Vol.9.Issue.4.2021 (Oct-Dec)

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

COMPARISON OF TREATING THE PROBLEM OF METHOD IN MULTICOLLINRARITY WITH APPLICATION ON THE STUDENTS OF THE SHIA RELIGIOUS SCHOOLS IN NAJAF BETWEEN THE METHOD OF PARTIAL LEAST SQUARES AND THE MAIN MATRIXES

Hasan Mahdi Abbas Al Qaseer[1], **Assist. Prof Dr. Fadel Hamid Hadi Al-Husseini**[2]

[1]Al-Qadisiyah University, College of Administration and Economics,
Department of Statistics, Qadisiyah, Iraq

[2]Supervisor, Al-Qadisiyah University, College of Administration and Economics
Department of Statistics, Qadisiyah, Iraq

Email: stat.post21@qu.edu.iq

DOI:10.33329/bomsr.9.4.30

ABSTRACT

In this study, we compare between pc and Partial Least Squares, which is considered as one of important methods of multicollinearity MSE criteria is used for choosing the best estimation Method. Multicollinearity is considered as one of the serions problem in regression analysis, so in this paper we give a big accuracy intention to this problem and to find its solutions

Keywords: multicollinrarity. Method of partial least squares. Main matrixes. Shia religious schools

Theoretical Study

Regression analysis is a widely used statistical method Its main goal is to constantly seek to know and explain relationships different between phenomena by identifying the relationships between the variables, which Explains the relationship between one variable called the dependent variable (variable response) and another variable It causes variance in the dependent variable and accepts the (variables predictive) It also describes the relationship between the dependent variable and the explanatory variables in the form of a mathematical model .The accuracy of this mathematical

model depends on the assumptions of the analysis, as some .These assumptions are related to the functional relationship, while others are related to the variable .These assumptions are several, the most important of which is the problem of multi-linearity, where Numerous research and studies have confirmed the great interest in the problem of polygamy Linearity and finding a negative solution to it, the most important of which are the main matrixes that return To the year 1901 when Karl Person. As a way to solve .The problem of multiple linear relationship and adopted an exploratory method can be used From them to reach the progress and understand the interrelationship between the variables Principal matrixes analysis is a method that aims to find factors are linear combinations called a few principal matrixes that are derived from . The original variables to be replaced so that they are eligible to explain most .The total variance of the original values and these principal matrixes are orthogonal .That is, there is no connection between them, and the main matrixes can be written as follows:

$$\mathrm{pc_i} = \mathrm{a_{1i}x_1} + \mathrm{a_{2i}x_2} + \cdots + \mathrm{a_{mn}x_m}$$

$$\mathrm{pc_i} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{a_{ji}\, x_j}\ (\mathrm{i, j} = 1{,}2 \ldots . \mathrm{m})$$

whereas

pc_i: represents the main matrix aij: represents the coefficient of j in the principal matrix i which represents the values of the characteristic vectors aij accompanying the characteristic roots λ_j for the matrix used. Using the matrix method, $Pc = xA$ And the matrix A in The above formula has columns is used to represent the characteristic vectors accompanying the matrix Used and arranged according to root amounts Characteristic λ_m……. $\lambda_2\rangle\ \lambda_1\rangle$. Each column of the matrix is A. It represents one of the main matrixes and the idea of calculating the matrixes depends .The main on the properties of the distinctive roots and the accompanying vectors .The characteristic of the correlation matrix or the matrix of covariance and covariance x_s. According to its units of measurement, it is the same as the original variables or different as (s) represents all the studied variables from x_1, x_2, …….x_n, then if the xs variables have the same units of measurement in these aji values of the characteristic vectors Status The coefficients are the main matrixes of the matrix of covariance and covariance v for the xs variables represented by:

$$v = \begin{Bmatrix} v_{11} & v_{12} & \ldots v_{1m} \\ v_{21} & v_{22} & \ldots v_{2m} \\ v_{m1} & v_{m2} & \ldots v_{mm} \end{Bmatrix}$$

And to find the distinctive roots λ We subtract from the diagonal values of the matrix v Then we set the value of its modulus to zero, and we get the characteristic equation for the matrix v

$$|v - \lambda| I = 0$$

And the form of the characteristic equation is a (polynomial) in (λ) class m

$$\lambda^{\mathfrak{m}} + c_{m-1} - \lambda^{m-1} + \cdots \ldots \ldots C_1 \lambda C_0 = 0$$

When solving this equation we get m from the roots $\lambda_1\ \lambda_2$ ….. λ_m.These roots are arranged so that λ_1, It is the largest value followed by λ_2

