International Journal of Statistics and Applications 2018, 8(2): 72-78
DOI: 10.5923/j.statistics.20180802.06

Ridge Parameter in Quantile Regression Models. An Application in Biostatistics

Ali Sadig Mohommed Bager

Department of Statistics and Econometrics, The Bucharest University of Economic Studies, Muthanna University, Iraq

Abstract In quantile regression, usually the explanatory variables in medical data are highly correlated with each other. Thus, the influence of one variable can't be differentiated from the others. Also, multicollinearity generates unstable regression coefficients with undesirable large variances. It is possible that the estimated coefficients may have the incorrect signs and present problems in the interpretation. In this paper ridge regression is applied to solve the problem of multicollinearity. An optimum ridge coefficient for the ridge regression parameter can be estimated through Bayesian approach. The Bayesian approach is a method to stabilize the ridge parameter. The quantile regression is the best method to predict an extreme value. This study discusses the use of ridge regression in quantile regression with a parameter ridge. Variance inflation factor (VIF) is used to determine the best ridge coefficient. The results show that the quantile regression with ridge regression was suitable for the study of the causes of genetic anemia (Thalassemia) in children.

Keywords Quantile regression, Ridge regression, Bayesian approach, Genetic blood diseases (Thalassemia)

1. Introduction

Thalassemia is an inherited disease that is transmitted from parents to children across genes and affects the ability to produce hemoglobin in the human body, leading to severe anemia. The most common type is found among children with anemia in all parts of world. This disease is caused by genetic acquired factors. Iron deficiency anemia affects a large number of children, and significantly more than other types of anemia in poor and developing countries, where there is an increased incidence of some infections [1]. The World Health Organization has calculated that about 7% of the world's population carry a hemoglobinopathy gene [2]. In the Mediterranean area, there are 15 to 25 million of healthy carriers [3]. Iraq is one of the countries with health, environmental and economic difficulties [4]. The statistics announced by government agencies on blood diseases confirmed high rates of patients with Thalassemia in the last few years, especially among children. The increase in the incidence of infection is due to several reasons, including the remnants of war and its impact on generations as well as the genetic factor and lack of health awareness among people.

Statistics show that the of Thalassemia patients has increased significantly in recent years. After the 2003 war that the rates of this genetic blood diseases in children in Iraq has risen to 22 cases per 100 thousands children compared with 2002, when it only appeared in 4 children out of 100,000 children. This number is much higher compared to neighboring countries. Children with Thalassemia need a life-long treatment of regular blood transfusion and iron chelation. Thalassemia has negative impact on many aspects of the life for children.

Thus, it is necessary to acknowledge the causes of the disease and the variables that cause infection with the disease. In this case, the Quantile regression models help define the relationship between a set of proposed predictor variables and the specific quantiles of the response variables analogous to a linear regression, which investigate the mean value of the response variables for a given set of predictor variables [5]. Quantile regression leads to a more comprehensive analysis by estimating the changes in specific quantiles of the response variables with respect to predictor variables, providing relationship at different points in the conditional distribution of the response variable. Since the pioneering research of Koenker and Bassett [6], quantile regression models have been the topic of major theoretical interest as well as many practical applications in many different areas such as: economics, survival analysis, microarray study, growth chart, finance, biomedical studies. A comprehensive account of these recent applications can be found in Koenker [7], and Cade and Barry [8]. The ridge parameter in quantile regression is employed to address the most important challenges that may arise with medical data, such as multicollinearity [9]. This problem arises because of the high correlation between the independent variables that lead to weak estimates. It is

* Corresponding author:
nader.ali62@yahoo.com (Ali Sadig Mohommed Bager)
Published online at http://journal.sapub.org/statistics

important to note the fact that in analysis of factors affecting the incidence of Thalassemia, the variables in the phenomenon under study are linked to each other in one way or another, either directly or indirectly. The existence of these relationships have several effects on the results of the analysis obtained. Nevertheless, existing literature shows that this issue is not adequately addressed.

