Improve "2SLS" Method by Genetic algorithm with application

Alaa H Sabri ¹ and Sabah M Ridha ²

¹ Al-Muthanna University, College of Science, Department of Mathematics and Computer Applications, Iraq
² Baghdad University, College of Administration & Economics, Department of Statistics, Iraq

Email: alaa.sabri@gmail.com

Abstract. This paper explores potential power of Genetic Algorithm for optimization by using new MATLAB, most of the robust methods based on the idea of sacrificing in one side versus promotion another, the artificial intelligence mechanisms try to balance sacrifice and promotion to make the best solutions in a random search technique. In this paper, a new idea was introduced to improve the estimators of parameters of linear simultaneous equation models resulting from the 2SLS method by using a class of genetic algorithm which called binary genetic algorithm (GA) and better estimates were obtained using two robust different criteria.

Keywords: 2SLS, LSEM, binary genetic algorithm, GA

1. Introduction

Although Simultaneous Equation Models (SEM) have traditionally been used in the economic world, each equation in an SEM should represent some underlying conditional expectation that has a causal structure. The relationships between the variables are used to create the model, but these will depend on the criteria chosen (10) two structural equations fall out of the individual’s optimization problem: one has work as a function of the exogenous factors (7), demographics, and unobservables; the other has crime as a function of these same factors. The completeness of the system requires that the number of equations equal the number of endogenous variables (21). The leading method for estimating simultaneous equations models is the method of instrumental variables (IV). Therefore, the solution to the simultaneity problem is essentially the same as the IV solutions to the omitted variables and measurement error problems. The mechanics of Two-Stage Least Squares (2SLS) are similar because we specify a structural equation for each endogenous variable; we can immediately see whether sufficient IVs are available to estimate either equation (8). If the disturbances appearing in the various structural equations are not independently distributed, lagged endogenous variables are not independent of the current operation of the equation system, which means these variables are not really predetermined. If these variables are nevertheless treated as predetermined in the 2SLS procedure, the resulting estimators are not consistent (2). However, in finite samples under certain situations. Even when 2SLS is used, bias remains because an estimate of Reduced from is used since the true parameters are unknown (9). The three operators. Selection, crossover and mutation, make GA an important tool for optimization. The exploitation and exploration aspects of GAs can be controlled.
almost independently. This provides a lot of flexibility in designing a GA. this methodology is applicable even in those cases in which we do not know the form of heteroscedasticity and Least squares methodology is not applicable\(^{(13)}\). The MATLAB package comes with sophisticated libraries for matrix operations, general numeric methods and plotting of data, therefore MATLAB become first choice of programmer to implement scientific, graphical and mathematical applications and for the GA implementation\(^{(14)}\) and we could use the method of GA successfully in more flexible circumstances\(^{(6)}\).

2-Linear Simultaneous Equation Models (LSEM)

Consider 2 interdependent variables (endogenous variables) which depend on 4 independent variables (exogenous variables). Suppose that each endogenous variable can be expressed as a linear combination of the other endogenous variables, the exogenous variables, and white noise that represents stochastic interference. Thus, let us modify the income–money supply model as follows: \(^{(2)}\)

\[
\begin{align*}
Y_{1t} & = \beta_{10} + \beta_{12} Y_{2t} + \gamma_{11} X_{1t} + \gamma_{12} X_{2t} + u_{1t}, \quad (1) \\
Y_{2t} & = \beta_{20} + \beta_{21} Y_{1t} + \gamma_{23} X_{3t} + \gamma_{24} X_{4t} + u_{2t}, \quad (2)
\end{align*}
\]

Where

- \(Y_1\) = income
- \(Y_2\) = stock of money
- \(X_1\) = investment expenditure
- \(X_2\) = government expenditure on goods and services
The variables \(X_1\) and \(X_2\) are exogenous.

