VAR (Vector Auto Regression) approach for analysis of econometric series

T.M.J.A. Cooray

Abstract

Typically, economic models are fitted using least-squares regression or maximum-likelihood estimation methods. Regression estimation methods relate to one or more right-hand side (independent) variables to each left-hand side (dependent) variable. Simple time series methods fit the time series to its own past values and more complex time series methods can relate several time series to each other. The combinations of regression include time series cross-sectional methods and transfer functions. Transfer functions enable you to incorporate the variability of the right-hand side variables into the forecasts of the dependent variable for more realistic forecasts. one major issue is the tentative identification (specification) of the model. In this paper, the use of Vector Auto regression models (VAR) analysis in the identification of a reduced form of an econometric model is proposed. The VAR method is also briefly described in this section. The theoretical properties of the VAR estimates are presented in the next section. In latter part of the paper, the VAR analysis is applied to a set of simulated time series and a set of actual macroeconomic time series.

Introduction

The development and application of time series analysis in econometric fore-casting has occurred rapidly during the past two decades. In recent years, the focus in this area has shifted from univariate or single equation to multivari-ate and simultaneous equation models. In particular, there has been a great deal of study on dynamic equation systems, rational structural form models, and vector autoregressive-moving average models. The development of econometric time series modeling, "classical" econometric models are still one of the major tools used by many commercial economic forecasting firms to provide national economic

forecasts. In this section "classical" econometric models are considered as the simultaneous equation systems originally proposed by Klein (1950), and studied extensively by a number of econometricians. In typical applications of the classical econometric models, a simultaneous equation system often consists of a set of linear lag regression equations with white noise disturbances.

For typical national economic forecasting systems, the number of variables and equations included in the systems is often large. Therefore it is impossible to perform a joint parameter estimation of the full system as recommended in modern time series econometric models. Even though the classical econometric model is often referred to as a system of equations, it is important to note that in typical applications of large scale econometric models, the use of "system" or "joint model" comes in at the forecasting stage, rather than at the model estimation stage. In terms of model estimation, the ordinary least squares (OLS) method is usually applied to each equation in the system individually.

Vector auto regression (VAR) model

In the structural equation approach, the equation of model is basically using economic theory to model the behavioral relationship among the variables of interest. Unfortunately, economic theory is not often rich enough to provide a dynamic specification that identifies all of these relationships. Furthermore, estimation and inference are complicated by the fact that endogenous variables may appear on both the left and right sides of the equations in the model.

These problems lead to alternative, non-structural approaches to modeling the relationship among several variables. The vector autoregressive (VAR) is commonly used for forecasting systems of interrelated time series and for analyzing the dynamic impact of random disturbances on the system of variables. The VAR approach sidesteps the need for structural modeling by treating every variable as endogenous in the system as a function of the lagged values of all endogenous variables in the system. The term autoregressive is due to the appearance of the lagged values of the dependent variable on the right-hand side and the term vector is due to the fact that a vector of two (or more) variables is included in the system model. The mathematical representation of a VAR system is

$$[Y]_t = [A][Y]_{t-1} + \ldots + [A'][Y]_{t-k} + [e]_t \text{or}$$

$$\begin{bmatrix} Y_t^1 \\ Y_t^2 \\ Y_t^3 \\ \cdots \\ Y_t^p \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1p} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2p} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{p1} & A_{p2} & A_{p3} & \cdots & A_{pp} \end{bmatrix} \begin{bmatrix} Y_{t-1}^1 \\ Y_{t-1}^2 \\ Y_{t-1}^2 \\ \cdots \\ Y_{t-1}^p \end{bmatrix} + \ldots + \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1p} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2p} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{p1} & A_{p2} & A_{p3} & \cdots & A_{pp} \end{bmatrix} \begin{bmatrix} Y_{t-k}^1 \\ Y_{t-k}^2 \\ Y_{t-k}^2 \\ \cdots \\ Y_{t-k}^p \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ \cdots \\ e_{pt} \end{bmatrix}$$

Where p is the number of variables be considered in the system, k is the number of lags be considered in the system, $[Y]_t$, $[Y]_{t-1}$, ...$[Y]_{t-k}$ are the 1xp vector of variables, and the [A], ... and [A'] are the pxp matrices of coefficients to be estimated, $[e]_t$ is a 1xp vector of innovations that may be contemporaneously correlated but are uncorrelated with their own lagged values and uncorrelated with all of the right-hand side variables.