Many researchers have pursued different aspects and ways of solving multicollinearity. In 1975, Hoerl and Kennard developed a new method by adding a positive value (k) to the elements of the information matrix, the so called ridge parameter [10].

The ridge estimator is used to solve the identification problem that can be viewed from a Bayesian perspective (Congdon, [11]). In this paper we will use Bayesian methods to choose the ridge parameter.

In 2013, Cule, E. and De Iorio, M. identified numerous genetic associated with diverse phenotypic traits. However, the identified associations generally explain only a small proportion of trait heritability and the predictive power of models incorporating only known-associated variants has been small, considering the joint effect of many genetic variants simultaneously. The study found ridge regression, a penalized regression approach that has been shown to offer good performance in multivariate prediction problems [12].

The variance is higher in the OLS method if in the multiple regression model, the linear correlation between explanatory variables is high [13]. The results of this study were obtained by using ridge regression study. Also, the authors showed that the best results are achieved with the ridge regression method of Hoerl and Kennard's. Bager et al. (2017) aimed to determine the most important macroeconomic factors which affect the unemployment rate in Iraq by using a ridge regression method, one of the most widely used methods for solving the multicollinearity problem [14].

This paper aims to use the ridge parameter in the quantile regression to address important challenges, such as the multicollinearity problem. A Bayesian method is used for estimation of the ridge parameter in quantile regression. Because this technique could provide an important research tool, it may help to find the most appropriate value for the parameter. The paper would also contribute to develop a robust statistical model to identify the most important factors affecting patients with thalassemia across various quantile ratios. The model helps rank the relevant factors affecting patients. The findings of the study can be potentially useful for the identification of the causes of the disease in order to reduce its incidence.

The paper structure is the following: section two provides ridge regression method, section three includes the ridge and quantile regression models, section four covers the test of the multicollinearity problems, section five is the application to the case study and section six includes the conclusions.

2. Ridge Regression Method

The history of multicollinearity dates back at least to the paper of Frisch (1934) who introduced the concept in order to describe a situation where the variables dealt with are subject to two or more relations. One way to approach this problem is called the ridge regression, first introduced by Horel and Kennard [15] who suggested an alternative method to the standard method of ordinary least squares (OLS). The ordinary ridge regression method (ORR) has become one of the most applied solutions for addressing the problem of semi-multicollinearity.

The method implies adding a small positive constant (K) to the main diagonal elements of the information matrix (X^TX). This positive value, known as the ridge parameter, decodes the links between the explanatory variables. The ORR method can be written as follows:

$$\hat{\theta}_{ORR} = ((X^T X) + kI_n)^{-1} X^T y$$
$$= ((X^T X) + kI_n)^{-1} X^T X \hat{\theta}_{ols} \quad (1)$$

k>0: ridge parameter in identity matrix

When k=0, the ordinary ridge regression method converts to the ordinary least square method.

$$(I + k(X^T X)^{-1})^{-1} \hat{\theta}_{ORR} = \hat{\theta}_{OLS} \quad (2)$$

When introducing the ridge parameter k, the variation of estimated parameters is reduced. Although the ordinary ridge regression method is biased, it produces a mean square error (MSE) lower than the mean square error obtained with OLS method.

$$MSE(\hat{\theta}_{ORR}) < MSE(\hat{\theta}_{Ols}) \quad (3)$$

The ridge estimator is regarded as the minimize of the residual sum square as the following formal (Tibshirani, [16]):

$$(\hat{\Theta}_{ORR}) = \arg\min \sum_{i=1}^{n} (y_i - \mathrm{X}\Theta) + k_{ridge} \sum_{j=1}^{p} (\Theta)^2 \quad (4)$$