The income equation, a hybrid of quantity-theory–Keynesian approaches to income determination, states that income is determined by money supply, investment expenditure, and government expenditure. The money supply function postulates that the stock of money is determined (by the Federal Reserve System) on the basis of the level of income in addition to the variables already defined, \(X_3\) = income in the previous time period and \(X_4\) = money supply in the previous period. Both \(X_3\) and \(X_4\) are predetermined. It can be readily verified that both Eqs. (1) and (2) are overidentified.

3-Two-Stage Least Squares (2SLS) method

To apply 2SLS, we proceed as follows: In Stage 1 we regress the endogenous variables on all the predetermined variables in the system. Thus,

\[
\begin{align*}
\hat{Y}_{1t} & = \hat{\pi}_{10} + \hat{\pi}_{11} X_{1t} + \hat{\pi}_{12} X_{2t} + \hat{\pi}_{13} X_{3t} + \hat{\pi}_{14} X_{4t} + \hat{u}_{1t}, \quad (3) \\
\hat{Y}_{2t} & = \hat{\pi}_{20} + \hat{\pi}_{21} X_{1t} + \hat{\pi}_{22} X_{2t} + \hat{\pi}_{23} X_{3t} + \hat{\pi}_{24} X_{4t} + \hat{u}_{2t}, \quad (4)
\end{align*}
\]

A useful extension of linear regression is the case where \(y\) is a linear function of two or more independent variables\(^{(19)}\).

In Stage 2 we replace \(Y_1\) and \(Y_2\) in the original (structural) equations by their estimated value from the preceding two regressions and then run the OLS regressions as follows:

\[
\begin{align*}
Y_{1t} & = \beta_{10} + \beta_{12} \hat{Y}_{2t} + \gamma_{11} X_{1t} + \gamma_{12} X_{2t} + \hat{u}_{1t}, \quad (5) \\
Y_{2t} & = \beta_{20} + \beta_{21} \hat{Y}_{1t} + \gamma_{23} X_{3t} + \gamma_{24} X_{4t} + \hat{u}_{2t}, \quad (6)
\end{align*}
\]

Where
\[ \ddot{u}_{1t} = u_{1t} + \beta_{12}\dot{u}_{2t}, \]
\[ \ddot{u}_{2t} = u_{2t} + \beta_{21}\dot{u}_{1t}. \]

And the proxy variables \( \hat{Y} \) which are close to the endogenous variables, the proxy variables are highly correlated with the exogenous variables but uncorrelated with the error ones. \(^{(11)}\)
The estimates thus obtained will be consistent.

4- Genetic algorithm (GA)

Genetic algorithms (Holland, 1975) perform a search for the solution to a problem by generating candidate solutions from the space of all solutions and testing the performance of the candidates. The search method is based on ideas from genetics and the size of the search space is determined by the representation of the domain. \(^{(4)}\)

In a genetic algorithm, each individual of a population is one possible solution to an optimization problem, encoded as a binary string called a chromosome. A group of these individuals will be generated, and will compete for the right to reproduce or even be carried over into the next generation of the population. Competition consists of applying fitness function to every individual in the population; the individuals with the best result are the fittest. The next generation will then be constructed by carrying over a few of the best individuals, reproduction, and mutation.

Reproduction is carried out by a “crossover” operation, similar to what happens in an animal embryo. Two chromosomes exchange portions of their code, thus forming a pair of new individuals. In the simplest form of crossover, a crossover point on the two chromosomes is selected at random, and the chromosomes exchange all data after that point, while keeping their own data up to that point. In order to introduce additional variation in the population, a mutation operator will randomly change a bit or bits in some chromosome(s). Usually, the mutation rate is kept low to permit good solutions to remain stable. The two most critical elements of a genetic algorithm are the way solutions are represented, and the fitness function, both of which are problem-dependent. The coding for a solution must be designed to represent a possibly complicated idea or sequence of steps. \(^{(18)}\)

The basic genetic algorithm (GAs) is outlined as below

Step I [Start] Generate random population of chromosomes, that is, suitable solutions for the problem.

Step II [Fitness] Evaluate the fitness of each chromosome in the population.

