ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 11, Issue 1, pp: (27-32), Month: April 2024 – September 2024, Available at: www.paperpublications.org

A Robust Estimator for Causal Inference: Integrating Two Stage Least Squares with Principal Component

Toba Temitope Bamidele[1], Alabi Olatayo Olusegun[2]

[1,2] Department of Statistics, Federal University of Technology, Akure, Ondo State, Nigeria

DOI: https://doi.org/10.5281/zenodo.12671069

Published Date: 06-July-2024

Abstract: **Multicollinearity remains a significant concern in Simultaneous Equation Models, as the inherent interdependence of variables within the system can lead to high levels of correlation among the independent variables. This can have substantial implications for the reliability and interpretation of parameter estimates in the Simultaneous Equation Models. As such, this paper propose an extended two stage least squares (2sls) estimator by introducing principal components which will eliminate the concern of high correlation among the variables in the simultaneous equation model. The extended 2sls estimator (2sls-pc estimator) transforms the predictors to principal components before producing an estimate. In a bid to compare the classical two stage least squares (2sls) with the two stage least squares with principal components (2sls-pc), simultaneous equation model with three equations predicting Final Consumption Expenditure, Gross Domestic Investment and Gross Domestic Product were modelled. The 2sls-pc estimator addressed the high collinearity between the dataset and produced more significant estimate than the classical 2sls estimator while retaining all the information from the original dataset.**

Keywords: **Endogeneity, Endogenous, Exogenous, Multicollinearity, Principal Components, SEM, Two stage Least square.**

I. INTRODUCTION

Simultaneous Equation Models (SEM) are a class of econometric models used to analyze relationships between multiple interdependent variables [1]. Unlike single-equation regression models, where the dependent variable is determined by a set of independent variables, in a simultaneous equation system, the variables can be both dependent and independent, influencing each other simultaneously. The key characteristics of SEM include the interdependence of variables, the presence of endogenous variables, and the need for specialized estimation techniques. Endogenous variables are determined within the system of equations, and this interdependence can lead to identification issues, which must be addressed to ensure the model parameters can be uniquely determined from the observed data. SEM has the ability to capture the interdependence of variables, address endogeneity, and improve estimation accuracy compared to single-equation methods [2]. Two-Stage Least Squares (2SLS) is widely used as the estimation technique for Simultaneous Equation Models (SEM). The key advantage of 2SLS is its ability to produce consistent and unbiased parameter estimates in the presence of endogeneity, which is a common problem in SEM [1]. Additionally, 2SLS can be more efficient than other methods, such as Indirect Least Squares (ILS), when the system is over identified [2]. However, the effectiveness of 2SLS depends on the availability and validity of instrumental variables, and the potential impact of data quality on the reliability of the estimation results. The reference to data quality is an important factor to consider. SEM assumes that the observed variables are measured without error. It requires complete data for all variables in the model, as missing data can lead to biased and inefficient parameter estimates, as well as potential issues with model identification [3]. Furthermore, high levels of multicollinearity among the variables in the SEM can lead to unstable and unreliable parameter estimates [3, 4].

Multicollinearity remains a significant concern in Simultaneous Equation Models (SEM), as the inherent interdependence of variables within the system can lead to high levels of correlation among the independent variables. This can have substantial implications for the reliability and interpretation of parameter estimates in SEM [5, 6]. When two or more independent variables in the SEM are highly correlated, it becomes challenging for the model to disentangle their unique effects on the dependent variables [7]. This can result in unstable and unreliable parameter estimates, characterized by large standard errors, wide confidence intervals, and low statistical significance [8, 6]. Moreover, multicollinearity can also impact the overall model fit and interpretability, as the high correlations among the independent variables can make it difficult to assess the individual contributions of each variable to the dependent variable [5]. This can lead to challenges in interpreting the results and understanding the underlying causal mechanisms within the SEM. Researchers has considered removing one or more of highly correlated variables from the model as remedial action, but his could lead to the removal of essential variables with more information about the predicted variable.