3. Ridge and Quantile Regression Models

In quantile regression there should be no multicollinearity in predictor variables. We will address this problem by using the ridge regression which can overcome the multicollinearity problem. An optimum ridge coefficient can be estimated through many methods. The quantile regression model has the following form [7]:

$$(\hat{\Theta}) = \arg\min \sum_{i=1}^{n} \rho_\Gamma \left(y_i - x_i^T \Theta_\Gamma \right), with i = 1, 2, ..., n$$
$$\rho\Gamma(.) = \begin{cases} (1-\Gamma)\left(y_i - x_i^T \Theta_\Gamma \right) if \left(y_i - x_i^T \Theta_\Gamma \right) < 0 \\ \Gamma\left(y_i - x_i^T \Theta_\Gamma \right) if \left(y_i - x_i^T \Theta_\Gamma \right) \geq 0 \end{cases} \quad (5)$$

The quantile regression with ridge regression use the ridge coefficients to build the quantile regression model. The solution of the ridge coefficient can be written as the following equation [17]:

$$(\hat{\Theta}) = \arg\min \sum_{i=1}^{n} \rho_{\Gamma}\left(y_i - x_i^T \Theta_{\Gamma}\right) + k_{ridge} \sum_{j=1}^{p} \left(\Theta_{\Gamma \cdot j}\right)^2 \quad (6)$$

where (k) ridge parameter.

3.1. Estimator Parameter (k)

The multiple linear regression model can be expressed as:

$$Y = X\theta + \varepsilon \quad (7)$$

where:

Y: Represents the vector of the dependent variable with dimension (n × 1)

x: Represents the vector of the independent variable with dimension (n × p)

θ: Represents the vector of parameter regression with dimension (P × 1)

ε: Represents the vector of random error with dimension (n × 1).

The ordinary least square estimator (OLS) of the regression coefficients θ is defined as [18]:

$$\hat{\theta} = \left(X^T X\right)^{-1} X^T y \quad (8)$$

Suppose the diagonal matrix $X^T X$ has the orthogonal conversion *p*; it results that $X^T X = p\Lambda p$ and the matrices *p* and *Λ* are matrices eigenvalue and eigenvector for matrix $X^T X$. $p^T X^T Xp = \Lambda = diagonal(d_1, d_2, ..., d_p)$ where d_i is the eigenvalue of number *j* of the matrix $X^T X$ and $p^T p = pp^T = I$. The orthogonal (canonical form) version of the multiple regression model (7) is [19] [20]:

$$Y = H\alpha + \varepsilon \quad (9)$$

Where:

$$\alpha = P^T \theta, H^T H, H = XP \quad (10)$$

$$\hat{\alpha}_{ols} = \left(P^T X^T XP\right)^{-1} P^T X^T y \quad (11)$$

$$= \left(P^T X^T XP\right)^{-1} P^T X^T X\hat{\theta}$$

$$= \left(P^T X^T XP\right)^{-1} P^T X^T XPP^T \hat{\theta}$$

$$= P^T \hat{\theta} \quad (12)$$

Then the estimator $\hat{\alpha} = P^T \hat{\theta}(k)$.

From equation (8) we can write:

$$\hat{\alpha}(k) = \left(I + k\left(H^T H\right)^{-1}\right)^{-1} \hat{\alpha}$$

$$= \left(I + k\left(P^T X^T XP\right)^{-1}\right)^{-1} \hat{\alpha}$$

where: $P^T X^T XP = \Lambda = diagonal(d_1, d_2, ..., d_p)$.

The equation above simplified:

$$\hat{\alpha}_j(k) = \left(I + k\left(H^T H\right)^{-1}\right)^{-1} \hat{\alpha} \quad (13)$$

$$j=1,2,\ldots\ldots p$$

From the above equation the value (k) can be chosen [21].

