A New Kind of Two-Stage Least Squares Based on Shapley Value Regression

Mishra, SK

30 December 2017

A New Kind of Two-Stage Least Squares Based on Shapley Value Regression

Sudhanshu K. Mishra
C-91 (GF) Avantika, Rohini Sector-1,
Delhi – 110085 (India)
Contact: mishrasknehu@hotmail.com

Abstract

The Two-Stage Least squares method for obtaining the estimated structural coefficients of a simultaneous linear equations model is a celebrated method that uses OLS at the first stage for estimating the reduced form coefficients and obtaining the expected values in the arrays of current exogenous variables. At the second stage it uses OLS, equation by equation, in which the explanatory expected current endogenous variables are used as instruments representing their observed counterpart. It has been pointed out that since the explanatory expected current endogenous variables are linear functions of the predetermined variables in the model, inclusion of such expected current endogenous variables together with a subset of predetermined variables as regressors make the estimation procedure susceptible to the deleterious effects of collinearity, which may render some of the estimated structural coefficients with inflated variance as well as wrong sign. As a remedy to this problem, the use of Shapley value regression at the second stage has been proposed. For illustration a model has been constructed in which the measures of the different aspects of globalization are the endogenous variables while the measures of the different aspects of democracy are the predetermined variables. It has been found that the conventional (OLS-based) Two-Stage Least Squares (2-SLS) gives some of the estimated structural coefficients with an unexpected sign. In contrast, all structural coefficients estimated with the proposed 2-SLS (in which Shapley value regression has been used at the second stage) have an expected sign. These empirical findings suggest that the measures of globalization are conformal among themselves as well as they are positively affected by democratic regimes.

Key words: Simultaneous equations model, Two-Stage Least Squares, Instrumental Variables, Collinearity, Shapley Value Regression, Democracy Index, Index of Globalization.

JEL Code: **C30, C36, C51, C57, C61, C71, F63**

1. Introduction: A simultaneous equations linear econometric model is YA+XB+U=0, where $Y_{n,m}$ are m number of current endogenous variables each in n observations, $X_{n,k}$ are k number of predetermined variables each in n observations, $U_{n,m}$ are structural disturbances, $A_{m,m}$ is the structural coefficient matrix associated with Y, and $B_{k,m}$ is the structural coefficient matrix associated with X. The matrix A has full rank such that A^{-1} exists. All the main diagonal elements of A are a-priori known to be -1, and the other elements may be either zero or unknown (that need to be empirically estimated). The coefficient matrix B associated with X may have some of the elements that are a-priori known to be zero and the others are unknown (that need to be empirically estimated). If n>k and X has a full column rank of k, it is possible to proceed to estimation of the unknown elements of A and B. On account of the fact that the current endogenous variables (Y) are random and correlated with the structural disturbances (U), estimation of A and B by Ordinary Least Squares (OLS) is problematic due to the stochastic regressor problem (Theil, 1971, p. 452; Intriligator, 1978, pp. 375-377) non-conformal to the Gauss-Markov conditions.

Post-multiplying YA+XB+U=0 by A^{-1} we get $YAA^{-1} + XBA^{-1} + UA^{-1} = 0$, or $Y = X\Pi + V$, where $\Pi = -BA^{-1}$ and $V = -UA^{-1}$. The equations in the system $Y = X\Pi + V$ is said to be in the reduced form. The matrix of the reduced form coefficients, Π, may be estimated as $P=(X'X)^{-1} X'Y$ by using OLS. But this is the only first stage of estimation. The real crux of the problem lies in obtaining estimated A and B from P. For any particular i^{th} equation, we have $Pa_i=-bi$. It may be shown that for a particular i^{th} equation if we partition the column vector a_i of A into $[a_{i\alpha} \mid a_{i\beta}]$ and the column vector b_i of B into $[b_{i\alpha} \mid b_{i\beta}]$ such that $a_{i\alpha}$ contains m_1 unknown elements (while $a_{i\beta}$ contains $m_2=m-m_1$ known elements) and similarly $b_{i\alpha}$ contains k_1 unknown elements (while $b_{i\beta}$ contains $k_2=k-k_1$ known elements), then we have a system of linear equations (dropping the subscript i for convenience)

$$\begin{bmatrix} P_{\alpha\alpha} & P_{\alpha\beta} \\ P_{\beta\alpha} & P_{\beta\beta} \end{bmatrix} \begin{bmatrix} a_\alpha \\ a_\beta \end{bmatrix} = - \begin{bmatrix} b_\alpha \\ b_\beta \end{bmatrix}, \text{ whence } P_{\beta\alpha} a_\alpha = - \begin{bmatrix} b_\beta & P_{\beta\beta} a_\beta \end{bmatrix}$$

This system is in k_2 equations and m_1 unknown quantities (a_α). If a_α can successfully be estimated, we may also obtain b_α by substitution. This is called the 'identification problem'. If $k_2 < m_1$ or even if $k_2 \geq m_1$ but $P_{\beta\alpha}$ does not have full column rank, the equation i is said to be 'under-identifiable' and consequently, one cannot obtain the estimated values of the structural coefficients, $a_{i\alpha}$ and $b_{i\alpha}$.

