Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 24, Issue 1 (2022) ISSN$_{Print}$: 1816-9171

The Multilinearity With Principal Component Regression

Fadel Hamid Hadi Al Husseini
Hassan Mahdi Abbas Al Qaseer
Al-Qadisiyah University - College of Administration and Economics
Corresponding Author: Hassan Mahdi Abbas Al Qaseer

***Abstract* :** This study focus to building a regression model without economic problems, the our model is focus on studying the time period for obtaining a degree of Ijtihad in religious sciences. The results showed through the variables involved in building The model under study, we find that there is a very large correlation between the explanatory variables, and thus this interference leads to the emergence of high correlations between the explanatory variables, and therefore the model will have a weak predictive efficiency, due to the exaggeration of the variance of the estimated parameters, and to overcome this problem a model was employed Regression Principal Components ..
Keywords Multilinearity, Effective principal , principal regression model.

INTRODUCTION: The analysis of the regression model focuses on describing the relationship between the dependent variable and a set of explanatory variables (Yan and Gang, 2009), but most of the studied regression models suffer from a set of standard problems, and among these problems is the multicollinearity that arises when there is a strong relationship between Two or more explanatory variables. And one of the important conditions that must be met in the regression model is the rank condition (intriligator, 1996).

Rank (X)=p

Whereas, X: a matrix of order ($(n * p)$ for the observations of the independent variables, and therefore in the event that the explanatory variables are (independent variable) linearly independent.

It is possible to find the inverse of the information matrix $(X^t X)$ and thus it is possible to estimate the estimations of the parameters of the regression model and the estimations of this model are efficient, with a high ability to generalize to the studied population and with an accurate predictive ability, but if it is a perfect relationship between two or more explanatory variables, then This will result in a violation of the rank condition i.e. that

I. $rank(X) < p$

Therefore, it is not possible to find the inverse of the information matrix $(X^t X)$,therefore there is no possibility to estimate the parameters of the model, and this case is called perfect multilinearity.

(perfect multicollinearity) But if the determinant of the information matrix is not equal to zero, but rather close to it (that is, $|X^t X| \approx 0$, where this situation arises, when the variables tend to move together, increasing or decreasing, or in the case of using time-returned variables (lagged variable) in this case it is possible to estimate the parameters of the model, but these estimates are inaccurate and not representative of the reality of the problem being studied, because the variance of the estimated parameters will be very large and therefore will give us inaccurate information about the significance of the (t) test. But building the model is unable to show the effect of these variables separately due to the overlap of these variables among themselves. It is known that a regression model is similar to the model that studies the relationship between a dependent variable and a set of explanatory variables. Therefore, the regression model is subject to standard problems, one of these problems is the problem of multiplicity, and therefore the search for a good treatment method for this problem will be reflected in the accuracy of the estimators of the model parameter as well as the accuracy of the regression equation prediction. One of the classic methods used to treat this problem is to search for the variable or variables that caused the problem of multilinearity to appear and delete it, but perhaps these variables that will be deleted are important variables in building the model and have explanatory weight in it. Each method has properties and advantages that distinguish it from the other. In the regression model that is exposed to the problem of multilinearity, we will employ the main compounds method to treat this problem, in order to avoid this problem that causes the variance of some parameters of the regression model to inflate and bias the parameters of this model, without decreasing any From the model information, (keep all the independent variables included in the studied regression model. This paper is divided into four section

II. HEADING S

1. **Abstract**
2. **Introduction**
3. **HEADING S**
4. **Indentations and EQUATIONS**
5. **Figures and Tables**

6. **Conclusion**
7. **Acknowledgements**
8. **REFERENCES**

III. INDENTATIONS AND EQUATIONS

We will transform the related independent variables into new unrelated variables, representing the main components and estimating its features using the methods of estimating the parameters of the regression model and taking advantage of these capabilities in estimating the parameters of the original regression model for the original variables (Mardia et al 1979).