3.2. Bayesian Approach to Estimator Parameter (k)

Ridge regression has a close connection to Bayesian linear regression. The first review of the Bayesian derivation and interpretation of the ridge estimator along the line was provided by Loesgen [22]. The ridge estimator proposed to solve the identification problem can be viewed from a Bayesian perspective [8]. To the estimate ridge parameter, we depend on the result. Lindely and Smith [19] express the formula into Bayesian for estimating the linear regression model through an appreciation multi stages and use multivariate normal prior distribution when the parameter is unknown at every stage of the analysis. In this case, the prior is considered to be marginal distribution: $\alpha \sim N\left(0, \frac{\sigma^2}{k} I_n\right)$.

The posterior distribution can be rewritten to depend on Bayesian estimator k as follows [1], [23]:

$$\hat{\alpha}_j(k) = \left[\hat{\sigma}^2\left(d_1^{-1} + k^{-1}\right)\right]^{-1} \hat{\alpha}_j \quad (14)$$

where

$$\hat{\alpha}_j = \frac{y_i\left[I - X\left(X^T X\right)^{-1} X^T\right] y\left[y^T y - n\hat{\sigma}^2\right]}{n\left(tr\left(X^T X\right)\right)} \quad (15)$$

4. Test Multicollinearity Problems (Farrar & Glauber Test)

This test is based on a Chi square test (x^2) to determine whether or not there is a multicollinearity problem in the estimated model.

The test formula is:

$$x^2 = -\left[N - 1 - \frac{1}{6}(2m + 5)\right] \log|D| \quad (16)$$

where:

N: represents the sample size.

m: represents the number of explanatory variables.

log|D|: represents the normal logarithm to determinant the correlation matrix.

Then we compared between the table values in a degree of freedom ((m(m-1))/2) and specific significance level. The null hypothesis is rejected if the calculated value is greater

than the tabular value. It isn't rejected if the calculated value is smaller than the tabular value [24].

5. The Sample and Data Analysis

The inasmuch prevalence of genetic blood diseases (Thalassemia) varies among areas of Iraq with the highest incidence of infections in the middle Al-Forat region and southern Iraq, according to the Iraqi Ministry of Health (National Cancer Council). This study was conducted to determine the factors affecting the children infected with some genetic anemia (Thalassemia). The sample was taken for 152 patients from centers for the treatment of blood diseases in that areas during 2016. This paper used the quantile regression model at four quantile levels (0.25, 0.50, 0.75, 0.95) respectively, because quantile regression has attractive properties. The variables that are thought to affect the disease were selected after the consultation of doctors specialists and practitioners in genetic anemia (Thalassemia). The following table (Table 1) describes the variables:

Table 1. Descriptive variables

Type	Symbol	Measuring unit	Definition of variable
Dependent variable	y	g/L	Hemoglobin (hp %): the oxygen-carrying pigment and predominant protein in the red blood cells.
Independent variable	X1	Year	Real age: represents the age of the patient.
	X2	10^6/ml	Complete Blood Count (C.B.C): is a blood test used to evaluate the overall health and detect a wide range of disorders, including anemia, infection and leukemia.
	X3	%	Mean Corpuscular Volume (MCV): mean cell volume (MCV) is a measure of the average volume of the red blood corpuscles (or red blood cells).
	X4	10^3/ml	Platelet Count: is a lab test to measure how many platelets there are in the blood. Platelets are parts of the blood that help the blood clot. They are smaller than red or white blood cells.
	X5	Centimetre	Size of Spleen: the altered spleen size is changed in some disorders, including infestation genetic anemia (Thalassemia). The size of the spleen can be extremely variable ranging from 7 cm to 14 cm, otherwise it is indicative of a satisfactory condition.
	X6	1-A	Blood type: 1 if blood type is A 0=otherwise
	X7	2-B	Blood type: 1 if blood type is B 0=otherwise
	X8	3-AB	Blood type: 1 if blood type is AB 0=otherwise
	X9	4-O	Blood type: 1 if blood type is 0 0=otherwise
	X10	10^3/ml	White Blood Cell (WBC): the number of leukocytes in the blood is often an indicator of disease, and thus the WBC count is an important indicator of the occurrence of the disease.
	X11	1=Yes 0=No	The father relatives with mother: is mean the patients have inbreeding between father and mother. (i.e. people with common grandparents or people who share other fairly recent ancestors)