2. The Two-Stage Least Squares for Identified structural Equations: As mentioned by Anderson (2005) it was shown by Theil (1953) and Basmann (1957) that for any particular structural equation if $k_2 \geq m_1$ and $P_{\beta\alpha}$ is not deficient in column rank one may first obtain the OLS-based expected values of Y by the relationship $\hat{Y} = XP$ and substitute the estimated $\hat{y}$ for observed y in the equation wherever it is a regressor variable. However, the regressand y would remain as it was (and will not be substituted by $\hat{y}$). Thus, among the regressor variables $\hat{y}$ would be used as an instrumental variable (Reiersøl, 1945) representing y. This approach (nick-named as the Two-Stage Least Squares method or 2-SLS) would render the use of OLS free from the blemishes of stochastic regressor problem since $\hat{y}$ is highly correlated with y as well as $\hat{y}$ is net of the error (otherwise, error is contained in the observed y). Therefore, at the second stage, the regressor variables are the estimated endogenous variables together with the predetermined variables relevant and specified in a particular equation of the model. Under certain conditions (that errors across the equations in the model are uncorrelated), the Two-Stage Least squares is a celebrated method of estimation of the parameters of a structural equation in the simultaneous equations model framework. It can be used for estimating the parameters of any (exactly identified as well as over-identified) structural equations one by one and hence for estimating all the equations of a simultaneous equation model.

3. Susceptibility of Estimated structural Parameter to the Deleterious Effects of Collinearity: Since, at the second stage, the 2-SLS uses the estimated values of some endogenous variables as well as some predetermined variables (pooled together) as regressors, it is not unlikely that collinearity among the regressor variables should crop up. This is likely due to the fact that the estimated values of endogenous variables are themselves the linear functions of the predetermined variables in the model. Collinearity among the regressor variables may not affect the explanatory power of the regressors adversely, but it surely affects standard errors of the estimated parameters. It is not unlikely that the estimated parameters are so inaccurate that their signs may be wrong (Smith and Brainard, 1976). In any case, collinearity may lead to a type II error or failure to reject a false null hypothesis of no effect of the explanatory variable(s).

4. Shapley Value Regression and its Application at the Second Stage of 2-SLS: The concept and properties of Shapley values came from the cooperative game theory with collusions and found their application in estimating regression parameters in collinear conditions. The method is described as follows. Let y_i be the dependent variable of the i^{th} structural equation and let Z = $[\hat{Y}_i \mid X_i]$ be the explanatory variables incorporating the estimated current endogenous variables (at the first stage of the 2-SLS) and the predetermined variables appearing in the i^{th} equation. In total, let Z have t number of variables. Let $\mathbf{z}_i \subset Z$ in which $z_i \in Z$ is not there or $z_i \notin \mathbf{z}_i$. Thus, $\mathbf{z}_i$ will have only t-1 variables. We draw r (r=0, 1, 2, ... , t-1) variables from $\mathbf{z}_i$ and let this collection of variables so drawn be called P_r such that $P_r \subseteq \mathbf{z}_i$. Also, $\mathbf{z}_i = \mathbf{z}_i \cup \emptyset$. Now, P_r can be drawn in L=tCr ways. Also, let $Q_r = P_r \cup z_i$. Regress (least squares) y on Q_r to find R^2_q. Regress (least squares) y on P_r to obtain R^2_p. The difference between the two R squares is $D_r = R^2_q - R^2_p$, which is the marginal contribution of z_i to y. This is done for all L combinations for a given r and arithmetic mean of D_r (over the sum of all L values of D_r) is computed. Once it is obtained for each r, its mean is computed. Note that P_r is null for r=0, and thus Q_r contains a single variable, namely z_i. Further, when P_r is null, its R^2 is zero. The result is the arithmetic average of the mean (or expected) marginal contributions of z_i to Z. This is done for all z_i; i=1, t in order to obtaining the Shapley value (S_i) of z_i; i=1, t (Mishra, 2017a). Once the Shapley values are worked out, one may derive the conventional regression coefficients by an optimization exercise (Lipovetsky, 2006; Mishra, 2016). Shapley value regression significantly ameliorates the deleterious effects of collinearity on the estimated parameters.

