International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 17, 2019, pp 115-128

Social Variable Selection Model: Theoretical Background and its Implications

Sourav Kumar Das[1], Jagannath Biswas[2] and Soumik Kanti Ghosh[1]

[1]Department of Economics, Lalbaba College, University of Calcutta, West Bengal, India

[2]Department of Mathematics, Lalbaba College, University of Calcutta, West Bengal, India

[Received July 15, 2018; Revised March 12, 2019; Accepted April 25, 2019]

Abstract

In social science research, variables are the simplest factors from both micro and macro points of view. Generally, the independent variables are designated as 'treatments' or 'causes' and the dependent variables represent 'effects'. Since these two types of economic variables are inter-related, there have been studies on formulation and justification of mathematical equations relating the two sets of variables. In this paper Social Variable Selection (SVS) model has been introduced to examine both intra- and inter- relationship among. Further, SVS model has been empirically evaluated through cross section data on consumption pattern across tribes and across districts. An attempt has also been made to examine the underlying theory.

Keywords: Social Variable Selection model, Economic variables, Cross section data analysis.

AMS Classification: 05C20, 91C99, 65J05.

1. Introduction

Labelling of dependent and independent variables have always posed interesting research problems in a way that identifies a general cause (or a group of causes), and its (their) implied effect. It pertains to classifying a given set of variables into two subsets as either independent or dependent**.** In social science the process of selection of variables has been mainly based on researcher's objectives. There are no appropriate acceptable rules in social science for selecting major independent variables those effects manifest themselves on one or more dependent variables.

Kapetanios (2006) in his study of a regression model, selected the variables by using a non-standard Optimization of Information criterion. On the other hand, Stock and Watson (1989, 91) project the indices of leading and coincident economic indicators using the time series econometrics to select the best variables to use as components of the leading index. Finally they present an explicit time series model that implicitly defines a variable that can be thought of as the overall state of the economy.

Wang et al. (1996) study a model of Poisson regression and independent Poisson mixtures for selection of variables. Sin and White (1996) introduced mis-specified parametric models for selecting variables. Hunter and Li (2005) uses algorithms for finding differentiable functions for variable selection. Foroni and Marcellino (2011) assess the forecasting performance of various variable selection methods. Akaike (1973) had been estimating the number of factors in factor analysis, estimating the degree of a polynomial describing the data and finally selecting the variables by multiple regression equations.

Donoho and Johnstone (1994), with the help of ideal spatial adaptation, provide information about how best to adapt a spatially variable estimator, whether piecewise constant, piecewise polynomial, variable knot spline, or variable

bandwidth kernel, to the unknown function. Again De Mol et al. (2006) consider Bayesian regression with normal and double-exponential priors as forecasting methods based on large panels of time series data. Finally a class of variable selection procedures for parametric models via nonconcave penalized likelihood was proposed in Fan and Li (2001).

In this paper we make an attempt to develop an equation involving a few selected 'independent' [economic or otherwise qualified and relevant] variables and the dependent economic variable in a way that load of 'independent' variables can be maximized. The idea is to select a few most significant explanatory independent variables, conforming to our desired degree of accuracy. In the process, we introduce "Social Variable Selection Model" and study its validity and relevance empirically.

2. Social Variable Selection Model

We start with the basic premises that not all independent variables are of same importance in respect of selecting them to make an explicit function for understanding the nature and variability of the dependent variable. We introduce here a model named Social Variable Selection Model [SVSM] which will enable us to select the variables as per their economic importance.

Not to obscure the essential steps of reasoning, we proceed as follows.

Let a social factor V be a function of n social variables $(u_1, u_2, u_3, \dots, u_n)$.

Definition 1: According to the inter-social dependency axiom, two of the above-listed social variables 'u_i' and 'u_j' are said to form a *directed* interaction path if there is a 'directed path' from 'u_i to u_j', meaning thereby that knowledge of the social variable 'u_i' provides perfect knowledge of the social variable 'u_j'. This directed path is expressed by the symbol 'u_(i)→ u_(j)'.