Table 2. Descriptive statistics

Variable	Count	Mean	Deviation	Variance
X1	152	6.436	3.293	10.845
X2	152	17948.485	29427.270	865964220.251
X3	152	19.370	6.344	41.580
X4	152	19713.418	190275.975	36204946736.025
X5	152	4.556	2.473	5.342
X6	152	0.0724	0.25995	0.068
X7	152	0.1382	0.34621	0.120
X8	152	0.1250	0.33181	0.110
X9	0	0	0	0
X10	152	8543.818	1309.257	1714154.235
X11	152	0.503	0.50151	0.252
Y	152	6.425	1.925	3.990

The Table 2. above shows the descriptive statistics of the variables of the model, in order to describe the nature of the variables under study. The following is an analytical presentation of these measures for each variable of the model. Show that the mean age patients in the sample was 6.43 years with the standard deviation 3.293 years, while the variance is 10.845 years. The results also show that the mean value for the complete blood count (C.B.C) was 17984.485(10^6/ml) and a deviation 29427.270(10^6/ml), while the variance 865964220.251(10^6/ml).

The mean corpuscular volume (M.C.V) is 19.370%, while the deviation was 6.344% and the variance 41.580%. The results also show that the mean value for the mean corpuscular volume (M.C.V) was 19713.418(10^3/ml) and a deviation 190275.975(10^3/ml). Mean size of spleen in the sample was 4.556c.m, with the deviation 2.473c.m, while the variance 5.342c.m. The results of the descriptive statistics show that the mean of the white blood cell (W.B.C) was 8543.818(10^3/ml) and a deviation was 1309.257(10^3/ml).

Finally, the mean the fathers' relatives with mothers was 0.503 that is mean about half of the patients have inbreeding between father and mother.

5.1. Testing the Multicollinearity Problems

We calculated the Chi square test (x^2). To apply this test, these steps were followed:

a. test hypothesis:

H_0=X_J *nonexistence multicollinearity problem*

H_1=X_J *existence multicollinearity problem*

b. calculate the Chi- square test (x^2):

$$x^2_{Calculate} = -\left[152 - 1 - \frac{1}{6}\left(2(8)+5\right)\right]\log(0.4298) \quad (17)$$

$x^2_{Calculate}$=(54.0735).

By comparing the calculated value to the tabular value, it is evident that it is greater, leading to rejection of null hypothesis and acceptance of alternative hypothesis, thus, proving that the model suffers from multicollinearity problem. Therefore, the parameters of the quantile regression model cannot be estimated directly, because the estimation will be inaccurate. First, the issue must be treated by using the ridge regression method.

5.2. Determine the Causative Variables of the Multicollinearity Problems

In order to determine the variables causing the problem of linear multiplicity the Variance Inflation Factor (VIF) will be used, which measures the inflation of the parameter estimates for all explanatory variables in the model. These indicators were computed for the regression parameters of all the explanatory variables of the model. The multicollinearity between the explanatory variables was determined with the following results:

Table 3. Variance Inflation Factors (VIF)

Variable	VIF
X1	39.4220
X2	14.2726
X3	11.2650
X4	37.0707
X5	1.2230
X6	2.2144
X7	9.6127
X8	10.6360
X9	0.0000
X10	1.3575
X11	1.1972

We notice from Table 3 that the values of the VIF for some of the explanatory variables (X1, X2, X3, X4, X8) are greater than 10, this means on the presence of the multicollinearity problem between explanatory variables.