Since there are only lagged values of the endogenous variables appearing on the right-hand side of the equations, simultaneity is not an issue and OLS yields consistent estimates. Moreover, even though the innovations may be contemporaneously correlated, OLS is efficient and equivalent to GLS since all equations have identical repressors.

For example, suppose that p = 2 y and x are jointly determined by a VAR and let a constant be the only exogenous variable. Assuming that the VAR contains two lagged values of the endogenous variables, it may be written as

or

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} y_{t-2} \\ x_{t-2} \end{bmatrix} + \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}$$

$$\begin{cases} y_t = C_1 + \sum_{i=1}^{k} a_{1i} x_{t-i} + \sum_{i=1}^{k} b_{1i} x_{t-1} + e_{1t} \\ x_t = C_2 + \sum_{i=1}^{k} a_{2i} y_{t-i} + \sum_{i=1}^{k} b_{2i} x_{t-1} + e_{2t} \end{cases}$$

Where a_{ij}, b_{ij}, and C_i are the parameters to be estimated. Where e_i's are the

stochastic error terms and are called innovations or shocks in the language of VAR.

Before estimate the VAR, we have to decide the maximum lag lengths, k to generate the white noise of error terms. This is an empirical question. How to determine the maximum lag lengths in the VAR model. We can base on the smallest value of Akaike (AIC) or Schwarz (BIC) of the VAR to determine the appropriate lags.

Some problems with VAR modeling

A VAR model is a-theoretic because it uses less prior information. Recall that in simultaneous equation models exclusion or inclusion of certain variables plays a crucial role in the identification of the model. Because of its emphasis on forecasting, VAR models are less suited for policy analysis. Suppose you have a three-variable VAR model and you decide to include eight lags of each variable in each equation. You will have 24 lagged parameters in each equation plus the constant term, for a total of 25 parameters. Unless the sample size is large, estimating that many parameters will consume a lot of degree of freedom with all the problems associated with that.Strictly speaking, in an m-variable VAR model, all the m variables should be (joint) stationary. If they are not stationary, we have to transform (e.g., by first-differencing) the data appropriately. If some of the variables are non-stationary, and the model contains a mix of I(0) and I(1), then the transforming of data will not be easy.

Since the individual coefficients in the estimated VAR models are often difficult to interpret, the practitioners of this technique often estimate the so-called impulse response function. The impulse response function traces out the response of the dependent variable in the VAR system to shock in the error terms, and traces out the impact of such shocks for several periods in the future.

Co- integration and an error correction (EC) model

A linear combination of two or more time series will be non-stationary if one or more of them is non-stationary, and the degree of integration of the combination will be equal to that of the most highly integrated individual series. However, if there is a long-run relationship between the time series, the outcome

may be different. Consider the consumption theory, the long-run relationship in regression is

$$Consum_t = \beta_0 + \beta_1 Income_t + e_t$$

If the theory is correct, in the long-run, ignoring short-run dynamics and the differences between the permanent and actual measures of the variables, consumption and income will grow at the same rate. Thus, although the two series are nonstationary, they appear to be wandering together. For this to be possible, it must be a stationary process, if it were not, the two series could drift apart indefinitely, violating the theoretical relationship.

More generally, in the regression model,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + e_t$$

If there exists a relationship between a set of the variables, Y_t, X_{1t}, X_{2t}, ..., X_{kt}, the error terms can be thought of as measuring the deviations between the components of this model:

$$e_t = Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t} - .... - \beta_p X_{pt}$$

In the short run the divergence between the components will fluctuate, but if the model is genuinely correct there will be a limit to the divergence. Hence although the time series are non-stationary, e_t will be stationary. When two or more non-stationary time series are linked in such a way, we say that the two or more series are cointegrated. If there are more than two variable in the model, it is possible that there may be multiple cointegrating relationships, the maximum number in theory being equal to p-1.To test for cointegration, it is necessary to evaluate whether the error term is a stationary process. Therefore, the notion of cointegration, which was given a formal treatment in Engle and Granger (1987), makes regression involving I(1) variables potentially meaningful.