$\lambda_1 > \lambda_2$ .....$> \lambda_m$ And that each root is distinct λ_j The value of a j discriminant overlap corresponds to and find these vectors use the equation

$$|v - \lambda I|a = 0$$

We choose aji so that the values of the scalar discriminant vector are equa a'a = 1 l. Then we enter a ji as an equation for the variables x j for the main matrixes. When the variables xs have different units of measure, it is recommended to convert x into standard variables that have a mean = zero and variance = 1 before finding a *ji.* The main matrixes that have a significant effect are selected by testing the cumulative ratio of the explained variance for each matrix as al rawyi in 1987 mentioned that the number of the main matrixes chosen is Distinctive roots greater than oneλ>1 As for the partial least squares (PLS) method, it has been shown in many studies It includes chemistry and social sciences, and the researcher (1966; wold) is the first to find this method and it has become popular in medicine, especially in clinical treatments, where the number of observations is low, the number of patients with Too many variables(x's) Represented by Agrarat Mart number Supported variants(y's) This method was developed by Friedman (1993; Friedman) The partial least squares regression method is based on the covariance and covariance matrix cov(x,y). As this method is called determination of factors by latent variables, which in turn is the best model for the dependent variable (y's) It can be described as follows: Let us have a rounded matrix of predictive variables (x's) multiplied by a random vector of U Which that:

$$X'U = \begin{cases} \sum x_{i1}u_i \\ \sum x_{i2}u_i \\ \sum x_{ip}u_i \end{cases}$$

Through which the matrix can be obtained β Which shows the type and form of the relationship between each dependent variable and each predictive variable as follows

	y_1	y_2	y_j
x_1 x_2	$\beta 11$	$\beta 12$	$\beta 1$ j
⋮	$\beta 21$	$\beta 22$	$\beta 2$ j
⋮	⋮		
x_k	βk_1	βk_2	βk_j

The partial least squares method is a linear combination of squares .The smallest matrix of correlation and covariance among predictive variables .And the dependent variables, and the part on which the squares method depends Partial least in the correlation and variance matrix is the intercept port Block Cross That is, the correlations between the predictive variables and the dependent variables. This method also provides scores factors in the form of linear sets between the original predictive variables, so it will not be There are correlations between the factors of the variables used in the model Regression prediction can explain the work of the partial least squares method. In terms of calculating the regression coefficients and matrix of occult variables T And the loaded columns P. As an initial step, the matrix of variables is transformed Predictive X and Y dependent variables matrix to standard form So that F, E, respectively, and the steps of this algorithm are as follows (Abdi,2003)

Taking random initial values for the vector U

find the vector W through the relationship W = E′U

Finding the vector t, which is one of the columns of the matrix T for the variables From the matrix X, told = Ew

Convert the vector t to the normalized form by multiplying by $1/\|t\|$

tnew =$1/\|t\|$.t old

$\|t\|$=$\sqrt{t_1^2 + t_2^2 + t_3^2 + \cdots + t_m^2}$

Find the weighted vector c of the Y matrix. $c = F't$

Find the vector U which is one of the columns of the matrix U of the variables extracted for matrix y

$$U_{\text{old}} = FC$$

Convert the vector U to the normalized form by multiplying by$1/\|u\|$

U_{new} =$1/\|t\|U_{\text{old}}$

$\|u\|$=$\sqrt{u_1^2 + u_2^2 + u_3^2 + \cdots + u_m^2}$

Compare the values of t new in step (4) with the values of U new in step (7). If the values are close, then the values of . are calculated β In the event of a lack of convergence, Calculation of the value of the bearing P for matrix X using the following equation

$$P = E't$$

Then the two new matrices F_1,E_1 are calculated by subtracting the vector t From each of the matrices F, E, as follows:-

$$E_1 = E - tp'$$

$$F_1 = F - btc'$$

Substituting the value of P into the equation $P = E't$

Substituting the value of c into the equation $C = F'$t

$$E_1 = E - t(E't)'$$

$$= E - ttE$$

$$F = F - bt(F't)'$$

$$= F - btt'F$$

After that, the algorithm is repeated again with the same steps, but using The two new matrices E_1, F_1 with the knowledge that in each iteration is done Compare the values of the vector (t) for the last cycle with the values of the vector (t) for the previous cycle. The process of iteration continues until

all columns are obtained matrix (T) whose number is equal to the rank of the matrix (x), and then the matrix of parameters is found β and as follows