Step III [New population] Create a new population by repeating following steps until the new population is complete.

a) [Selection] Select two parent chromosomes from a population according to their fitness. Better the fitness, the bigger chance to be selected to be the parent.

b) [Crossover] With a crossover probability, cross over the parents to form new offspring, that is, children. If no crossover was performed, offspring is the exact copy of parents.

c) [Mutation] With a mutation probability, mutate new offspring at each locus.

d) [Accepting] Place new offspring in the new population.

Step IV [Replace] Use new generated population for a further run of the algorithm.

Step V [Test] If the end condition is satisfied, stop, and return the best solution in current population.

Step VI [Loop] Go to step 2.

The genetic algorithms performance is largely influenced by crossover and mutation operators. The block diagram representation of genetic algorithms (GAs) is shown in Fig.1. \(^{(15)}\)
Figure 1 shows block schematic of various stages to perform genetic algorithms (GAs) optimization

5- Genetic Algorithm for Regressors’ Selection (GARS)

GA starts with a set of solutions taken from a population which is constituted by chromosomes. These solutions are then used to create a new population. However, GA has an edge on traditional algorithms because of its advantages such as not needing a derivative and other supporting information, and being able to find global optimum points without being stuck with local optimum points. In GA, the search is carried out on a potential solution set and the solutions are evaluated until the best solution is found. (6)

Uses binary encoding to identify which independent variables should be included in the model. No transformation is applied to the independent variables before including them. Each GA individual consists of a string of m binary cells: if the i-th cell (i=1,...,m) has value 1, then Xi is included in the model, otherwise not. Every candidate solution is then evaluated with respect to a fitness function. The AIC criteria have been considered as possible fitness function. After randomly initializing the population and evaluating the population with respect to the chosen fitness function, the population is evolved through generations using stochastic uniform sampling selection scheme, single point crossover with pc=0.8, uniform mutation with pm=1/NBITS and direct reinsertion of the best recorded candidate solution. The algorithm stops when the population has been evolved for MAXGEN generations. The best solution is then reported. Even for a bigger search space, GARS is still capable of selecting the models with smaller AIC value than the ones selected by the other and the
complete model. In case that the expert is interested in a model with good forecasting capabilities, the model selected by GARS for AIC should be considered first. \(^{(16)}\)

5- The proposed method (2SLS-GA)

Model selection and validation has a crucial role in statistics. The selection of a statistical model usually requires a detailed a-priori analysis of the empirical framework and competence on the behalf of the researcher. At first, the researcher should specify the functional form (linear or not), the number and which variables to include in the model and the statistical distribution of the stochastic component. However, classical approaches have some shortcomings, such as the strong path-dependence and their difficulty to explore the whole models space \(^{(20)}\). In this paper we propose some evolutionary approaches, based on genetic algorithms, in order to overcome these shortcomings. Genetic algorithms allow a better exploration of the whole solution space through the evolution of a population of candidate models to the problem under investigation. The method for regression modeling based on improved genetic algorithm is proposed in first Stage of 2SLS.

We choose one representatives is Akaike information criterion (AIC)

For the regression model \(^{(12)}\)

\[
AIC = n \log \left( S_p^2 \right) + 2p
\]

Where \(n\) is the sample size;

\(P\) is the number of independent variables in the regression equation;

\(S_p^2\) is the residual variance

Thus the fitness function for model selection problem is assumed as the reciprocal of the rule function. \(^{(15)}\)

Now, consider the case where \(k\) is large. In such a case, it is often desirable and necessary to select a subset of \(K = \{1, \ldots, 2k\}\). Let \(P\) be any subset of \(K\) having \(|P| = p\) members. Let \(X_P\) be the sub matrix of \(X\) containing only those columns Whose indices are in \(P\). Using OLS method, it is then possible to estimate a new coefficient vector, \(b_P\), with the same goal of estimating dependent variables. The question then becomes how to select \(P\) so that the resulting model is in some way good or desirable. \(^{(1)}\)

The contribution of our paper to model building is a powerful procedure of selecting regressors which permits a very good model selection performance using a simple information criterion. In building a multiple regression model, a crucial problem is the selection of regressors to be included. If a lower amount of regressors are selected in the model, the estimate of the parameters will not be consistent and if a higher amount is selected, its variance will increase. \(^{(5)}\)

In our work Stage (1) of 2SLS method we regress the endogenous variables on all the predetermined variables in the system.