Testing for cointegration: augmented engle -granger (AEG) test

Because the type of model to be estimated might depend on whether a "dependent" variable may be cointegrated with an "independent" variable, it is important to test whether two or more variable are cointegrted. Engle and Grange

Cointegrating regression AEG test: For the two-variable case:

First, run the OLS regression as:

$$Y_t = \beta_0 + \beta_1 X_t + e_t$$

and obtain the estimated error terms:

$$\hat{e}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t$$

Second, estimate the following Augmented Dickey-Fuller (ADF) regression:

$$\hat{e}_t = \phi\ \hat{e}_{t-1} + \sum_{i=1}^{k} \alpha_i \Delta \hat{e}_{t-i} + u_t$$

Where k is the pre-selected order of lags for the white noise residuals. The test statistic is the t-statistic for ϕ, but the t-distribution is not appropriate, Dickey-Fuller (1979) has provided critical values.

Error correction model

In addition to learning about a potential long-run relationship between two series, the concept of cointegration enriches the kinds of dynamic models. If Y_t and X_t are I(1) process and are not cointegrated, we might estimate a dynamic model in first differences. As an example, consider the equation

$$\Delta Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j \Delta X_{1t-j} + \sum_{j=1}^{h} \alpha_j \Delta Y_{t-j} + e_t$$

Where et has zero mean given $\Delta Y_{t-1}, \ldots, \Delta Y_{t-h}, \Delta X_t, \Delta X_{t-1}, \ldots, \Delta X_{t-k}$. If we view this as a rational distributed lag model, we can find the impact propensity, long run propensity, and lag distribution for ΔY as distributed lag in ΔX. If Y and X are cointegrated, then the obtained estimated error term must be stationary, i.e., I(0). Now if we include the lagged estimated error term as

$$\Delta Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j \Delta X_{1t-j} + \sum_{j=1}^{h} \alpha_j \Delta Y_{t-j} + \delta\ Z_{t-1} + \varepsilon_t$$

Where $Z_t = \hat{e}_t = Y_t - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j \Delta X_{1t-j} - \sum_{j=1}^{h} \hat{\alpha}_j \Delta Y_{t-j}$ is the one-period lagged value of the estimated error of the cointegrating regression obtained from OLS estimation, this term is called the error correction term. The principle behind this model is that there often exists a long run equilibrium relationship between two economic variables. In the short run, however, there may be disequilibrium. With the error correction mechanism, a proportion of the disequilibrium is corrected in the next period. The error correction process is thus a means to reconcile short-run and long run behavior.

Therefore, in the error correction model, the right hand side contains the short-run dynamic coefficients (i.e., α_i, β_i) as well as the long-run coefficient (i.e., δ). The absolute value of δ decides how quickly the equilibrium is restored.

The error correction model of the consumption function becomes:

$$\Delta C_t = \beta_0 + \sum_{j=1}^{k} \beta_j \Delta Y_{t-1} + \sum_{j=1}^{h} \alpha_j \Delta C_{t-j} + \delta Z_{t-1} + e_t$$

The error correction term, $Z_t = C_t - \sum_{i=0}^{a} \gamma_i Y_{t-i} - \sum_{i=1}^{b} \vartheta_i C_{t-i}$, is obtained from the OLS regression.