Definition 2: A *directed* graph is formed of all or a subset of the 'n' social variables, say, (u_(i_1), u_(i_2),…,u_(i_r)) if between any two consecutive variables, there is a directed path from the former to the latter. Diagrammatically, this is given by:

u_(i_1)→u_(i_2)→….→→u_(i_r)

Definition 3: A *loop* based on (u_(i_1), u_(i_2),…,u_(i_r)) is a directed graph as in Definition 2 with the additional property that 'u_(i_r)' also forms a directed interaction path with 'u_(i,1)'. In this case, we have the following diagrammatical representation:

u_(i_1)→u_(i_2)→….→→u_(i_r) →u_(i_1).

Note that in a 'loop', all the social variables involved preserve some kind of 'symmetry'. In other words, in the above, all of $u_{i_1}, u_{i_2}, \dots, u_{i_r}$ are equivalently strong from the perspective of loop-formation.

It is tacitly assumed that there are some loops present in the given collection of n social variables.

We now examine another loop formed as:

$u_{j_1} \to u_{j_2} \to \cdots \to\to u_{j_m}$ →u_(j_1).

All of these m social variables make themselves equivalent in the sense of loop formation.

Proceeding in the same way, we can check all possible loops available in our collection consisting of the original n variables.

Now from the first loop, for convenience let us assume that only one of those r variables namely $u_{i_{r_1}}$ is also present in the second loop as $u_{j_{m_1}}$ and none of remaining $(u_{j_1}, u_{j_2}, \dots, u_{j_{m_1-1}}, u_{j_{m_1+1}}, \dots, u_{i_r})$ is equal to one of

$(u_{i_1}, u_{i_2}, \dots, u_{i_{r_1-1}}, u_{i_{r_1+1}}, \dots, u_{i_r})$. So the variable which is present as $u_{i_{r_1}}$ in first loop and as $u_{j_{m_1}}$ in second loop, may be said to be more prominent social factor which exerts its influence on a vast area of society through the effectiveness over those remaining $(r-1)+(m-1)=(r+m-2)$ variables in those two loops simultaneously.

On the other hand, if we select a variable u_k , where $k \in \begin{cases} \{i_1, i_2, \dots, i_{r_1-1}, i_{r_1+1}, \dots, i_r\} \, for \; 1st \; loop \\ \{j_1, j_2, \dots, j_{m_1-1}, j_{m_1+1}, \dots, j_m\} \, for \; 2nd \; loop \end{cases}$

among the remaining $(r+m-2)$ variables as an independent variable, then that would not be a representative of the entire pool of (r + m) social variables. That would only serve as equivalent to at most either $(r-1)$ or $(m-1)$ variables.

So to maximize the effectiveness of all or a large number of social factors by selecting minimum number of variables, we have to choose the variables like $u_{i_{r_1}}$ or $u_{j_{m_1}}$.

This is more or less one aspect of SVS Model and Selection of variable(s).

3. Working procedure of SVS Model

In this section, we will try to explain the working procedure of the Social Variable Selection (SVS) model.

We take common variables of all loop-making variables and write down them as

$$\alpha_1, \alpha_1, \alpha_1, \dots, \alpha_1(p_1 times), \alpha_2, \alpha_2\,, \alpha_2\,, \dots, \alpha_2(p_2 times), \dots, \alpha_s, \alpha_s, \alpha_s, \dots, \alpha_s(p_s\, times).$$

Here α_1 is present in p_1 number of loops, α_2 is present in $p_2$number of loops etc. Then we will select the variable with maximum frequency, and after selecting one we will take the next one with maximum frequency [among those available] and will proceed according to our desired degree of accuracy.

That means, our selected variables will be arranged in the order $\{\alpha_t\}$, where {t} is a decreasing sequence consisting of $\{p_1, p_2, \dots, p_s\}$

Now we will consider the correlation coefficient of each of the n social factors with V. For each social factor forming a loop, we locate the co-ordinates $(l, r*)$ in the 2D plane, where l = no. of loops made by the selected variable & r^* = 10 times the correlation coefficient r of the factor with *V*.