Sujata study [9] demonstrated that the Principal Component Regression outperforms Multiple Linear Regression when it comes to handling out-of-sample data in presence of multicollinearity. The study iterated that a common practice for addressing moderate to severe multicollinearity is to eliminate independent variables which can result in the loss of valuable information. On the other hand, Principal Component Regression proves to be a superior alternative for dealing with multicollinearity issues while preserving the richness of the data. By reducing the number of predictors using Principal Component Analysis (PCA), Principal Component Regression (PCR) can help mitigate multicollinearity issues and potentially improve the stability, interpretability and applicability of the regression model.

In addition, a study in a bid to assess the best prediction model for ozone concentration between Multiple Linear Regression (MLR) and Principle Component Regression (PCR), collected an hourly data on ozone, nitrogen dioxide, nitrogen oxide, temperature, relative humidity and wind speed from 2012 to 2014. Principle Component Analysis (PCA) was used in order to reduce multicollinearity problem, prior to the implementation of MLR. The hybrid model of PCR was selected as best-fitted models as it had higher R squared values compared with MLR model [10].

These studies are few of many studies that has established that Principal Component Analysis is very useful in addressing multicollinearity in Ordinary Least Square Regression. As such, this paper propose an extended 2sls estimator by introducing principal component which will eliminate the concern of high correlation among the variables in the simultaneous equation models.

II. PROPOSED ESTIMATOR

Models

$$y_1 = \beta_{1,0} + \beta_{1,1}y_2 + \beta_{1,2}y_3 + \cdots + \beta_{1,n-1}y_n + \beta_{1,n}y_{n+1} + \beta_{1,n+1}x_1 + \beta_{1,n+2}x_2 + \cdots + \beta_{1,n+m}x_m + u_1$$

$$y_2 = \beta_{2,0} + \beta_{2,1}y_1 + \beta_{2,2}y_3 + \cdots + \beta_{2,n-1}y_n + \beta_{2,n}y_{n+2} + \beta_{2,n+1}x_{m+1} + \beta_{2,n+2}x_{m+2} + \cdots + \beta_{2,n+m}x_{2m} + u_2$$

$$\vdots$$

$$y_n = \beta_{n,0} + \beta_{n,1}y_1 + \beta_{n,2}y_2 + \cdots + \beta_{n,n-1}y_n + \beta_{n,n}y_{2n} + \beta_{n,n+1}x_{M-m} + \beta_{n,n+2}x_{M-m+1} + \cdots + \beta_{n,n+m}x_M + u_n$$

where,

y_i for $i = 1,2, \dots, n$ are the endogenous variables,

y_i for $i = n+1, n+2, \dots, 2n$ are the instrument variables,

x_i for $i = 1,2, \dots, m, m+1, m+2, \dots, 2m, \dots M$ are the exogenous variables,

$\beta_{i,j}$ are the co-efficient of the variables, and

u_i for $i = 1,2, \dots, n$ are the error terms.

Stage 1 Least Square Estimation

Let each of the instrument variables $(y_{n+1}, y_{n+2}, \dots, y_{2n})$ and the exogenous variables $(x_1, x_2, \dots, x_M)$ be a px1 matrix, where p is the number of elements in each variable. If there exist a px1 all-ones matrix J such that matrix $W = (J, y_{n+1}, y_{n+2}, \dots, y_{2n}, x_1, x_2, \dots, x_M)$.

Then,

$$\hat{\alpha}_i = (W'W)^{-1}W'y_i \tag{1}$$

and,

$$\hat{y}_i = W\hat{\alpha}_i \tag{2}$$

for $i = 1,2,\dots,n$

Stage 2 Least Square Estimation with Principal Component

The estimated values from eq. 1 are substitute into the models as shown below:

$$y_1 = \beta_{1,0} + \beta_{1,1}\hat{y}_2 + \beta_{1,2}\hat{y}_3 + \cdots + \beta_{1,n-1}\hat{y}_n + \beta_{1,n}y_{n+1} + \beta_{1,n+1}x_1 + \beta_{1,n+2}x_2 + \cdots + \beta_{1,n+m}x_m + u_1$$

