Journal of Statistical and Econometric Methods, vol. 2, no.1, 2013, 33-41
ISSN: 2051-5057 (print version), 2051-5065(online)
Scienpress Ltd, 2013

Using PC Regression for Multicollinear Model With Lagged Variable

Habib Ahmed Elsayir[1]

Abstract

This paper aims at identifying a most frequently multivariate technique,Principal Components Analysis(PCA), to solve a multicollinear single equation econometric model .results of the method used were compared to Ordinary Least Squares(OLS) and(Two Stages Least Squares(2SLS) to see if satisfactory results can be obtained. The proposed technique was applied to annual time series economic data, mainly total value added in agriculture. The method seemed to have little usefulness in the model; this might be referred to the nature and the number of the explanatory variables under concern. The method seemed to have few applications in economic fields and recommended when the number of explanatory variables included in the model is very large relative to sample size, or when multicollinearity exists.

Keywords: Lagged variables, Least squares estimators, Multicollinearity, Principal components analysis.

1 Introduction

Given a large complex number of variables: $x_1, x_2, \dots, x_p$,reduction of such variables whose inter-relationships are complex to a much smaller set of new variables (say k), whose inter-relationships are simple, but which contain most of the variation in the original variables may be of interest. The question of how many dimensions there are or how much independence that really is in the set of these (k) variables, then arises. Some times in econometrics, for instance the number of predetermined variables be included in a function, is a large relative to the size of the sample. In addition, in a multiple regression

[1](PhD Statistics) Dept. of Mathematics, Al Qunfudha University College, Umm AlQura University, Saudi Arabia,
e-mail: Habibsayiroi@yahoo.com

Article Info: *Received :* December 16, 2012. *Revised :* January 12, 2013.
Published online : March 1, 2013

form when there is severe multicollinearity (*Evgenia Vogiatzi (2002*), reliability of estimates may not be sensibly assessed. In such cases, need for a method to overcome these problems arises. This study aims at using a proposed method, mainly principal component regression, a method for coping with multicollinearity among independent variables (see the the discrption and the applications in Frank Wood (2009), Tarım , B.Dergisi(2011), (*Fekedulegn, B. Desta; etal (2002)* **,** compared to OLS estimates to overcome the underlying problem and to see if better estimates may be gained based on an econometric model data.

2 Model Specification

2.1 Introduction

Econometric model building has been applied and widely accepted as a standard approach to forecast and policy simulation in modern economies. It is possible to construct large and disaggregated models that can monitor economic performance to be useful guidelines for policy decision.A number of studies have been made early on the Sudanese economy, but few of these attempt to construct a full scale and single econometric models (see *M.S.Marzouk(1975) A.Elsheikh(1979),A.Elsheikh(1983),*and *Osman,S.N.(1997)),* since most of the studies were concerned with a general description of the Sudanese economy and its major development problems.

In this study, an econometric single equation model is introduced for an endogenous variable in terms of three exogenous variables, where:

AG: total value added in agriculture in millions of (L.s).

X good: total exports of goods in millions of (L.s).

CONS: personal consumption expenditures in millions of (L.s).

AG (-1): lagged total value added in agriculture in millions of (L.s).

The agricultural activity contributes to total GDP by 47.4%(*Bank of Sudan(1997*).The value added in the primary sector is related to private consumption expenditures (basic demand for food), total exports (foreign demand for basic agricultural crops) and lagged government investment in agricultural schemes .Thus total value added in agriculture is determined by total exports of goods ,personal consumption, and lagged total value added in agriculture .This relation can be formulated in the known regression model as follows:

$y = \beta_{0+}\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu$

Where Y is agricultural value added.

In economic variables, the problem of multicollinearity may arise for various reasons:

I) Tendency of economic variables of moving together overtime:

For instance, periods of booms or rapid economic growth, the basic economic magnitude grow, although some tend to lag behind others. Thus income, consumption investment, prices, savings, employment tend to rise in periods of economic expansion and decrease in periods of recession. Growth and trend factors in time series are most serious causes of multicollinearity(*Abd.Gadir(1998*).