Now we draw the circles of radius $\sqrt{l^2 + r^{*2}}$, and the greater circles will give us the ultimate selected variables. Of course, the number of selected circles will depend on the desired degree of accuracy.

4. Implications of SVS model

Below we examine SVS model and its application with the help of the following case studies. The primary data of the Monthly Per Capita Consumption Expenditure of the tribal communities of Puruliya, Bankura and Paschim Midnapur districts of West Bengal have been analyzed by SVS model.

As we know Monthly Per Capita Consumption Expenditure depends mainly upon the Income distribution of the households.

$C_E = f(I)$ ……(i)

Now again Income is a function of Land Holding, Common Property Resources, Education and Occupation and these variables are functions of one another as given below:

$I=f_1(L_h,C_r,E_d,O_p)$……(ii)

$L_h=f_2(Po)$…….. (iii)

$C_r=f_3(I)$…… (iv)

$E_d=f_4(I)$…… (v)

$O_p=f_5(E_d,C_r)$…… (vi)

$Po= f_6 (E_d,I)$…… (vii)

In the above, C_e is the Monthly Per Capita Consumption Expenditure, I is the Per Capita Income, L_h is the per head Land Holding, C_r is the per head entitlement of Common Property Resources, E_d is the educational attainment, O_p is the occupation and P_o is the household size.

This relationship can be better understood with the help of the following diagram.

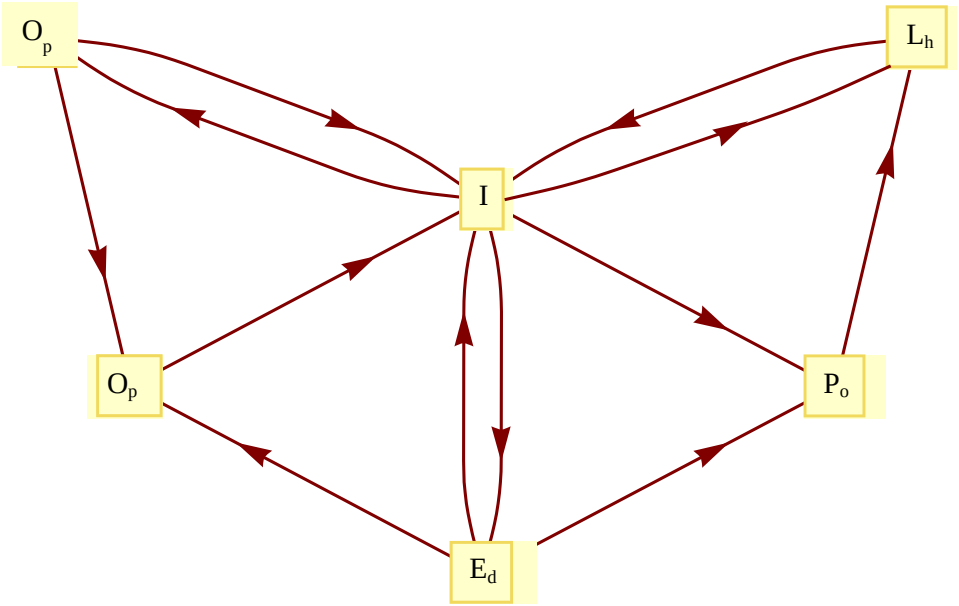

Figure 1: Relationship between the variables

In this model, the numbers of loops made by the variables are given below:

$I \rightarrow 5$, $L_h \rightarrow 2$, $Po \rightarrow 2$, $O_p \rightarrow 2$, $E_d \rightarrow 3$, $C_r \rightarrow 2$

Number of loops in this model represents the ground-level inter-relationship or economic domination upon other variables and larger correlation co-efficient with Ce makes a variable more distinguishable. Now let us see the pairwise correlation coefficients among the independent variables.

Table 1: Correlation coefficients involving economic variables.

	Income	Land Holding	Population	Occupation	Education	CPR
Income	1.000					
Land Holding	0.103	1.000				
Population	-0.308	0.369	1.000			
Occupation	0.165	-0.157	0.074	1.000		
Education	0.143	0.129	-0.003	0.054	1.000	
CPR	-0.039	0.018	0.067	0.005	-0.040	1.000

Source: Sources: Field survey & authors' own calculations

From the above table it is clear that the economic variables do not exhibit any significant pairwise correlation. Now let us see the correlation coefficient between the Monthly Per Capita Consumption Expenditure and the economic variables.