5. Simultaneous Equations Model of the relationship between Democracy measures and Globalization Measures: In this study we construct a simultaneous equation model in which five measures of different aspects of a regime (ranging between the two poles of full democracy and authoritarianism) aim at explaining six indicators of globalization. We have modelled the relationships in which globalization indicators are endogenous variables while the indicators of political regime are predetermined (exogenous) variables, and estimated the model with the data borrowed from Mishra (2017b).

The Economist Intelligence Unit (EIU), a British business within the Economist Group has published the Democracy Index for 2006, 2008 and 2011 and for every year afterwards. The index is based on 60 indicators grouped in five different categories or dimensions of regime ranging from democracy to authoritarianism. These five categories are: Electoral process and pluralism (EPP), Functioning of government (FOG), Political participation (PPN), Political culture (PCL) and Civil liberties (CVL). Subsequently, these five measures of different aspects of democracy are suitably weighted and aggregated to yield an overall index (OSC, or the Index of Democracy with the score value in the range of zero to ten). On the basis of the score value (OSC) the political systems of different countries may be classified into Full democracies (score value in 8-10 range), Flawed democracies (score value in 6 to below-8 range), Hybrid regimes (score value in 4 to below-6 range) and authoritarian regimes (score value below 4). In the present work we have used EPP, FOG, PPN, PCL and CVL for the year 2006 that pertain to 116 countries (Mishra, 2017b).

As to the measures of different aspects of globalization, we have used the KOF (KOF, 2017; Dreher, 2006; Dreher et al., 2008) indices for 2006-2014. During this period, different countries (116 in number having provided a long series of data on different aspects of globalization) have scored the highest levels of overall globalization in different years. We have considered only those years and the data pertaining to them for our study at hand (Mishra, 2017b). The indices of globalization are six in number. They are: (1). E1 - actual economic flows such as trans-border trade, direct investment and portfolio investment, (2). E2 - relaxation of restrictions on trans-border trade as well as capital movement by means of taxation, tariff, etc., (3) S1 - trans-border personal contacts such as degree of tourism, telecom traffic, postal interactions, etc., (4) S2 - flow of information, (5) S3 - cultural proximity, and (6) P - the measure of trans-national political set up.

Our simultaneous equation model is given in the schematic form as under:

1. $E1 = f(E2, S1, FOG, PCL, CVL) = a_{12}E2+a_{13}S1+b_{12}FOG+b_{14}PCL+b_{15}CVL+const_1 + u_1$
2. $E2 = f(S2,S3,P,EPP,PPN) = a_{24}S2+a_{25}S3+a_{26}P+b_{21}EPP+b_{23}PPN+ const_2 + u_2$
3. $S1 = f(E1,S3,FOG,PCL,CVL) = a_{31}E1+a_{33}S1+a_{35}S3+b_{32}FOG+b_{34}PCL+b_{35}CVL+const_3 + u_3$
4. $S2 = f(E2,FOG,PCN,PCL,CVL) = a_{42}E2+_{b42}FOG+b_{43}PPN+b_{44}PCL+b_{45}CVL + const_4 + u_4$
5. $S3 = f(P,EPP,FOG,PPN,PCL) = a_{56}P+b_{51}EPP+b_{52}FOG+b_{53}PPN+b_{54}PCL + const_5+ u_5$
6. $P = f(E1,E2,S1,S2,S3) = a_{61}E1+ a_{62}E2+a_{63}S1+a_{64}S2+a_{65}S3 + const_6 + u_6$

6. Empirical Findings: The reduced form coefficients (based on OLS) are given in Table-1. At the second stage, we have estimated the structural parameters (A and B) by OLS (i.e. conventional 2-SLS) and presented them in Table-2. As the proposed alternative at the second stage, we have also estimated the structural parameters by Shapley value regression and presented them in Table-3. Optimization has been done by the Host-Parasite Co-Evolutionary algorithm, which is a powerful biologically inspired population method of global optimization (Mishra, 2013). Under the current endogenous parameters matrix (A), the elements in the principal diagonal (minus unity) pertain to the dependent endogenous variables. A zero in the cell denotes that the variable has not been included in the particular equation. In Table-4 we have presented the R^2 obtained for different equations. It is seen that conventional 2-SLS at the second stage gives the R^2 values that are identical to those obtained for the reduced form equations. However, the R^2 values for the proposed 2-SLS (in which OLS is replaced by the Shapley value regression) are a little smaller than those given by the conventional 2-SLS based on OLS. This cost has to be paid for treating the collinearity problem that has devastating effects on the coefficients of the structural equations.