$$Y_i = X_i\beta + U, \;\; i = 1, \dots\dots, n \qquad (1)$$

where Y_i is the dependent variable and U is the limit of the random error that is distributed according to the normal distribution with an arithmetic mean equal to zero and a constant variance σ^2 i.e. $U \sim N(0, \sigma^2)$ and x_i^T is the set of explanatory variables to be estimated The regression model shown in (1) is affected by economic problems, and one of these problems is the problem of multicollinearity, and one of the methods proposed to treat this problem is the method of regression of the main compounds, which is a linear orthogonal combination of independent variables (X) known As follows

$$F = X\Lambda \qquad (2)$$

Whereas F represents the matrix of principal components of order $(n * q)$ either Λ is an orthogonal matrix of standard characteristic vectors corresponding to the characteristic roots of the information matrix (X^tX) and its order $(q * q)$ and its elements θ_ij (i=1,……). .n), and its columns Λ_j (j=1,2,…..q) and this matrix makes the information matrix a diagonal matrix, assuming that the characteristic values of the matrix $(X^tX\ )$ are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$$

The term is to reconcile the regression model shown in equation (1) on the principal component F, that is, an expression of the variable T_i as a function of the orthogonal principal components instead of the independent variables (X) that are interrelated with each other. Since the matrix Λ is orthogonal, that is,

$$(\Lambda\Lambda^t = 1\ )$$
$$(F = X\Lambda) * \Lambda^t$$
$$F\Lambda^t = X\Lambda\Lambda^t$$
$$X = F\Lambda^t \qquad (3)$$
$$Y_i = F\Lambda^t\beta + U,$$

Suppose $\Lambda^t\beta = \varphi$ We will find that the previous regression model will take the following form

$$Y_i = F\varphi + U, \qquad (4)$$

After estimating the corresponding feature vector, the main components included in the construction of the regression model, which are a linear combination of independent variables, and these main components (the new variables) are unrelated, that is, independent among themselves $cov(F_i, F_j) = 0$.

The estimated feature vector φˆ is distributed according to the normal distribution with mean φ and variance $\vartheta^{-1}\sigma^2$)where ϑ is a diagonal matrix of order(q*q) and its elements are the characteristic roots of the information matrix $(X^tX\ )$. The variance resulting from fitting the regression line with the main component can be calculated as follows: $\varphi^2\lambda_j$

Therefore, the percentage of variance in the values of the response variable (Y), which is explained by the principal component F_j , is

$$\left(\frac{\varphi^2\lambda_j}{\sum_{i=0}^{N} \quad {Y_i}^2}\right) * 100\% \qquad (5)$$

In order to estimate the vector of the parameters of the regression model $\hat{\beta}$ for the original variablesX_i , we can benefit from the relationship between the original parameters βˆand the parameters of the regression model φˆ resulting from the regression of the dependent variable (Y), according to the following

$$\Lambda^t\,\hat{\beta} = \hat{\varphi}$$

1=$\Lambda\Lambda^t$

$$\Lambda\Lambda^t\,\hat{\beta} = \Lambda\hat{\varphi}$$
$$\hat{\beta} = \Lambda\hat{\varphi} \qquad (6)$$

Since the parameter $\hat{\beta}$ will be distributed according to the normal distribution with mean) $\Lambda\hat{\varphi}$ and variance ($\vartheta^{-1}\Lambda\sigma^2$) that is,

$$\hat{\beta} \sim N(\Lambda\hat{\varphi}\,,\, \vartheta^{-1}\Lambda\sigma^2)$$

So the mean$\hat{\beta}$ is equal to $E(\hat{\beta}\,) = \sum_{j=1}^{q} \quad \theta_{ij}\varphi_j$ and the parameter variance $\hat{\beta}$ is equal to

$$var(\hat{\beta}) = \sigma^2 \sum_{j=1}^{q} \frac{{\theta_{ij}}^2}{\lambda_j} \qquad (7)$$

From the relationship (7) we also note that the variance of the parameters of the original regression model, estimated $\hat{\beta}$, also depends on the characteristic roots of the matrix $(X^t X)$, and therefore it is affected by the presence of small characteristic roots that amplify their variances. Through the definition of the main components, we find that The small characteristic roots that contribute to amplifying the variance The parameters of the regression model always correspond to the last principal components of the matrix $(X^t X)$. From pickling in this paper we will exclude the main compounds that correspond to the values of the characteristic roots less than (1).

According to the proposal of (Jeffers (1967), Chatterje.and price (1991.In the same regard, the researcher (Morison (1976)) mentioned that it is possible to rely on the main components that explain (75%) of the variance of the independent variables.