5.3. Ridge Quantile Regression Analysis

This method allows the estimation of the quantile regression models established by ridge coefficients. These coefficients are selected based on the Bayesian method for parameter (k) estimation. This method was used to find the best value of the ridge parameter in accordance with formula (15). Using R package the following result was reached: k = 0.02578. Table 4 shows the results for each quantile coefficients.

The results presented in Table 4. for the coefficient estimation at the quantile levels 0.25, 0.50, 0.75 and 0.95 show the following :

a. At the quantile level (0.25): the variables X1 and X3 have a significant effect on the dependent variable according to p-value and R-square (0.3958); this means the independent variables can explain 39.5% of the variation in the dependent variable. This quantile level is weak in interpreting the data of the phenomenon under study. The VIF values for X1 and X3 are 1.1316 and 1.1354, respectively. That is an indicator that the multicollinearity problem is solved.
b. The results were quantile level 0.50 showed that the

variables X1, X3, X10 have a significant effect on the dependent variable. The R-square value (0.7121) indicates that the independent variables can explain 71.2% of the variation in the dependent variable. This level is also weak in interpreting the data of the phenomenon under study. The VIF values for X1, X3, X10 are 1.2870, 1.1323 and 1.1383, respectively. This is an indicator that the multicollinearity problem is solved.

c. The results at the quantile level 0.75, indicate that the variables X1, X2, X4, X5 and X11 have a significant effect on the dependent variable (Incidence of anemia in children, Thalassemia). The R-square (0.8567) shows that the independent variables, can explain 85.6% of the variation in the dependent variable. This quantile level is strong and the variables have a direct correlation with the response variable except three variables (X3, X6, X7, X8 and X10). These variables have weak effect according to this measure, all the results are logical by the medical side. The VIF values for X1, X2, X4, X5 and X8 are 1.1258, 1.0718, 1.057, 1.1451, and 1.0207. This indicates that the multicollinearity problem is solved.

d. At quantile level 0.95, the variables X1, X5, X8 and X10 have a significant effect on the adopted variable. The value of R-square (0.6952) shows that the independent variables can explain 69.5% of the variation in the independent variable. This ratio is small and can't be used to interpret the results.

Finally, the last step of our analysis is to compare the performance of the quantile levels in ridge quantile regression. This comparison would help to better select the best quantile level. Through comparison we found that the quantile level at 0.75 is the best as it has a high R-square (0.8567) according to the interpretation of the data regarding the phenomenon under study.