Table 2: Correlation coefficient between MPCE and economic variables.

	Income	Land holding	Population	Occupation	Education	Cpr
Purulia	0.858715	0.099268	-0.261075	0.154876	0.112229	-0.074422
Bankura	0.954371	-0.016092	-0.391132	0.172106	0.074760	-0.177310
Midnapur	0.867122	0.133051	-0.387832	-0.044714	0.239546	-0.186958

Source: Sources: Field survey & authors' own calculations

So if we consider a two-dimensional co-ordinate system with vertical axis as 10 times correlation co-efficient (r*) with Ce and in the horizontal axis the number of loops (l), then the corresponding co-ordinates will be (l, r*).

Here the study has been based on demand side analysis because from demand side one can estimate the supply side. Again SVS model is a better indicator than regression analysis because in SVS model we analyze both intra and inter dependency among the variables.

For the case of PURULIA:

$A_1(5, 8.58715)$, $A_2(2, 0.99268)$, $A_3(2, 2.61075)$, $A_4(2, 1.5487875)$

$A_5(3, 1.12229)$, $A_6(2, 0.744225)$

The symbols represent: A_1 for Income, A_2 for Land Holding, A_3 for Population, A_4 for Occupation, A_5 for Education and A_6 for Common Property Resources.

Now for the selection of variables we measure $\sqrt{(l^2 + r*^2)}$ to distinguish the greater circles, where the distance of each A_i from the origin are respectively 9.936757274, 2.232803973, 3.288771133, 2.529573624, 3.203050864, 2.133980049.

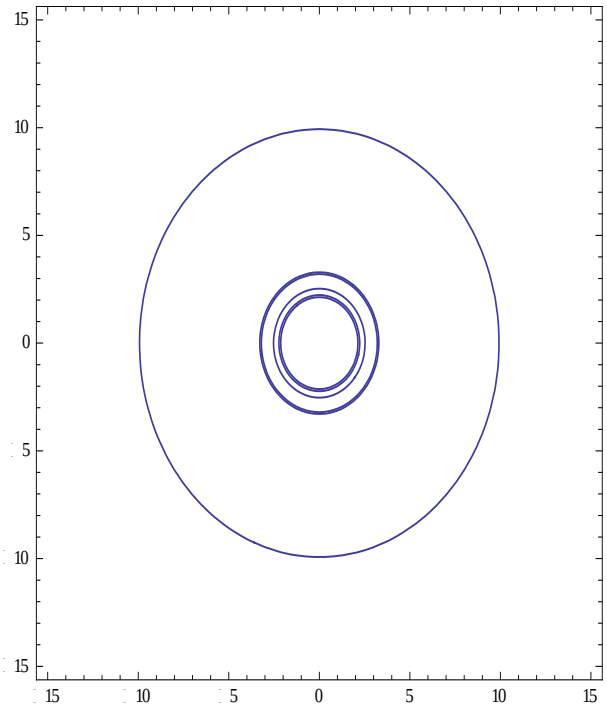

Figure 2: Selection of Variables

Obviously the 1st, 3rdand 5th circles exhibit themselves as most distinguished. So in this case, the selected variables are Income (I), Population (Po) and Education (Ed). It is also evident from the figure that all of A_2, A_4, A_6 lie inside the inner circles whose radii are nearly equal to each other.