$$y_2 = \beta_{2,0} + \beta_{2,1}\hat{y}_1 + \beta_{2,2}\hat{y}_3 + \cdots + \beta_{2,n-1}\hat{y}_n + \beta_{2,n}y_{n+2} + \beta_{2,n+1}x_{m+1} + \beta_{2,n+2}x_{m+2} + \cdots + \beta_{2,n+m}x_{2m} + u_2$$

$$\vdots$$

$$y_n = \beta_{n,0} + \alpha_{n,1}\hat{y}_1 + \beta_{n,2}\hat{y}_2 + \cdots + \beta_{n,n-1}\hat{y}_n + \beta_{n,n}y_{2n} + \beta_{n,n+1}x_{M-m} + \beta_{n,n+2}x_{M-m+1} + \cdots + \beta_{n,n+m}x_M + u_n$$

Let

$$X_1 = (\hat{y}_2, \hat{y}_3, \dots, \hat{y}_n, y_{n+1}, x_1, x_2, \dots, x_m) = \begin{pmatrix} a_{111} & a_{112} & \cdots & a_{11k} \\ a_{121} & a_{122} & \cdots & a_{12k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1j1} & a_{1j2} & \dots & a_{1jk} \end{pmatrix}$$

$$X_2 = (\hat{y}_1, \hat{y}_3, \dots, \hat{y}_n, y_{n+2}, x_{m+1}, x_{m+2}, \dots, x_{2m}) = \begin{pmatrix} a_{211} & a_{212} & \cdots & a_{21k} \\ a_{221} & a_{222} & \cdots & a_{22k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{2j1} & a_{2j2} & \dots & a_{2jk} \end{pmatrix}$$

$$\vdots$$

$$X_n = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n, y_{2n}, x_{M-m}, x_{M-m+1}, \dots, x_M) = \begin{pmatrix} a_{n11} & a_{n12} & \cdots & a_{n1k} \\ a_{n21} & a_{n22} & \cdots & a_{n2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{nj1} & a_{nj2} & \dots & a_{njk} \end{pmatrix}$$

Generally,

$$X_i = \begin{pmatrix} a_{i11} & a_{i12} & \cdots & a_{i1k} \\ a_{i21} & a_{i22} & \cdots & a_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{ij1} & a_{ij2} & \dots & a_{ijk} \end{pmatrix} \tag{3}$$

where,

a_{ijk} is the j[th] value in the k[th] variable of the i[th] equation.

The standardized data matrix Z is given as:

$$Z_i = \begin{pmatrix} z_{i11} & z_{i12} & \cdots & z_{i1k} \\ z_{i21} & z_{i22} & \cdots & z_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{ij1} & z_{ij2} & \dots & z_{ijk} \end{pmatrix} \tag{4}$$

The standardized variables in the matrix are obtained as:

$$z_{ijk} = \frac{a_{ijk} - \bar{a}_{ik}}{s_{ik}} \tag{5}$$

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 11, Issue 1, pp: (27-32), Month: April 2024 – September 2024, Available at: www.paperpublications.org

where,

a_{ijk} is the j[th] value in the k[th] variable of the i[th] equation,

$\bar{a}_{ik}$ is the mean of the k[th] variable of the i[th] equation,

S_{ik} is the standard deviation of the k[th] variable of the i[th] equation.

The covariance matrix Σ_i is given as:

$$\Sigma_i = \frac{Z_i{}' Z_i}{n} \tag{6}$$

The eigen values λ_{ij} and eigen vectors v_{ij} of the covariance matrix Σ_i is computed as:

$$\left|\Sigma_i - \lambda_{ij}\mathrm{I}\right| = 0 \tag{7}$$

$$\left(\Sigma_i - \lambda_{ij}\mathrm{I}\right)v_{ij} = 0 \tag{8}$$

where,

λ_{ij} is the j[th] eigen value in the i[th] equation,

v_{ij} is the j[th] eigen vector in the i[th] equation,

I is an identity matrix.