Table 4. Ridge quantile regression analysis

Quantile levels	Independent Variable	Regression Coefficient	Standard Error	T-test	p - value	VIF
Q (0.25)	Intercept	3.7589				
	X1	0.7347	0.2584	2.8429	0.0051	1.316
	X2	0.0434	0.1568	0.2769	0.7823	1.0768
	X3	1.0090	0.1396	7.2277	0.0000	1.1354
	X4	0.0284	0.2504	0.1134	0.9098	1.0639
	X5	0.0794	0.0453	1.7523	0.0819	1.1527
	X6	0.0552	0.0420	1.3145	0.1889	1.5999
	X7	0.0890	0.0478	1.8599	0.0631	1.0941
	X8	0.0211	0.04251	0.4982	0.6184	1.9258
	X10	0.0710	0.0478	1.4839	0.1400	1.4425
	X11	0.0148	0.0429	0.3453	0.7304	1.0251
	R-square	0.3958				
Q (0.50)	Intercept	17.3574				
	X1	0.9296	0.2518	3.6913	0.0003	1.287
	X2	0.1371	0.1528	0.8972	0.3711	1.0743
	X3	0.0999	0.0441	2.2631	0.0251	1.1323
	X4	0.0423	0.2440	0.1735	0.8625	1.0605
	X5	0.01330	0.1359	0.09782	0.9222	1.1488
	X6	0.0676	0.0407	1.660	0.0971	1.5264
	X7	0.0587	0.0463	1.2660	0.2057	1.8433
	X8	0.0235	0.04119	0.5726	0.5670	1.6963
	X10	0.0803	0.0366	2.1939	0.01334	1.1383
	X11	0.0178	0.0418	0.4267	0.6702	1.0229
	R-square	0.7121				
Q (0.75)	Intercept	18.2781				
	X1	1.0199	0.2557	3.9892	0.0001	1.1258
	X2	0.4231	0.1551	2.7279	0.0325	1.0718
	X3	0.0792	0.1381	0.5735	0.5672	1.1293
	X4	0.5801	0.2477	2.1400	0.03971	1.0572
	X5	0.0971	0.0448	2.1671	0.0319	1.1451
	X6	0.0747	0.0411	1.813	6.9965	1.4585
	X7	0.0536	0.0469	1.1443	2.5268	1.6415
	X8	0.0488	0.0416	1.1546	2.4844	1.5098
	X10	0.0644	0.0473	1.3604	0.1758	1.1342
	X11	0.0653	0.0325	2.009	0.0421	1.0207
	R-square	0.8567				
Q (0.95)	Intercept	16.5410				
	X1	0.7482	0.2620	2.8562	0.0049	1.1229
	X2	0.0223	0.1589	0.1404	0.8885	1.0694
	X3	0.0179	0.1415	0.1263	0.8997	1.1264
	X4	0.5191	0.2538	0.20453	0.03845	1.0540
	X5	0.0949	0.0459	2.0675	0.00562	1.1413
	X6	0.07686	0.0439	1.748	0.0806	1.3954
	X7	0.0952	0.0500	1.9030	0.0572	1.4763
	X8	0.1202	0.0444	2.7058	0.0069	0.6732
	X10	0.1070	0.0485	2.2074	0.0289	1.1302
	X11	0.0378	0.0435	0.8693	0.3862	1.0186
	R-square	0.6952				

6. Conclusions

In this paper, the multicollinearity issues in quantile regression models were the subject under research, in an attempt to find practical solutions to deal with the violation issue of a regression model assumption. The solution adopted in our research is the ridge regression and for estimating the parameter of the ridge we use the Bayesian approach. We tested this approach for identifying the factors affecting anemia in children (Thalassemia) across various quantile ratios.

The study showed that the use of the ridge quantile regression method in the cases when the independent variables are affected by multicollinearity is one of the successful ways to solve this issue. Therefore, applying the ridge quantile regression method in other studies is

recommended, since it provides better estimators than the ordinary regression methods when the independent variables are related, without omitting any of the independent variables.

By applying the ridge regression method at quantile level 0.75, we found that there were five variables with a significant impact on the anemia in children (Thalassemia) in Iraq at the statistically significant level less for 0.05%: real age, complete blood count, platelet count, size of spleen, the relatives between father and mother, where was the p-value as (0.0001, 0.0325, 0.03971, 0.0319, 0.0421) respectively. As for the rest of the variables (mean corpuscular volume, Blood type, white blood cell) the study shows that they are weak and have no significant statistical effect.

The results are explained by the fact that Iraq is one of developing countries still facing health, environmental and economic difficulties.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the contribution of all the physicians that helped collecting and analysing the data for this study.

REFERENCES

[1] R. Najdecki, I. Georgiou, and D. Lolis, "The thalassemia syndromes and pregnancy, molecular basis, clinical aspects, prenatal diagnosis," Ginekologia Polska, vol. 69, no. 8, pp. 664-668, 1998.

[2] "Obesity and Overweight factsheet from the WHO. Health." World Health Organization (WHO), 2017.

[3] S. T. Miller, E. A. Macklin, C. H. Pegelow, T. R. Kinney, L. A. Sleeper, J. A. Bello, and K. Ohene-Frempong, "Silent infarction as a risk factor for overt stroke in children with sickle cell anemia: a report from the Cooperative Study of Sickle Cell Disease," The Journal of Pediatrics, vol. 139, no. 3, pp. 385-390, 2001.