$\beta = E'U\ (TEEU\ )-1T\ 'F$

Application side: The data in this research was obtained, as it was represented by four basic elements in the religious schools in the Shiite seminary in Najaf. number of study hours. Teaching methods. Case Study . The lectures, which represent independent variables, each variable contains (93) observations and a dependent variable y as follows:

$y_i = B_0 + B_1\ x1 + B_2\ x_2 + B_3\ x_3 + B_4\ x_4 + e_i$

As shown in the following table:

Table (1)

X_1	number of study hours
X_2	Teaching methods
X_3	Case Study
X_4	The lectures

Y : quality specialty ei : Random and distributive error Medium Natural (5) contrast σ_2 Perform a general regression analysis between the dependent variable y With the rest of the four influencing variables using Ready programming v15 minitab The results of the analysis were as follows:

$y_i = 74.0 + 1.41x_1 + 0.390x_2 - 0.031x_3 - 0.266x_4$

Table 2: values of the regression coefficients B, VIF, and the statistical laboratory T

predictor	coef	sEcoef	T	P	VIF
constant	73.97	72.39	1.02	0.337	
X1	1.4111	0.7695	1.83	0.104	38.5
X2	0.3896	0.7478	0.52	0.616	254.4
X3	-0.0314	0.7797	-0.04	0.969	46.9
X4	-0.2656	0.7326	-0.36	0.726	282.5

S=2.527 R-sq=98.1 % R-sq (adj)=97.2%

To confirm the existence of a linear overlap between the predictive variables, . was used I sold ways to check it out, including:

1-1 Using the Inflation Variance(VIF)

From observing the values of the variance inflating factors in the above table, we find that they are very high. Marquardt (1970) found that there is a multiplicity of linear relationship between the predictive variables in the event that the values of the factors include the variance VIF greater than

(4) or (10) and through the previous analysis it is noted that these very values. This indicates that there is a large linear overlap between the predictive variables.

1.2 Find the determinant of the matrix $|x'x|$

This criterion was suggested by Webster & Mason (1975). It states that if 0 = $|x'x|$ This indicates a complete overlap Among the predictive variables, whether or not 1 = $|x'x|$This indicates independence Predictive variables among themselves either as the value of $|x'x|$ Between zero and one, this indicates that there is a near perfect linear overlap, and through the eigenvalues of the matrix ($x'x$) , the determinant of the matrix ($x'x$) was calculated, which was equal to

$$|x'x| = \prod_i^p = iL_j = 0.000014$$

These values are very small t and this is another evidence of the existence and close to zero Large linear overlap between the predictive variables.

1.3 Using the off-diagonal elements of the correlation matrix :

This method was proposed by Mason & Gunst (1980, Mason & Gunst) whereby through the correlation matrix ($x'x$) And explained below:

$$x'x = \begin{Bmatrix} 0.000 & 0.229 & 0.824 & 0.245 \\ 0.229 & 1.000 & 0.139 & 0.973 \\ -0.824 & 0.139 & 1.000 & 0.030 \\ -0.245 & 0.97 & 30.030 & 1.000 \end{Bmatrix}$$

We note that there is a perfect correlation between all the variables, which indicates the presence of overlap .There is a large linearity between these variables, which leads to the inaccuracy of the results of the analysis .We also note another evidence of the overlap between the predictive variables is the value of R^2 Represents (the coefficient of determination) where it appeared very large.

2. Method of the main components

In order to address the problem of multilinearity, the components method was used Principal based on the covariance and covariance matrix to calculate . The coefficients of the main components aij are as follows:

$$v = \begin{Bmatrix} 34.603 & 20.923 & 31.051 & 24.167 \\ 20.923 & 242.141 & 13.878 & 253.417 \\ -31.051 & 13.878 & 41.026 & 3.167 \\ -241.167 & 253.417 & 3.167 & 280.167 \end{Bmatrix}$$

Then we find the characteristic roots:

$\lambda_1 = 517.80$

$\lambda_2 = 67.50$

$\lambda_3 = 12.41$

$\lambda_4 = 0.24$

From the characteristic roots we calculate the characteristic vector matrix:

$$\begin{Bmatrix} -0.068 & 0.646 & 0.567 & 0.506 \\ -0.679 & 0.020 & 0.544 & 0.493 \\ 0.029 & 0.755 & 0.404 & 0.516 \\ 0.731 & 0.108 & 0.468 & 0.484 \end{Bmatrix}$$