II) The use of lagged models

In some models, lagged values of explanatory variables are used as separate independent factors in the relationship. For instance, in consumption functions, it has become usual to include the past as well as the present value of income. However, in investment functions past levels of economic activity are included as explanatory variables. So the successive values of a certain variable are intercorrelated, for example income in the current period may be determined by its previous value, and so on. Thus multicollinearity is almost certain to exist in distributed lag models *(Koutsoyiannis,A(1977),Abd.gadir(1998*). Taking the above considerations into account it is obvious that some degree of collinearity is expected to appear in most economic relationships models.
It should be noted that although multicollinearity problem is usually related with time series, it is quite frequent in cross-section data as well. But it tends to be more common and more serious problem in time series.
After the test, an appropriate procedure should be followed to overcome the problem of multicollinearity .The solutions which may be adopted if multicollinearity exists depend on the severity of multicollinearity as well as other information about the data and the purpose of the study. If multicollinearity has serious effects on the coefficients estimates of important factors, one of the following corrective solutions should be adopted ,M. El-Dereny and N. I.Rashwan(2011) *Abd.Gadir(1998),Draper&Smith(1981):*

1. Application of methods incorporating quantitative information, such as: restricted least squares, pooling of cross-section and time series data, Durbin's version of generalized least squares, mixed estimation technique proposed by Theil and Goldberger.
2. Increase of the sample size.
3. Substitution of lagged variables for other explanatory variables in the model.
4. Introduction of additional equations in the model.
5. Application of the principal components method.

3 Model Estimation and Testing

A Sudanese econometric model was estimated with annual observations covering the period 1956 to 1997.The model is also estimated on the basis of degree of correlation. In the equation, the numbers in parentheses below the right –hand side coefficients are their respective standard errors and R^2 is the coefficient of determination where DW is the Durbin-Watson statistic.

3.1 Model I

Applying the data for the period from 1970 to 1997, using (ECONOMETRIC) computer software dropping the constant since it is insignificant-the model is:

$$AG = \underset{(0.001255)}{-0.002426}Xgood + \underset{(0.30262)}{0.681382}CONS - \underset{(-0.76180)}{0.295517}AG(-1)$$

Sign t statistic AR(1) = 0.00
R-squared = 0.999999
D.W. = 2.00839
Prob.value = 0.00
F-statistic =6525954

After the model parameters have been estimated, the fitted model should be subject to diagnostic checks and tests of goodness of fit (*Box Jenkins (1976*). There is variety of criteria to judge the model performance.

The high R^2 (0.999) suggests that dependent variable (AG) almost totally determined by the defined explanatory variable i.e. the variation in the agricultural sector is almost fully explained by the existing explanatory variables; total value added, personal consumption, and lagged agriculture value added. Further, the results show no autocorrelation after using autoregressive technique of the flirt order; AR (1) according to D.W.statistic (2.008839), the model also shows relatively small standard errors. The sign of the coefficient of the first predetermined variable is different from what is expected to be. The correlation matrix for the above model variables are shown in table (3-1).

Further investigations to the above model resulting in presence of linear relationship between the two explanatory variables (XGood) and (CONS), where R=0.9962.

The whole diagnostics denotes presence of multicollinearity which may be due to the tendency of economic variables of moving together overtime. The simple correlation between the two predetermined variables (R=0.9962) is nearly equal to the overall multiple correlation of the relationship, so multicollinearity is seemed to be harmful.

The method of principal components can be used to overcome the consequences of multicollinearity.

To understand the use of principal components in regression analysis, it is useful to consider the information available in, for instance, the last or *pth* component. The eigen value of that component will express how much of the total variance in the X it explains. For standardized data magnitudes, if the eigen value of the last principal component (which must be smallest) is almost one, then the simple correlation among the X variables must be close to zero. With low or zero simple correlations, the length of the principal component axes within the ellipse of concentration is nearly the same and multicollinearity is not a problem.

At the other extreme side, if the numerical value of the last eigenvalue is close to zero, then the length of its principal axis within the ellipse of concentration is very small. Since this eigenvalue can also be considered, the variance of the *pth* component, the variance of this component is close to zero. For standardized data with zero mean, the mean of the *pth* principal component is zero and its variance is close to zero. The values of the coefficient of the principal components can give information on this relationship.

When multicollinearity exists, the examination of the last few principal components will provide information on which variables are causing the problem and on the nature of the interrelationship among them.