National economic time series of Sri Lanka

In this section, we study a macro econometric forecasting system employed by the Accounting and Statistics of Annul Bulletin published by Central Bank of Sri Lanka. We shall use this study to illustrate the problems with the inclusion of contem-poraneous input variables in the identification of a transfer function model and their effects on the forecasting performance of the model. Following the methods are developed by the author. The data used in this study began in the first quarter of 1991 and ended in the fourth quarter of 2008, a total of 72 observations for each variable. Among the 72 observations in each series, the first 68 observations will be used for model identification and parameter estimation. The last 4 observations will be used solely for the comparison of the forecasting performance of the models. One of the most important applications of this econometric model is to forecast the quarterly economic time series of Sri Lanka.Among the dependent variables in the behavioral equations of the econometric model, we shall only include 6 of them in this study. For convenience of reference, the abbreviations and definitions of the 6 dependent variables and their relevant explanatory variables are listed below. The time series plots for the

6 dependent variables are shown in Figure 1. As shown in the time series plots, all the series to be studied are non stationary and possess strong seasonality. Since the series KBF, M, X, and MON have greater variability over time, logarithmic transformed data will be employed in this study.

Interested economical variables

M Imports of goods and services, in constant million rupees

X Exports of goods and services, in constant million rupees

MON Money demand, in constant million rupees

E Foreign exchange rate (Rs/US$)

GDP Gross domestic product, in constant million Rs

GNP Gross national product, in constant million Rs

KBF Fixed capital stock on private sector, in constant million Rs

Model 1

KBF Fixed capital stock on private sector, in constant million Rs

$\ln(KBF) = f(\ln(X_{-1}), \ln(X_{-2}), \ln(GDP))$

Model 2

Imports of goods and services (M)

$\ln(M) = f(\ln(X), \ln(X_{-1}), \ln(E), \ln(E_{-1}))$

Model 3

Exports of goods and services

$\ln(X) = f(\ln(E) \ln(M_{-1}), \ln(X_{-4}))$

Criterion of model performance

To evaluate the performance of different models, we shall employ the root mean squared error (RSME) for within-sample and post-sample of each equation. The RMSE is defined as

$$RMSE=\left\{\frac{1}{m}\sum_{t=1}^{m}(y_t-\hat{y}_t)^2\right\}^{1/2}$$

where $\hat{y}_t$ is the fitted or predicted value of y_t based on an estimated model, and m is the number of observations used in the computation. According to the definition of RMSE, the within-sample RMSE is an estimate of the standard deviation of random errors if the parameter estimates of the model are unbi-ased, and the post-sample RMSE is a measure of forecast performance using the estimated model.

Conclusion

The analyses shown above reveal several interesting points that are worth further discussion. When the VAR analysis is employed, a number of transfer function models degenerate to ARIMA models. This result indicates that the association between the explanatory variables and the dependent variable is not as strong as the original hypotheses of the models suggested (or what the classical regression models indicated). This is not too surprising if we take the economic environment of Sri Lanka into consideration. Sri Lankan has a highly regulated economy. A number of foreign and domestic events also have had important impacts on Sri Lankans' economy. All these factors contribute to major disturbances which might weaken the potential relationships between the dependent variables and their explanatory variables. As the free economic environment becomes more mature and the political situation becomes more stable, we may find transfer function models more useful in modelling Sri Lankan economic time series.

References

Box, G.E.P. and Jenkins, G.M. (1971). ***Time series analysis : forecasting and control.*** Holden Day, San Francisco.

Central Bank of Sri Lanka, ***Economic progress of independent Sri Lanka,*** Central Bank of Sri Lanka, Annual Bulletin.

Dickey and Fuller, 1979, ***Distribution of the estimators for autoregressive time series with a unit root,*** Journal of the American Statistical Association, 74.

Engle, R.F. (1982). ***Autoregressive conditional heteroscedasticity with estimates of the***

variance of United Kingdom inflation. Econometrica,.50, p,987-1007.

Engle and Granger, (1987), ***Cointegration and error correction: representation, estimate, and testing,*** Econometrics 55, p,251-276.

Gujarathi,D.M, Sangeetha, (2007), ***Basic econometrics***, Tata McGraw-Hill Publishers, New Delhi.

Klein, L. R. (1950). ***Economic fluctuations in the United States,*** 1921-41. Cowless Commission Monograph 11. John Wiley, New York.