To clarify how to build a regression model for the dependent variable (Y) on the remaining major compounds after excluding some of them with one of the criteria mentioned above, we will assume that among the number (q) of the characteristic roots of the matrix $(X^t X)$ it turns out that there are (r) of them with values The number (q-r) of them is large, while the number (qr) of them has small values, so it is possible to exclude the number (q-s) of the main components F, and then adapt the Tobit regression model for the dependent variable (Y) to the remaining main components.

by orthogonality, the values of the estimators for the parameters (φ) will not differ whether all or part of the main components are used. Therefore, the parameters of the reduced regression model are estimated φ_r

$$\varphi_r = \vartheta^{-1}{}_r \, F^t{}_r \, Y \qquad (8)$$

where is a diagonal matrix i.e. $\vartheta_r = diag(\lambda_1, \lambda_2, ... , \lambda_r)$ where ϑ_r the feature vector (ϑ_r) is obtained by segmenting the matrix

$\Lambda = [\Lambda_r, \Lambda_{q-r}]$

$,\theta_{12} ,\theta_{13},\ldots\ldots \theta_{1r}.$ $\Lambda_1 = \theta_{11}$ $\Lambda_r = [\Lambda_1, \Lambda_2, \ldots \ldots \Lambda_r]$ and ,

In this case, the main compounds can be reduced as follows

$F = [F_r, F_{q-r}]$ Where the compounds that will be included in te analysis are

$F = [F_1, F_2, \ldots \ldots, F_r]$ Likewise, the domain of the main component features vector is determined by the components that will be included in the analysis. $[\varphi_r, \varphi_{q-r-1}]$= After obtaining the estimations of the feature vector φ_r can be φ_r used to estimate the parameters of the original model which can be expressed as follows

$$\hat{\hat{\beta}}_i = \Lambda_r \hat{\varphi}_r \qquad (9)$$

That is, the regression coefficients of the independent variables of the regression model can be obtained from the use of the regression coefficients of the remaining principal components, as follows

$$\hat{\hat{\beta}}_i = \sum_{j=1}^{r} \theta_{ij} \, \hat{\varphi}_j \qquad (10)$$

Equation (9) is a summary of the process of calculating the estimations of the parameters of the regression model after treatment from the autocorrelation of the main components method

In order to obtain the capabilities of the main component regression model, the R . language program will be employed.

Real data

The questionnaire was sent by e-mail to 40 mujtahids in the Hawza of Najaf, and their answers were received by e-mail due to the conditions of the Corona epidemic. Statistical operations were carried out and the results were as shown in the results section.

Al-Mujtahid is a title derived from Ijtihad. It is given to the seminary science student who, through his studies, reaches the stage of being able to elicit and derive legal rulings from religious texts. A religious science student in the seminary after completing a number of study levels: introductions, surfaces, and external research, which last for long years, may reach 25 Or 30 years, and so on until he reaches the point where he is able to elicit and derive the legal ruling from religious texts, and only then is he called the title of mujtahid or jurist, which is a title derived from jurisprudence, and it is called the person who studied and worked in religious sciences in general, and with the science of jurisprudence and its affiliates in particular. So that it reached the point of being able or able to elicit legal rulings from its religious texts and the ability to give fatwas, so it refers to the same meaning of a mujtahid, but in another wording. The term jurist is also shared by all Islamic sects.

To test the Multi collinearity problem ,there are many testes , in this study ,we will focus on

identified through using the Farrar-Glauber test
Multicollinearity Diagnosis . Methods
There are several proposed techniques to detect the problem of multilinearity, including the Variance Inflation Factor (VIF) method.
This test was presented by (Ferrar & Glauber m 1967) and VIF is one of the basic and broad methods for detecting the existence of the problem of multilinearity. Detecting the problem of polylinearity،
And equation (10) measures the so-called variance inflation factor: (Net, 2015, 58) (Obeid, 2017, 250-249(

Where $VIF = C_{jj} = \left(1 - \mathbf{R^2}_{\mathbf{j}}\right)^{-1}$ (10)

⟧ variance inflation factor for explanatory variable VIF_j
The coefficient of determining the regression model of the explanatory variable $\mathbf{R^2}_{\mathbf{j}}$
over the remaining explanatory variables p-1

The VIF inflation factor measures the effect of multilinearity among the explanatory variables in the regression model, which is always greater than or equal to one, but there is no limit value for VIF to determine the severity of the multilinearity effect on the model. The largest value of VIF is often used as an indicator of non-linearity. Desirable and often if it exceeds (10) it is considered an indication of the possibility of an unacceptable effect of high polylinearity on the estimations of the ordinary least squares.