3.2 Model II

If the sample is adjusted to include 34 observations instead of 24, that is to cover the period from 1960 to 1977 to drop some estimated values, the model will be:

$$AG = -0.002414Xgood + 0.681754CONS - 0.295261AG(-1)$$
$$(0.001054) \qquad (0.681754) \qquad (0.062139)$$

Sign t statistic AR(1) = 0.00
R-squared = 0.999999
D.W. = 1.988390
Prob.value = 0.00
F-statistic =9932836

It is obvious from the above results that the same previous results hold and no additional information were gained. It should be noted that Durbin-Watson test is criticized (see *Johnston (1984*) when a lagged dependent variable appears among the regressors, and the combination of a lagged Y variable and a positively autocorrelated disturbance term will bias the D.W.statistic upward and thus gives misleading indications.Thus, if the explanatory variables include a lagged dependent variable, Durbin's h statistic must be used. *Elena Pesavento (2007).*The consequences of accepting H_0(that is the zero autocorrelation, when autocorrelation is present are almost more serious than the consequences of incorrectly assuming it to be absent. when the regressors are slowly changing series, as many economic series are, the true critical value will be close to the D.W. upper bound(*Johnston(1984*).

Instead of using the original regression variables in regression model, the principal components of the variables are used. The justification of this procedure is that the components are uncorrelated so that the contribution of each to the regression equation may be separated out simply. But the resulting regression is somewhat difficult to interpret, and there is no guarantee that the most important components in the regression equation will be those with the largest variances.

3.3 Model III: Two -Stage Least Squares Estimators

The 2SLS is applied to the single equation econometric model as an alternative to OLS.The model output is:

$$AG = \underset{(27.353)}{-95.591} + \underset{(0.00211)}{5.572Xgood} - \underset{(0.000374)}{5.496CONS} + \underset{(0.1146)}{1.021lagG} + U$$

Prob.value = 0.00
F-statistic =1061.938
Multiple R =0.99688
R-squared = 0.99376
Standard error = 1.988390

The R^2 for 2SLS is equal to that of OLS,and smaller standard error of regression in 2SLS than in OLS is seen, but less F statistic. Examining the correlation matrix of parameter estimates, (Table3-2), there is high multicollinearity between the predetermined variables. However, the 2SLS, in general, yields an estimated $\hat{Y}$ series which displays less correlation with the residual series than does the original Y series.

3.4 Model IV

The technique of principal component regression now to be applied to see if satisfactory results are attainable. It has been mentioned earlier that the technique of principal component aims at constructing a new variables $p_1, p_2, \ldots, p_k$ smaller than the number of the X's.i.e transformation of data through rewriting the data with properties the original data did not have (See *Mike Wulder (2005)*).If only one component is extracted and the regression has been run, then the following model is obtained:

$$AG = 5493618 + 383471.8p_1 + U$$
$$(92886.48) \quad (6721.125)$$

Sign t statistic = 0.00
R-squared = 0.987861
D.W. = 2.282747
Prob.value = 0.000
F-statistic =3255.239

It is noted that the results are not better than the previous ones, but the coefficient signs are positive ones.

3.5 Model V

As in the case of retaining two components, that represent the most of variation, the regression model is:

$$AG = 7778199 + 1128833p_1 + 3371555p_2 + U$$
$$(7327.468) \quad (2319.106) \quad (10473.19)$$

Sign t statistic = 0.00
R-squared = 0.999995
D.W. = 1.728734
Prob.value = 0.000
F-statistic =4270324

Again it is clear that no change was made. The standard error of regression here became smaller than in the precedent one and so is the sum squared residuals. Hence estimates of principal components regression are similar to that of OLS. The economic data seemed to illustrate a situation in which results are not clear cut that principal component estimates are better than the OLS.The whole results were sequentially presented in table(3-3).Thus by selecting one or two components, we can not be sure that the principal components estimates are in general better than the estimates of OLS applied to the original X's.It is only when the number of X's is too large relative to small size of the sample that one should adopt the principal component technique to obtain better results.However,the sample size seemed to be very large with respect to the number of variables.

4 Discussion

The application of this method for detection of multicollinearity is criticized on the grounds that it uses less information from the sample than the OLS method. The information of the sample may not be sufficient for reliable estimation of all the coefficients of the model. The greater the degree of multicollinearity, the less reliably OLS allocate the variation in Y among the explanatory variables. The suggested solutions for multicollinearity mentioned before involve the use of more information (such as restrictions or priori information) based on the knowledge about the explanatory variables.

The advantage of using principal components analysis that it both helps in understanding what variables are causing multicollinearity and provided a method for obtaining stable(though biased) estimates of the slope coefficients. If serious multicollinearity

exists, the use of these coefficients will result in larger residuals in the data from which they were obtained and smaller multiple correlation than when least squares is used, but the estimated standard error of the slopes could be predict better in a sample.
In the principal component method the multicollinear X's are transformed into orthogonal artificial variables .If these artificial variables can be given any specific economic meaning, then they can be used as variables in their own right, and transformations provides considerable solution to the problem of multicollinearity, because in this case a meaningful reduction is achieved in the number of the original model. however in most cases ,the constructed variables cannot be interpreted directly as economic variables, and their use implies utilization of only a part of all the information of the sample (the part incorporated into the principal components),that is, a reduction of information instead of application of the required increase in the information of the sample.