Table-1. Estimated Reduced Form Coefficients Matrix [Transposed P]

EQN	Regressand Variable	Reduced Form Regressor Variables (All Predetermined) Relating to the Political Regime					
		EPP	FOG	PPN	PCL	CVL	CONST
1	E1	0.12400	0.10555	0.07432	0.20298	-0.06958	42.89204
2	E2	0.09679	0.23024	-0.02864	0.24574	0.05692	26.76866
3	S1	-0.16699	0.12057	0.22784	0.48374	0.30252	-8.67294
4	S2	-0.07373	0.08948	0.19270	0.30243	0.20114	25.23083
5	S3	-0.06626	0.45089	0.36109	0.43824	0.11419	-30.03597
6	P	0.06859	0.01120	0.17536	0.21128	-0.04321	52.97000

.

Table-2. Estimated Structural Parameters Based on Conventional 2-SLS Estimation

EQN	Current Endogenous Variables: Transposed A Matrix						Predetermined Variables: Transposed B Matrix					
	E1	E2	S1	S2	S3	P	EPP	FOG	PPN	PCL	CVL	CONST
1	-1.0000	2.3547	0.6222	0.0000	0.0000	0.0000	0.0000	-0.5116	0.0000	-0.6766	-0.3918	-14.7428
2	0.0000	-1.0000	0.0000	0.0160	0.5052	0.0923	0.1251	0.0000	-0.2303	0.0000	0.0000	36.6456
3	-0.9095	0.0000	-1.0000	0.0000	0.8182	0.0000	0.0000	-0.1523	0.0000	0.3098	0.1458	54.9108
4	0.0000	-0.7618	0.0000	-1.0000	0.0000	0.0000	0.0000	0.2649	0.1709	0.4896	0.2445	45.6232
5	0.0000	0.0000	0.0000	0.0000	-1.0000	-2.6429	0.1150	0.4805	0.8245	0.9966	0.0000	109.9563
6	1.2483	-0.6065	-0.0036	0.4164	-0.0394	-1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.9434

.

Table-3. Estimated Structural Parameters Based on Shapley Value Regression at the 2nd Stage of 2-SLS												
EQN	Current Endogenous Variables: Transposed A Matrix						Predetermined Variables: Transposed B Matrix					
	E1	E2	S1	S2	S3	P	EPP	FOG	PPN	PCL	CVL	CONST
1	-1.0000	0.1783	0.1106	0.0000	0.0000	0.0000	0.0000	0.0740	0.0000	0.0929	0.0614	-30.2465
2	0.0000	-1.0000	0.0000	0.2071	0.1157	0.3142	0.0693	0.0000	0.1051	0.0000	0.0000	-51.7940
3	0.4186	0.0000	-1.0000	0.0000	0.1708	0.0000	0.0000	0.1327	0.0000	0.2071	0.1117	-62.6993
4	0.0000	0.2502	0.0000	-1.0000	0.0000	0.0000	0.0000	0.1087	0.1289	0.1635	0.0969	-44.1332
5	0.0000	0.0000	0.0000	0.0000	-1.0000	0.7641	0.1297	0.2512	0.2468	0.2919	0.0000	-109.4473
6	0.1968	0.1237	0.0862	0.1179	0.0599	-1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-35.4534

.

Table-4. R-Square Values for different equations at different stages of 2-SLS								
Stage	Estimation	R^2	EQUATION-1	EQUATION-2	EQUATION-3	EQUATION-4	EQUATION-5	EQUATION-6
First	Reduced form	PRSQ	0.24173	0.51836	0.51797	0.50351	0.48426	0.28982
Second	Conventional 2-SLS	SRSQ	0.24173	0.51836	0.51797	0.50351	0.48426	0.28982
Second	Shapley Value 2-SLS	SSRSQ	0.22952	0.48162	0.49500	0.49216	0.47196	0.27350

The structural coefficients in Table-2 (obtained by the conventional 2-SLS) reveal that some off-diagonal elements of the structural coefficient matrix (transposed A or A') are negative. The negatively signed off-diagonal elements in A' (in Table-2) suggest that the measures of globalization have a conflict among themselves. In particular, E1 (trans-border actual economic flows) has a negative effect on S1 (trans-border personal contacts), E2 (relaxation of restrictions on trans-border trade) has a negative effect on S2 (flow of information), P (trans-national political set up) has a negative effect on S3 (cultural proximity) and, especially in the last equation, E2, S1 and S3 have negative effects on P. These negative effects suggested by the elements of A' (in Table-2, based on the conventional 2-SLS) are quite unexpected. Similarly, the elements of the transposed B or B' (Table-2, row 1) suggest that FOG (Functioning of government), PCL (political culture) and CVL (Civil liberties) negatively affect E1 (trans-border actual economic flows). FOG (Functioning of government) shows up a negative effect on S1 (trans-border personal contacts), which is unlikely. In short, the structural coefficients obtained by the conventional 2-SLS are misleading.