The GA here used for selecting predetermined variables at random and evaluation the response models by the asymptotic Information Criterion (AIC) (Akaike, 1973) to generate initial population (solutions) and select one of them at random from the best \(n\) to improve estimate the parameters of linear SEM. We can applied the set of 4 independent variables (with 2 observations) above and check the models to obtained the random solutions. For each random solution that passed the Criterion with an acceptable value with respect to all random solutions for evaluation and using two robust different criteria mean absolute percentage error (MAPE) and median absolute error (MEDAE) \(^{(3)}\) to compare.
Each individual is evaluated with respect to an objective function (the fitness function) that measures the optimality of each model respect to the problem under investigation. The population is evolved, within an elitistic schema, by using the usual genetic operators (crossover, mutation, reinsertion) until a stopping criterion is satisfied. 

5- Results

The proposed method can be implemented according to real data (the Annual Report of the Council of Economic Advisers ,2007-2018) on the income–money supply modified model with sample size equal to 48 and obtain the results by using MATLAB2017b.

5-1: Results of the first equation

**Table1:** the results of 2SLS and the proposed method (2SLS-GA) for equation (1) where max iteration =100

| Method          | Estimations       | Variables selected to measure response according to AIC | MAPE   | MEDAE  |
|-----------------|-------------------|--------------------------------------------------------|--------|--------|
| 2SLS            | $\beta_{10} = 2.3637$ | $\beta_{12} = 0.0004$ $\gamma_{11} = 0.0018$ $\gamma_{12} = 0.0014$ | All    | 1.7955 | 133.2374 |
| The proposed method | $\beta_{10} = 2.7293$ | $\beta_{12} = 0.0004$ $\gamma_{11} = 0.0011$ $\gamma_{12} = 0.0006$ | $X_1$  | 1.1871 | 102.1789 |
| 2SLS-GA         |                   |                                                        |        |        |

5-2: Results of the second equation

**Table2:** the results of 2SLS and the proposed method (2SLS-GA) for equation (2) where max iteration =100

| Method          | Estimations       | Variables selected to measure response according to AIC | MAPE   | MEDAE  |
|-----------------|-------------------|--------------------------------------------------------|--------|--------|
| 2SLS            | $\beta_{20} = -76.8828$ | $\beta_{21} = -0.0323$ $\gamma_{23} = 0.0614$ $\gamma_{24} = 0.0229$ | All    | 2.1587 | 51.8462 |
| The proposed method | $\beta_{20} = -1454.157$ | $\beta_{21} = -0.7472$ $\gamma_{23} = 1.5893$ $\gamma_{24} = 1.0112$ | $X_2$  | 2.0225 | 50.6710 |
| 2SLS-GA         |                   |                                                        |        |        |
6- Conclusions

It is clear that the results of the proposed method (2SLS-GA) are better than the results of the traditional method (2SLS) using the two criteria (MAPE and MEDAE). This means that the genetic algorithm (GA) with mutation rate equal to (0.0625) has succeeded to improving the estimators of linear SEM.