$pc_1 = -0.068x_1 - 0.679x_2 + 0.021x_3 + 0.731x_4$

$pc_2 = 0.646x_1 + 0.020x_2 - 0.755x_3 + 0.108x_4$

$pc_3 = -0.567x_1 + 0.544x_2 - 0.404x_3 + 0.468x_4$

$pc_4 = 0.506x_1 + 0.493x_2 + 0.516x_3 + 0.484x_4$

The main components are:

Where we choose the distinct roots greater than one, which are the first roots .The second and third (then we calculate the multiple regression equation, which is as follows:

$y = 1.02557 + 0.01393pc_1 - 0.42407\ pc_2 - 0.63449\ pc_3$

Now we substitute each of Pc_1, Pc_2, and Pc_3 into the regression equation the above, we get the following regression equation:

$y = 1.02557 + 2.15701x_1 + 1.15367x_2 + 0.73568x_3 + 0.4823x_4$

The analysis of variance table for y is as follows:

Table 3: Analysis of variance for students of religious schools of the Hawza in Najaf by (PC) method

souree	DF	SS	MS	F	P
Regession Residnal Error	3	2664.18			
Total	9	52.74	888.06	151.54	0.000
	12	2716.92	5.86		

3- Partial Least Squares (Pls) Method

The second method in order to address the problem of multilinearity was used Partial Least Squares (pls) method (shown on the side Theoretical results were as follows:

souree	D_F	SS	MS	F	P
Regession	4	26665.83	666.458	104.3	
Residnal Error Total	8	51.09	6.386	6	0.000
	12	2716.92			

In order to identify the significance of the proposed linear relationship and test the extent of . The effect of explanatory variables on the dependent variable The hypothesis was tested .The special regression model is as follows:

$H_0 : B_1 = B_2 = B_3 = B_4$

$H_1 : B_1 \neq B_2 \neq B_3 \neq B_4$

It is clear from the results of Table (3) that the null hypothesis was rejected and the hypothesis accepted (α)1% This means that there are differences Significant among explanatory variables on the dependent variable The problem of multiple linear relationships is not present, as is the case with the results of Table (4). Rejection of the null hypothesis and acceptance of the alternative hypothesis at a significant (α)1% level That is, there is no problem of multiple linear relationships

Conclusions

It has been shown the importance of studying the problem of multiple linear relationship and its impact on the results regression analysis

1. The analysis hypotheses were tested for the data and we concluded through the tests that the data suffers from the problem of multilinearity, as it was revealed according to the Inflation of Variance Test (VIF) and the correlation matrix, where a complete correlation appeared between the variables, as well as through the correlation coefficient, where a very large value appeared
2. In our research, the problem of multiple linear relationship has been addressed in two ways: the main components (PC) partial least squares (PLS), where it was found through the statistical analysis MSE mean error squares that the method of the main components is more efficient than the method of partial least squares Owns the lowest value of MSE
3. As it was shown through the statistical analyzes and comparison with the calculated F-value with the tabular F, that there are significant differences between the explanatory variables on the dependent variable, and just as the study showed that there is no problem of linear multiplicity

References

[1]. Marquardt , D . W . (1970) , “Generalized Inverse, Ridge Regression , Biased Linear Estimation and Nonlinear Estimation”, *Technometrics* , Vol . 12 , pp (591-612) .

[2]. Mason, R.L., Gunst, R.F. and Webster, I.T., (1975), "Regression Analysis and problems of Multicollinearity", C*omm. In statistics* .VOL.(4) No.(3) pp(277-292).

[3]. Saikat Maitra and Jun yan , (2008) , “Principle component Analysis and Partial Least squares : Two Dimension Reduction Techiques for Regression” *Casualty Actuarial society* , pp (79-90).

[4]. Wold , H . (1966) “La Regression PLS . Paris : Technip Iterative Least squares” , In P . R . Kishnaiaah (Ed) . Multivairte Analysis , New york, Academic Press . PP (391 -420).

[5]. Al-Rawi, Khasha Mahmoud, 1987, “The Introduction to Regression Analysis”, Dar Al-Kutub for Printing and Publishing, University of Mosul.

[6]. Abdi , H . (2003). Partal least squares (pls) Regression .In Lewis – Beck , M . , A . Bryman , & T. Futing , editors , Encyclopedia of social oaks : Sage

[7]. Draper , N . R . and smtth , H . (1966) , “Applied Regression Analysis), University of Wisconsin and Rensselaer polytechnic Institute” .

[8]. Frank , I . E . and friedman , J . H . (1993) . “Astatistical view of chemometrcs Regression Tools” Technometrics , 35 , pp (109-148)