCA + QR and BP neural network predicted values with the real values in the test set (e); the schematic diagram of the neural network corresponding to this data set: a three-layer hidden layer with 10 neurons, 4 neurons and 4 neurons (f); the results of PCA + QR at quantiles of 0.2, 0.4, 0.6, 0.75 and 0.8 (g)

The first five principal components were 90.99% :

PCA1=0.282*size+0.414*mno2-0.279*mxph-0.478*cl-0.281*NO3-0.295*NH4-0.521*po4

PCA2=0.333*size+0.199*mno2-0.422*mxph+0.589*NO3+0.517*NH4

PCA3=0.587*size-0.444*mno2-0.447*mxph+0.132*cl-0.290*NO3-0.192*NH4+0.34*po4

PCA4=0.531*mno2-0.118*mxph+0.744*cl+0.102*NO3-0.359*NH4

PCA5=0.102*size+0.287*mno2+0.173*mxph-0.130*cl+0.283*po4

ISSN: 2664-9640

Then the five principal components and log (A7) were used as a data set, and 0.6 QR was performed:

$y = 0.00072 \cdot PCA1 + 0.00117 \cdot PCA2 + 0.00811 \cdot PCA3 - 0.00210 \cdot PCA4 + 0.00159 \cdot PCA5 + 0.05253$

It looks like maybe we can modify:

$y = 0.0018 \cdot size — 0.00042 \cdot MNO2 — 0.0034 \cdot MXPH + 0.00024 \cdot Cl — 0.002 \cdot NO3 + 0.0523$

It is easy to know that in the PCA + QR model, NH4 and PO4 have little effect on the number of algae. In other words, when the algae are larger, the number of algae in the corresponding level will be more, $e^{0.0018}$. Appropriate input of certain concentrations of MnO2, mxph, Cl and NO3 in the River can reduce the number of algae in the river.

4. Relevant conclusions

4.1. Advantages of the model PCA+QR

Compared with the least square method, the QR method is more robust to outliers. Moreover, the QR does not require strong assumptions on the error term, so the QR coefficient estimator is more robust for non normal distribution. Because there are many outliers in this paper, when using PCA + 0.6 QR, the mean square error of the model is the least

The prediction accuracy of QR with principal component analysis is even higher than that of BP neural network, which may be due to the use of principal component instead of data set for QR, which effectively reduces multicollinearity and auto-correlation.

Compared with the least square method, the accuracy of least square method is higher, and the least square method is easy to be affected by outliers when estimating regression parameters. Although QR has no better way to solve outliers, when the QR model with quantile of 0.6 is proposed, it is equivalent to eliminating outliers. It is not difficult to find through the QR result chart. When the quantile exceeds 0.8, the QR model will be established And the accuracy of that will drop dramatically.

4.2. Disadvantages of the model PCA+QR

(1) It is known that the QR model used here is 0.6 QR, so this also means that only 60% of the data is included in the regression curve.

In how to select the quantile, we can only try one by one, and can not achieve the effect of parameter training.

In fact, in literature [6], some researchers have combined QR with neural network, and achieved good results. Therefore, compared with the above combined model, this model is still insufficient, but the data set in this paper is less, and it is easy to over fit.

4.3. Improvment of the model PCA+QR

In order to solve the above problems, the most reasonable method is to add more samples into the training set of the model, because the sample size used in this paper is 198, in fact, for the neural network, it is a small amount of data, easy to appear over fitting. The combination of neural network and QR is still worth studying, but sometimes we need to pay attention to that some neural networks are not open source in R, so if we want to try, we can only try more basic neural networks, but if these neural networks are really applied to deep learning, it is not enough. At the same time, we can consider building a neural network, but it requires too much equipment.

For QR, this paper uses quantile linear regression. Correspondingly, we can try quantile nonlinear regression. There is no corresponding packet sum function on R, and there is a ready-made function on the network, but it is only applicable to the case of unary nonlinear.

ISSN: 2664-9640

References

[1]Laplace, P. S: Theorie analytique des probabilities, Editions Jacques Gabay, Paris, 1818.

[2]Koenker, R. and Bassett. G: The Asymptotic Distribution oP the Least Absolute Error Estimator, Journal oP the American Statistical Association, 1978, 73: 618-622.

[3] Li Yu'an. Introduction to quantile regression and its application [J]. Statistics and information forum, 2006, 21 (3): 35-38

[4]Zhang Fenglian. Discussion on the solution of multicollinearity in multiple linear regression [D]. South China University of technology, 2010

[5]Tao Chunhai, Wang Mengying. Analysis of factors influencing the proportion of personal health expenditure in China based on lasso regression model [J]. Statistics and decision making, 2017, 000 (021): 100-103

[6]He Yaoyao, Xu Qifa, Yang Shanlin, et al. Power load probability density forecasting method based on quantile regression of RBP neural network [J]. Chinese Journal of electrical engineering, 2013, 033 (001): 93-98, inserted 12

[7]He Yun. Research on the relationship between corporate governance and M & a performance of China's A-share listed companies -- Based on quantile regression analysis [J]. Journal of graduate students of Central South University of economics and law, 2017, 000 (001): p.33-41

[8]Qin Tianyan, Zhang Jiwei, lazati mulati, et al. Analysis on Influencing Factors of health literacy of college students in Lanzhou Based on Quantile Regression [J]. Modern preventive medicine, 2018, v.45 (08): 93-97