In contrast, a perusal of Table-3 (the structural coefficient matrices A' and B' obtained by the proposed Shapley value regression at the second stage) suggests that the coefficients associated with endogenous as well as predetermined variables (off-diagonal elements of A' and the elements B' - except the constant term) are all positive. They suggest that globalization measures are concordant with each other and the democratic regimes promote globalization. These results are in consonance with the research findings elsewhere (Mishra, 2017b).

7. Concluding Remarks: The 2-Stage Least squares method for obtaining the estimated structural coefficients of a simultaneous linear equations model is a celebrated method that uses OLS at the first stage for estimating the reduced form coefficients and obtaining the expected values in the arrays of current exogenous variables. At the second stage it uses OLS, equation by equation, on the structural variables included in the concerned equation in which the explanatory current endogenous variables are replaced by their expected counterpart. Thus, the explanatory expected current endogenous variables are used as instruments representing the explanatory observed current endogenous variables. It has been pointed out that since the explanatory expected current endogenous variables are linear functions of the predetermined variables in the model, inclusion of such expected current endogenous variables together with a subset of predetermined variables as regressors make the estimation procedure susceptible to the deleterious effects of collinearity, which may render some of the estimated structural coefficients with inflated variance as well as wrong sign. As a remedy to this problem, the use of Shapley value regression at the second stage has been proposed. For illustration a model has been constructed in which the measures of the different aspects of globalization are the endogenous variables while the measures of the different aspects of democracy are the predetermined variables. It has been found that the conventional (OLS-based) 2-

SLS gives some of the estimated structural coefficients with an unexpected sign. In contrast, all structural coefficients estimated with the proposed 2-SLS (in which Shapley value regression has been used at the second stage) have an expected sign. These empirical findings suggest that the measures of globalization are conformal among themselves as well as they are positively affected by democratic regimes.

References

Anderson, T.W. (2005).Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics*, 127 (1): 1-16.

Basmann, R.L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25(1): 77–83.

Dreher, A. (2006). Does Globalization Affect Growth? Evidence from a new Index of Globalization. *Applied Economics*, 38(10): 1091-1110.

Dreher, A., Gaston, N. and Martens, P. (2008). *Measuring Globalisation: Gauging its Consequences*. New York: Springer.

Intriligator, M.D. (1978). *Econometric Models, Techniques, and Applications*. Amsterdam: North-Holland.

KOF [Konjunkturforschungsstelle or Economic Research Centre of ETH Zurich]. (2017). 2017 Index of globalization. http://globalization.kof.ethz.ch/media/filer_public/2017/04/19/rankings_2017.pdf

Lipovetsky, S. (2006). Entropy criterion in logistic regression and Shapley value of predictors. *Journal of Modern Applied Statistical Methods*, 5(1): 95-106.

Mishra, S.K. (2013). Global Optimization of Some Difficult Benchmark Functions by Host-Parasite Coevolutionary Algorithm", *Economics Bulletin*, 33(1): 1-18.

Mishra, S.K. (2016). Shapley Value Regression and the Resolution of Multicollinearity. *Journal of Economics Bibliography*, 3(3): 498-515.

Mishra, S.K. (2017a). Almost equi-marginal principle based composite index of globalization: China, India and Pakistan. *Journal of Economic and Social Thought*, 4(3): 335-351.

Mishra, S.K. (2017b). Are Democratic Regimes Antithetical to Globalization? Working Paper. SSRN: https://ssrn.com/abstract=3088921.

Reiersøl, O. (1945). Confluence Analysis by Means of Instrumental Sets of Variables. *Arkiv for Mathematic, Astronomi, och Fysik. 32A*. Uppsala: Almquist & Wiksells.

Smith, G. and Brainard, W. (1976). The value of a priori information in estimating a financial model. Journal of Finance, 31(5): 1299-1322.

Theil, H. (1953). Repeated least-squares applied to complete equation systems. Central Planbureau, Memorandum.

Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.