Suppose V_i is a matrix containing all the eigen vectors of the i[th] equation i.e $V_i = (v_{i1}, v_{i2}, v_{i3}, \dots, v_{ij})$. Then, the principal component K_i obtained by transforming the standardized data matrix Z_i using the eigen vectors V_i is given as:

$$K_i = Z_i V_i \tag{9}$$

The two stage least square with principal components (PC-2SLS) estimator is:

$$\hat{\beta}_{pc-2sls\ i} = (K_i{}' K_i)^{-1} K_i{}' y_i \tag{10}$$

III. DATA AND METHODOLOGY

A. Data

Data on Gross Domestic Product (GDP), Final Consumption Expenditure (CON), General Government Final Consumption Expenditure (GOV), Gross Domestic Investment (GDI), Net Primary Income (NPI), Net Taxes on Products (TAX) and Manufacturing Output (MAN) in billion dollars were collected from 1981 to 2020 for this research.

B. Model

The simultaneous equation model below shows three equations predicting Final Consumption Expenditure (CON), Gross Domestic Investment (GDI) and Gross Domestic Product (GDP)

$$CON = \beta_{10} + \beta_{11} GDP + \beta_{12} GOV + \beta_{13} NPI + u_1$$

$$GDI = \beta_{20} + \beta_{21} GDP + \beta_{22} MAN + \beta_{23} TAX + u_2$$

$$GDP = \beta_{30} + \beta_{31} CON + \beta_{32} GDI + \beta_{33} GOV + u_3$$

where β_{ij} are the parameters to be estimated.

C. Method

Classical 2SLS estimator and PC-2SLS estimator were used to obtain the estimate of the simultaneous equation model for comparison.

i. Two Stage Least Squares (2SLS) Estimator

First Stage

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 11, Issue 1, pp: (27-32), Month: April 2024 – September 2024, Available at: www.paperpublications.org

As established in eq. 1 and eq. 2

$$\hat{\alpha}_i = (W'W)^{-1}W'y_i$$

and,

$$\hat{y}_i = W\hat{\alpha}_i$$

Second Stage

Define the matrix X_i as shown in eq. 3, then estimate $\hat{\beta}_i$ using ordinary least square:

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i \quad (11)$$

ii. Two Stage Least Squares with Principal Component (PC-2SLS) Estimator

First Stage

As established in eq. 1 and eq. 2

$$\hat{\alpha}_i = (W'W)^{-1}W'y_i$$

and,

$$\hat{y}_i = W\hat{\alpha}_i$$

Second Stage

As defined in eq. 10

$$\hat{\beta}_{pc-2sls\ i} = (K_i'K_i)^{-1}K_i'y_i$$

D. Results

Table 1: Table of Estimates

	Response Variable	Predictor Variable	Estimate	Std. Error	t value	Pr(>|t|)	VIF	
Two-Stage Least Square Estimates	CON	Intercept	-217.260	113.234	-1.919	0.063		Multiple R-squared = 0.926
		GDP	3.794	1.754	2.162	0.037	2005.649	Adjusted R-squared =0.9198
		GOV	-55.811	32.749	-1.704	0.097	4514.194	AIC = 416.0638
		NPI	39.259	22.761	1.725	0.093	614.154	BIC = 424.5082
	GDI	Intercept	-2.945	4.487	-0.656	0.516		Multiple R-squared = 0.8445
		GDP	-0.104	0.043	-2.452	0.019	10.114	Adjusted R-squared = 0.8316
		MAN	4.503	0.502	8.968	$1.05e^{-10}$	10.416	AIC = 330.0829
		TAX	-11.552	3.227	-3.580	0.001	7.106	BIC = 338.5273
	GDP	Intercept	17.918	9.577	1.871	0.069		Multiple R-squared = 0.8445
		CON	0.830	0.199	4.159	0.0002	31.482	Adjusted R-squared = 0.8316
		GDI	0.621	0.236	2.638	0.012	2.223	AIC = 393.4967
		GOV	2.833	1.933	1.466	0.151	27.644	BIC = 401.9411
Two-Stage Least Square with Principal Component Estimates	CON	Intercept	137.226	6.457	21.254	$< 2e^{-16}$		Multiple R-squared = 0.926
		GDP-PC	-79.775	3.869	-20.618	$< 2e^{-16}$	1	Adjusted R-squared =0.9198
		GOV-PC	80.530	17.232	4.673	$4.06e^{-05}$	1	AIC = 416.0638
		NPI-PC	1013.335	552.007	1.836	0.076	1	BIC = 424.5082
	GDI	Intercept	51.058	2.204	23.164	$< 2e^{-16}$		Multiple R-squared = 0.8445
		GDP-PC	-15.152	1.323	-11.448	$1.49e^{-13}$	1	Adjusted R-squared = 0.8316
		MAN-PC	34.114	7.201	4.738	$3.34e^{-05}$	1	AIC = 330.0829
		TAX-PC	59.479	9.171	6.485	$1.57e^{-07}$	1	BIC = 338.5273
	GDP	Intercept	196.336	4.870	40.319	$< 2e^{-16}$		Multiple R-squared = 0.8445
		CON-PC	102.602	3.063	33.497	$< 2e^{-16}$	1	Adjusted R-squared = 0.8316
		GDI-PC	-36.291	7.891	-4.599	$5.09e^{-05}$	1	AIC = 393.4967
		GOV-PC	56.767	37.688	1.506	0.141	1	BIC = 401.9411

The table above shows that the estimate for Multiple R-squared, Adjusted R-squared, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) are the same for both the 2sls and the 2sls-pc estimators because the Principal Components are the linear combination of the original data set. When all the principal components obtained from the dataset are retained, the principal component retain all the same information in the dataset.

Meanwhile, direct comparison of the standard errors is not straightforward given that these response variables relate to entirely different values in 2sls predictors and 2sls-pc predictors. Exploring the difference in the dataset of the two predictors, the Variance Inflation Factor (VIF) of the 2sls predictors from the table shows that the predictors suffers severe multicollinearity which was addressed in the 2sls-pc predictors before producing an estimate. The 2sls-pc therefore produced more significant estimate.

IV. CONCLUSION

This study aims to extended 2sls estimator by introducing principal component to eliminate the concern of high correlation among the variables in the simultaneous equation models. The extended estimator - 2sls estimator with principal components (2sls-pc estimator) addressed the high collinearity between the dataset and produced more significant estimate than the classical 2sls estimator while retaining all the information from the original dataset.

REFERENCES

[1] Greene, William H. Econometric Analysis. 7th ed., Prentice Hall, 2012.

[2] Wooldridge, Jeffrey M. “Introductory Econometrics: A Modern Approach”. 6th ed., Cengage Learning, 2016.

[3] Kline, Rex B. “Principles and Practice of Structural Equation Modeling”. 4th ed., Guilford Press, 2015.

[4] Tabachnick, Barbara G., and Linda S. Fidell. “Using Multivariate Statistics”. 6th ed., Pearson, 2013.

[5] Hair, Joseph F., et al. “When to Use and How to Report the Results of PLS-SEM”. European Business Review, vol. 31, no. 1, pp. 2-24, 2019

[6] Kline, Rex B. “Principles and Practice of Structural Equation Modeling”. 4th ed., Guilford Press, 2016.

[7] Wooldridge, Jeffrey M. “Introductory Econometrics: A Modern Approach”. 7th ed., Cengage Learning, 2020.

[8] Gujarati, Damodar N., and Dawn C. Porter. “Essentials of Econometrics”. 5th ed., McGraw-Hill, 2017.

[9] Suvarnapathaki, Sujata. “Use of Principal Component Analysis in Regression Problem-Dr Sujata Suvarnapathaki”. Journal, vol. 10, pp. 125-135, 2023.

[10] Binti Mohd Napi, Nur Nazmi, et al. “Multiple Linear Regression (MLR) and Principal Component Regression (PCR) for Ozone (O3) Concentrations Prediction”. IOP Conference Series: Earth and Environmental Science, vol. 616, p. 012004, 2020.