References

[1] Bradley C. Wallet, David J. Marchette, Jeffery L. Solka, and Edward J. Wegman (1996) "A Genetic Algorithm for Best Subset Selection in Linear Regression "To appear in the Proceedings of the 28th Symposium on the Interface .
[2] Damodar N. Gujarati and Dawn C. Porter (2008) " Basic Econometrics" Fifth Edition, www.mhhe.com .
[3] David A. Swanson, Jeff Tayman and T. M. Bryan (2010) " MAPE-R: A RESCALED MEASURE OF ACCURACY FOR CROSS-SECTIONAL, SUBNATIONAL FORECASTS " Riverside, CA 92521 USA ,email: David.swanson@ucr.edu .
[4] D. Michie, D.J. Spiegelhalter and C.C. Taylor(1994)" Machine Learning, Neural and Statistical Classification"MRC Biostatistics Unit, Institute of Public Health, University Forvie Site,Robinson Way, Cambridge CB2 2SR, U.K.
[5] Eduardo Acosta-González and Fernando Fernández-Rodríguez (2001) "MODEL SELECTION VIA GENETIC ALGORITHMS"JEL classification: C20; C61; C63.
[6] Emre Demir and Özge Akkuş (2015)" An Introductory Study on “How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model?” International Journal of Sciences: Basic and Applied Research (IJSBAR) ,Volume 19, No 2, pp 162-180 .
[7] Jeffrey M.wooldredge "Econometric Analysis of Cross Section and Panel Data" The MIT Press, Cambridge, Massachusetts, London, England .
[8] Jeffrey M.wooldredge "introductory econometrics a modern approach" The MIT Press, Cambridge, Massachusetts ,London, England .
[9] Jinyong Hahn and Jerry Hausman (2002) " Notes on bias in estimators for simultaneous equation models" Economics Letters 75 237–241.
[10] Jose J. López-Espín, Antonio M. Vidal b and Domingo Giménez (2012) " Two-stage least squares and indirect least squares algorithms for simultaneous equations models" Journal of Computational and Applied Mathematics 236 3676–3684 .
[11] Jose J. López-Espín a and Domingo Giménez (2012) " Obtaining simultaneous equation models from a set of variables through genetic algorithms" Procedia Computer Science 1 427–435 .
[12] KENNETH P. BURNHAM and DAVID R. ANDERSON (2004) " Understanding AIC and BIC in Model Selection" SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 2.
[13] M A IQUEBAL, PRAJNESHUand HIMADRI GHOSH (2012) "Genetic algorithm optimization technique for linear regression models with heteroscedastic errors",Indian Journal of Agricultural Sciences 82 (5): 422–5 .
[14] Mr. Manish Saraswat and Mr. Ajay Kumar Sharma (2013) "Genetic Algorithm for optimization using MATLAB", Available Online at www.ijares.info .
[15] Rahul Malhotra, Narinder Singh & Yaduvir Singh(2011) "Genetic Algorithms: Concepts, Design for Optimization of Process Controllers” Published by Canadian Center of Science and Education, www.ccsenet.org/cis .
[16] SANDRA PATERLINI and TOMMASO MINERVA (2007) "Regression Model Selection Using Genetic Algorithms".Rome PRIN,ISSN: 1790-5109 .
[17] Shi Minghua, Xiao Qingxian, Zhou Benda and Yang Feng (2017) "REGRESSION MODELLING BASED ON IMPROVED GENETIC ALGORITHM" ISSN 1330-3651 (Print), ISSN 1848-6339 (Online), DOI: 10.17559/TV-20160525104127.

[18] Sultan H. Aljahdali and Mohammed E. El-Telbany (2008) "Genetic Algorithms for Optimizing Ensemble of Models in Software Reliability Prediction" ICGST-AIML Journal, Volume 8, Issue 1.

[19] Steven C. Chapra (2012) "Applied Numerical Methods with MATLAB® for Engineers and Scientists" Third Edition, Berger Chair in Computing and Engineering Tufts University.

[20] Tommaso Minerva and Sandra Paterlini (2002) "Evolutionary approaches for statistical modeling" 0-7803-7282-4/02/$10.00 ©2002 IEEE, https://www.researchgate.net/publication/232620105.

[21] William H. Greene (2003) " ECONOMETRIC ANALYSIS " FIFTH EDITION , Upper Saddle River, New Jersey 07458.

[22] The Annual Report of the Council of Economic Advisers (2007-2018) "Economic Report of the President" https://www.whitehouse.gov/wp-content/.