[4] B. D. Benoist, E. McLean, I. Egll, and M. Cogswell, Worldwide prevalence of anaemia 1993-2005: WHO global database on anaemia. Worldwide prevalence of anaemia 1993-2005: WHO global database on anaemia, 2008.

[5] G. Bassett Jr., and R. Koenker, "Asymptotic theory of least absolute error regression," Journal of the American Statistical Association, vol. 73, no. 363, pp. 618-622, 1978.

[6] R. Koenker, and G. Bassett Jr., "Regression quantiles," Econometrica: Journal of the Econometric Society, pp. 33-50, 1978.

[7] R. Koenker, Quantile Regression (No. 38), Cambridge University Press, 2005.

[8] B. S. Cade, and B. R. Noon, "A gentle introduction to quantile regression for ecologists," Frontiers in Ecology and the Environment, vol. 1, no. 8, pp. 412-420, 2003.

[9] R. Alhamzawi, and K. Yu, "Bayesian Tobit quantile regression using g-prior distribution with ridge parameter," Journal of Statistical Computation and Simulation, vol. 85, no. 14, pp. 2903-2918, 2015.

[10] A. E. Hoerl, R. W. Kannard, and K. F. Baldwin, "Ridge regression: some simulations," Communications in Statistics Theory and Methods, vol. 4, no. 2, pp. 105-123, 1975.

[11] P. Congdon, Multilevel and Panel Data Models. Bayesian Statistical Modelling, 2nd ed., pp. 367-424, 2007.

[12] E. Cule, and M. De Iorio, "Ridge regression in prediction problems: automatic choice of the ridge parameter," Genetic Epidemiology, vol. 37, no. 7, pp. 704-714, 2013.

[13] A. Fitrianto, and L. C. Yik, "Performance of ridge regression estimator method on small sample size by varying correlation coefficients: a simulation study," Journal of Mathematics and Statistics, vol. 10, no. 1, pp. 25–29, 2014.

[14] A. Bager, M. Roman, M. Algedih, and B. Mohammed, "Addressing multicollinearity in regression models: a ridge regression application," Journal of Social and Economic Statistics, vol. 6, no. 1, 2017.

[15] A. E. Hoerl, and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55-67, 1970.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267-288, 1996.

[17] A. S. M. B. M. Harbi, and O. B. K. Mohammed, "using approach quantile regression to determine the factors affecting measuring capacity in Iraq," American Review of Mathematics and Statistics, vol. 5, no. 1, pp. 35-44, 2017.

[18] M. H. Odah, A. S. M. Bager, and B. K. Mohammed, "tobit regression analysis applied on Iraqi bank loans," American Journal of Mathematics and Statistics, vol. 7, no. 4, pp. 179-182, 2017.

[19] D. V. Lindley and A. F.Smith, "Bayes estimates for the linear model", Journal of the Royal Statistical Society. Series B (Methodological), 1-41, 1972.

[20] T. Kubokawa, and M. S. Srivastava, "Improved empirical Bayes ridge regression estimators under multicollinearity," Communications in Statistics-Theory and Methods, vol. 33, no. 8, pp. 1943-1973, 2004.

[21] A. P. Dempster, "Alternatives to least squares in multiple regression," Multivariate Statistical Inference, pp. 25-40, 1973.

[22] K. H. Loesgen, "A generalization and Bayesian interpretation of ridge-type estimators with good prior means," Statistical Papers, vol. 31, no. 1, pp. 147-154, 1990.

[23] S. Sclove, Least Squares with Random Regression Coefficient. Technical Report, Department of Economics, Stanford University, 1973.

[24] M. H. Odah, A. S. M. Bager, and B. K. Mohammed, "Studying the determinants of divortiality in Iraq. a two-stage estimation model with Tobit regression," International Journal of Applied Mathematics & Statistical Sciences, vol. 7, no. 2, pp. 45-54, 2018.