Therefore, Ce = f (I, Po, Ed)

Now we can generate the general equation in explicit form as

$$Ce = b_{1.234} + b_{12.34}I + b_{13.24}Ed + b_{14.23}Po, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(a)}$$

where $b_{1.234}, b_{12.34}, b_{13.24}, b_{14.23}$ are given by $P^{-1}Q$

where P = $\begin{bmatrix} n & \sum I & \sum Ed & \sum Po \\ \sum I & \sum I^2 & \sum Ed.I & \sum Po.I \\ \sum Ed & \sum Ed.I & \sum Ed^2 & \sum Po.Ed \\ \sum Po & \sum Ed.I & \sum Po.Ed & \sum Po^2 \end{bmatrix}$, Q = $\begin{bmatrix} \sum Ce \\ \sum Ce.I \\ \sum Ce.Ed \\ \sum Ce.Po \end{bmatrix}$

For the case of Bankura

The co-ordinates are

A_1 (5, 9.54371), A_2 (2,-0.1609225), A_3 (2,-3.911325)

A_4 (2, 1.7210625), A_5 (3, 0.747595), A_6 (2,-1.7731)

where A_1 etc represent the same explanatory variables.

Now measuring $\sqrt{(l^2 + r^{*2})}$ we get the radii like: 10.77415429, 2.006463568, 4.393001623, 2.63857085, 3.091746801, and 2.672804447

Obviously the 1st, 3rd and 5th circles provide distinguished features. So in this case, the selected variables are I, Po and Ed.

Therefore, Ce = f (I, Po, Ed)

So the equation will be same as equation (a).

For the case of Paschim Midnapur

The co-ordinates are

A_1 (5, 8.671225), A_2 (2, 1.33051), A_3 (2,-3.878325)

A_4 (2,-0.447135), A_5 (3, 2.3954575), A_6 (2,-1.869575)

where A_1 etc represent the same explanatory variables.

Now measuring $\sqrt{(l^2 + r^{*2})}$, we get the radii like: 10.00950264, 2.402135895, 4.363645816, 2.049373004, 3.839038504 and 2.737756505

Obviously the 1st, 3rd and 5th circles turn out to be distinguished. So in this case, the selected variables are I, Po and Ed.

In other words, Ce = f (I, Po, Ed) is the recommended regression function.

So, again the equation will be same as equation (a).

Therefore it transpires that tribal communities of each of the three districts exhibit the same kind of trend in the aspect of measurement of MPCE with the help of independent variables, i.e., interestingly same social factors have been selected through our SVS model in each of three tribal community areas.

SVS model is thus seen to indicate a procedure for subset selection and to develop a meaningful empirical relationship.

5. Conclusions

Selection of variables in social science has always been a topic of interest. Various models and analyses are available for this purpose. In this paper, we have introduced Social Variable Selection (SVS) model and procedure for selection of meaningful subset of the economic variables, using the intra- and inter-relationship between the economic/social variables and the dependent variable. The Social Variable Selection (SVS) model has been empirically evaluated through cross section data. The model also provides the scope of selecting any number of variables up to the desired degree of accuracy.

References

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in Second International Symposium on Information Theory, eds. B.N. Petrox and F. Caski. Budapest: AkademiaiKiado, page 267.

[2] De Mol, C., Giannone, D., Reichlin, L. (2006). Forecasting with a large number of predictors: is bayesian regression a valid alternative to principal components?. Discussion Paper 5829, Centre for Economic Policy Research.

[3] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81, 425-455.

[4] Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96, 1348-1360.

[5] Foroni, C., Marcellino, M. (2011). A comparison of mixed frequency approaches for modelling euro area macroeconomic variables. Mimeo. Foroni, C., Marcellino, M., Schumacher, C. (2011), MIDAS and unrestricted MIDAS models.Mimeo.Goffe,

[6] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms, The 32 Annals of Statistics, 33, 1617-1642.

[7] Kapetanios, G. (2006). Variable selection in regression models using non-standard optimisation of information criteria, Computational Statistics and Data Analysis, 52-1, 4-15.

[8] Sin, C.Y., White, H. (1996). Information criteria for selecting possibly misspecifed parametric models. Journal of Econometrics, 71-(1-2), 207-225.

[9] Stock, J., Watson, M. (1989). New indexes of coincident and leading economic indicators. NBER Macroeconomic Annual, 351-394.

[10] Stock, J., Watson, M. (1991). A probability model of the coincident economic indicators. In Leading Economic Indicators: New Approaches and Forecasting Records, edited by K. Lahiri and G. Moore. Cambridge University Press.

[11] Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates, Biometrics, 52, 381-400.