IV. Figures and Tables

Table 1. Farrar-Glauber test results

${\chi^2}_{Calculated}$	${\chi^2}_{Tabulated}$
$\chi^2_C = 1461$	$\chi^2_T = 107.31$

find that the calculated value of χ^2_C is greater than the tabular values of χ^2_T . Through By comparing the calculated chi-square results with the tabular chi-square results, we this test, it can be judged that our model under study suffers from the problem of Multilinearity .

Chart 1. Eigen values and Eigen vector for principal component

	Principal component							
Variables	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*	*F8*
	Eigen value λ							
	3.899	3.001	2.586	1.37	1,610	1.494	1.346	1.051
	Reduction of variance %							
	19.497	15.003	12.931	8.684	80.51	7.472	6.742	9.253
	Collecting of variance %							
	19.497	34.500	47.430	56.114	64.165	71.637	78.379	83.633
	θi1	θi2	θi3	θi4	θi5	θi6	θi7	θi8
X1	-.666	.362	.361	-.016	87.500	87.500	-.054	.040
X2	.487	-.536	-.146	.412	100.000	100.000	.143	-.037
X3	.533	.104	-.043	.248	-.511	.007	-.210	-.035
X4	.152	.252	.084	.666	-.157	-.225	-.051	-.114
X5	.748	.353	.000	-.358	.132	-.042	.012	-.053
X6	-.581	-.385	.119	-.155	.065	-.253	.447	-.167
X7	.720	-.264	.015	-.034	.299	-.114	-.386	.107
X8	.229	.765	-.043	.075	.053	-.255	.365	-.097
X9	.240	.843	-.129	.142	-.019	-.128	.169	.294
X10	.417	.132	.380	-.113	.395	.182	.588	-.185
X11	.565	.220	.254	-.370	.406	-.048	-.108	.348
X12	-.363	.310	.067	.258	.112	.672	.050	.426
X13	.443	.285	.167	-.183	-.189	.498	-.279	-.481
X14	.193	-.492	.529	.240	-.019	.025	.018	.303
X15	.398	-.336	-.012	.475	.348	.489	.153	-.252
X16	.366	-.446	-.416	-.153	-.228	.169	.235	.418
X17	-.419	-.082	.035	-.286	.556	-.066	-.357	-.090
X18	.126	-.127	.909	-.084	-.264	-.077	.036	.039
X19	.126	-.127	.909	-.084	-.264	-.077	.036	.039
X20	-.260	.273	.351	.540	.437	-.267	-.258	.050

From the results in Chart 1., there are eight important principal components in this phenomenon and the extraction variance for these eight principal components is %83.733 from the total variance. Therefore, the rest of the principal components corresponding to Eigen value less than 1 (twelve components) will be excluded, as they caused increase of the variance coefficients of principal component regression model, because the principal components contain all the independent variables

Chart 2. Regression model on important principal components

Principal component	**Regression coefficient**	**t-value**	**p-value**
F1	0.554	3.292	0.001
F2	0.422	4.848	0.000
F3	0.692	2.838	0.000
F4	0.116	14.871	0.000
F5	0.153	4.424	0.000
F6	0.672	3.443	0.002
F7	0.808	2.025	0.000
F8	0.826	2.151	0.000

From the results in Table (-2-) we found all important principal component have significant relationship with the response variable. But from the results in Table (-2-) ,we will obtain the original parameters estimation for independent variables without multicollinearity problem This is shown in Chart 3.

Chart 3 . Regression coefficients on original variables by principal component regression

Principal component	**Regression coefficient**	**t-value**	**p-value**
X1	.063	6.926	0.000
X2	-.046	5.648	0.000
X3	.043	2.606	0.005
X4	.034	3.584	0.000
X5	.002	9.000	0.000
X6	.075	8.300	0.000
X7	.021	1.266	0.187
X8	-.038	3.974	0.000
X9	-.070	2.877	0.006
X10	.109	2.354	0.007
X11	.070	8.856	0.000
X12	-.046	1.755	0.076
X13	.019	9.436	0.000
X14	.226	3.974	0.000
X15	-.052	1.983	0.452
X16	-.046	1.583	0.325
X17	-.064	4.038	0.000
X18	.378	2.000	0.004
X19	.378	2.000	0.002
X20	-.042	2.568	0.004