5 Conclusion

From the regression analysis model, it has been reached that principal components estimates are not better than the OLS estimates. In econometric model, the method is recommended when the number of explanatory variables been included in a function is very large relative to size of the sample, or when there is a need to check for problem of multicollinearity.In addition, principal components regression is extremely valuable technique only when the principal components have a "real meaning" in the practical system or future work can be conducted using principal components as working variables.

References

[1] Ahmed ElSheikh M.A.(1979).An Econometric Model for the Sudan.(Ph.D),Exter University,Exter,England.
[2] Ahmed ElSheikh M.A.(1983).Critique of Ten Year Plan: An Econometric Model. Khartoum.
[3] Atia,Abd.Gadir(1998).Econometrics: Theory and Practice,2nd ed., Alexanderia, University House(in Arabic).
[4] Bank of Sudan.The Annual 37 Report ,Several Years.
[5] Draper, N.R.and Smith, H.(1981).Applied Regression Analysis.2nd.ed.New York; John Wiley & Sons,Inc.
[6] Elana Pasavento(2007).Residuals-based tests for the null of no-cointegration :an analytical comparison. Journal of Time Series Analysis Volume 28 Issue1 Page111-137.
[7] Evgenia Vogiatzi (2002).Problems in Regression analysis and their Corrections In http:// www.geocities.com;qecon 2002/index html
[8] Fekedulegn, B. Desta; etalt, (2002):Coping with Multicollinearity: An Example on Application of Principal Components Regression in Dendroecology**:** U.S. Department of Agriculture, Forest Service, Northeastern Research Station. 43p. -721
[9] George E.P.Box&G.M.Jenkins (1976).Time Series Analysis Forecasting and Control: Holden-Day; Inc.California, USA.
[10] Johnston, J. (1984).Econometric Method.3rd.ed.Singapore: Mc Graw-Hill Book Co.

[11] Koutsoyiannis, A.(1977).Theory of Econometrics.2nd.ed.New York: Harper and Row Pub lishing Inc.
[12] M. El-Dereny and N. I. Rashwan .(2011) Solving Multicollinearity Problem Using Ridge Regression Models ,Int. J. Contemp. Math. Sciences, Vol. 6, 2011, no. 12, 585 - 600
[13] Mike Wulder , (2005)Jan1 0. Multivariate Statistics: Principal Components and Factor Analysis Mike Wulder - Personal Web Site.
[14] M.S.Marzouk.(1975).An Econometric Model of sudan.Journal of Development and Economics. North Holland Publishing Company, pp.337-348.
[15] Tarım Bilimleri Dergisi Tar. Bil. Der.Dergi (2011). Multivariate Multiple Regression Analysis Based on Principal Component Scores to Study Relationships between Some Pre- and Post-slaughter Traits of Broilers Journal of Agricultural Sciences,in www.agri.ankara.edu.tr/dergi.

Appendix

Table (3-1): Correlation Matrix of variables.

	AG	XGOOD	CONS
AG	1.000	0.9902	0.9986
XGOOD	0.9902	1.000	0.9962
CONS	0.9986	0.9962	1.000

Source: Model results.

Table (3-2): Correlation Matrix of Parameter Estimates of 2SLS

	CONS	LAGG	XGOOD
CONS	1.000	-0.9273	-0.9997
LAGG	-0.9273	1.000	0.9213
XGOOD	-0.9997	0.9213	1.000

Source: Model results.

Table (3-3): The Model Results.

Model	No. of Obs.	R-squared	S.e. Of Regression	Sum squared Residuals	D.W.	Prob. value	F-statistic
Model I (OLS)	24	0.99999	1419.94	40324599	2.008839	0.00	6525954
Model II (OLS)	34	0.99999	1159.19	40312099	1.988390	0.00	9932836
Model III (2SLS)	42	0.99376	45.02	40537.7	-	0.00	1061.93
Model IV (P.C.)	42	0.987861	126173.7	6.37	2.28274	0.00	3255.3
Model V (P.C)	42	0.999995	2478.36	2.40	1.7287	0.00	4270324