From the results in Table 3., the independent variables have positive or inverse relationship with the response variable, as follows: X1 (The period of time it took to complete the study of the external research stage) is a -significant variable from statistical point of view in the response variable. X2 (Age at the end of the external search stage of the respondent) is a -significant variable from statistical point of view in the response variable . X3 Social status and (Obtaining the stage of diligence) (The place of residence is a countryside or a city) is non-significant variable from statistical point of view in the response variable X4 (The place of residence is a countryside or a city) and (Obtaining the stage of diligence)X5 Your area of residence outside or inside Najaf and is a significant variable from statistical point of view in the response variable X6 Monthly income and (Obtaining the stage of diligence)) significant variable from statistical point of view in the response variable. . X7 The number of your children and (Obtaining the stage of diligence)) is a non-significant variable from statistical point of view in the response variable. X8 Academic level is a significant variable from statistical point of view in the response variable. X9 Number of reading hours per dayis is a significant variable from statistical point of view in the response variable.x10 Use of the Internet.x10 Use of the is a significant variable from statistical point of view in the response variable x11The efficiency of the professor in the introduction stage is a significant variable from statistical point of view in the response variable x12The efficiency of the professor in the intermediate stage called the surfaces and (Obtaining the stage of diligence)) is significant variable from statistical point of view in the response variable. x13The efficiency of the professor in the intermediate stage called the surfaces is a significant variable from statistical point of view in the response variablex14 The type of study is systematic or open and (Obtaining the stage of diligence)) is a significant variable from statistical point of view in the response variable. X15 role of the family and is a non-significant variable from statistical point of view in the response variable x16 The role of the community and (Obtaining the stage of diligence)) is a non-significant variable from statistical point of view in the response variable. x17 Discuss research with colleagues and(Obtaining the stage of diligence)) a significant variable from statistical point of view in the response variable;are in positive relationship x18 Teaching during your studies is a significant variable from statistical point of view in the response variable x19date of birth and (Obtaining the stage of diligence)) is a significant variable from statistical point of view in the response variable. x20 Type of housing with family, owned, rent, student dormitory and is a significant variable from statistical point of view in the response

V. CONCLUSION

The our model is suffer from multicollinearity problem. Therefore, we must use method more robust to overcoming this problem , the principal component regression has a good properties to analyzing the data until with multicollinearity problem. In our data principal component regression is effective model. From 20 principal component ,there are only eight principal components are high effective, it can explanation 83.633% from total variance . Also there are 17 independent variables are significant effect in the response variable . We recommend extension this work to a new models such as lasso principal component regression , adaptive lasso principal component regression and elstic-net principal component regression and others

VI. ACKNOWLEDGEMENTS

I Would like to express my special thanks of gratitude to my **Dr.Fadel Hamid Hadi** .as well as our **teaching staff** that they gave me the golden opportunity to do this wonderful project on the topic the multilinearity with principalcomponent regression., which also helped me in doing a lot of research and I come to know about so many new things.

I new really thankfulto them. Secondly I would also like a thank my .. family who helped me a lot in finishing this project within the limited It helped me increase my knowledge and skills. THANKS AGAIN TO ALL WHO SUPPORTED

VII. REFERENCES

1- Chatterjee, S. & Price, B. (1991). Regression diagnostics. New York, John Wiley.
Farrar, D. E. & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. Review of Economics and Statistics, 49, 92-107.

2- Jeffers, N. R. (1967). Two case studies in the application of principal component analysis. Appl.Statist., 16, 225-236.

3- Morrison (1976). Multivariate. statistical methods. The Wharton School, University of Pennsylvania. McGraw-Hill, Inc. New York St.

4- Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.

5- Greene, W. H., (2007). Econometric analysis, 7th ed. New York University. Intrilligator, M. D. (1996). Econometrics models, techniques and applications, Prentice Hall.

6- Yan, X. & Gang, S. X., (2009). Linear regression analysis, theory and computing. London, World Scientific Publishing. Co. Pte. Ltd.

7- R.X. Liu; J. Kuang; Q. Gong; X.L. Hou (2003). Principal component regression analysis with spss. , 71(2), 141–147.

8-] N.R. Draper, H. Smith, Applied Regression Analysis, 2nd ed., John Wiley & Sons Inc, New York, 1981

9- S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, “Multilinear discriminant analysis for face recognition,” IEEE Trans. Image Process., vol. 16, no. 1, pp. 212–220, Jan. 2007.

10-L. Lathauwer, B. De Moor, and J. Vandewalle, “A multilinear singular value decomposition,” SIAM J. Matrix Anal. Appl., vol. 21, no. 4, pp. 1253–1278, 2000.