FORECASTING THE LEVELS OF VECTOR AUTOREGRESSIVE LOG-TRANSFORMED TIME SERIES

Miguel A. Ariño[1*] Philip Hans Franses[2].

November 20, 1996

[1] IESE, Universidad de Navarra
Avda. Pearson 21
08034 Barcelona, Spain
aarino@iese.es

[2] Econometric Institute and Rotterdam Institute for Business Economic Studies
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
franses@few.eur.nl

Econometric Institute Report 9669/A
http://www.eur.nl/few/ei/reports/

Abstract

In this paper we give explicit expressions for the forecasts of levels of a vector time series when such forecasts are generated from (possibly cointegrated) vector autoregressions for the corresponding log-transformed time series. We also show that simply taking exponentials of forecasts for logged data leads to substantially biased forecasts. We illustrate this using a bivariate cointegrated vector series containing US GNP and investments.

Key Words: VAR time series, log-transformation, forecasting

1 Introduction

In the empirical time series analysis of economic variables it is common practice to transform the data using natural logarithms prior to the construction of econometric models that are often used for forecasting. Some of the motivations for this strategy are that this log-transformation reduces the impact of outliers, that first differenced log-transformed data correspond to growth rates, and that

*The first author has received partial financial support from CIIF, Centro Internacional de Investigaciones Financieras (International Center for Financial Research)

it reduces the often observed increasing variance of trending time series. Once a model has been constructed, and the parameters have been estimated, one can make forecasts for the log-transformed data. In some cases, however, one is interested in the forecasts of the levels of the time series (i.e. the untransformed data) instead of (functions of the) log-transformed data. In that case, as is well known from the results in Granger and Newbold (1976) for univariate time series, simply taking exponentials of the forecasts of the logged data yields biased forecasts. For the class of the univariate autoregressive [AR] model, Granger and Newbold (1976) derive expressions for unbiased forecasts of the levels. In the present paper we extend their results to the practically very relevant class of vector autoregressive [VAR] time series models. VAR models are often used in empirical economics to generate out-of-sample forecasts since their parameters are easy to estimate, and especially since such models provide a simple framework for the analysis of cointegration, see for example Johansen (1995). In the first part of Section 2 of our paper, we give explicit expressions for the out-of-sample forecasts of the levels of m time series when these series (in log-transformed format) are modeled by a VAR model of order p. To illustrate the details of our results, we present an example for two series in case $p = 1$ in the second part of Section 2. In Section 3, we give an empirical example concerning a bivariate US series containing GNP and investments, where we take into account that the log-transformed series are cointegrated. We conclude our paper in Section 4 with some remarks.

2 Forecasting levels

In this section we present explicit expressions for the forecasts of the levels of a time series, when the log-transformed vector time series follows a vector autoregressive model. To motivate our paper, consider the univariate time series X_t, for which one analyses Y_t with the latter being the series in logs, that is, $Y_t = \log X_t$, where log denotes the natural logarithmic transformation. Suppose that the log-transformed series can be modelled as $Y_t = M_t + \eta_t$ where M_t denotes the conditional expectation of Y_t, given the information set at time t, and where η_t is a standard white noise prcess. One may now want to use the so-called naive forecast of X_{t+k}, that is, the exponential of the forecast of Y_{t+k}:

$$\widehat{X}^*_{t+k} = \exp(\widehat{M}_{t+k}).$$

However, since the seminal work in Granger and Newbold (1976), we know that this forecast is not the expected value at time t of X_{t+k} which would be the unbiased forecast $\widehat{X}_{t+k}$ of X_{t+k}. In fact, the latter equals

$$\widehat{X}_{t+k} = \mathrm{E}_t[\exp(M_{t+k} + \eta_{t+k})].$$

where E_t is the expectation operator at time t. The naive forecast is seen to be biased since the expected value of the exponential of the white noise process is unequal to zero.

In this section we first develop expressions for the unbiased k-step ahead forecast of a m-dimensional time series, of which the log-transformed series follows a VAR(p) model and we show how the naive forecasts are to be corrected to obtain unbiased forecasts. Next, as an example, we give the expressions for $m = 2$ and $p = 1$ for illustrative purposes.

2.1 Forecasting an m-dimensional level time series

Let $\mathbf{X}(t)$ be an m-dimensional vector time series $\mathbf{X}'(t) = (X_1(t), \cdots, X_m(t)$ such that $\mathbf{Y}(t)$ with $\mathbf{Y}'(t) = (Y_1(t), \cdots, Y_m(t))$ with $Y_j(t) = \log X_j(t)$, follows the VAR(p) model

$$\mathbf{Y}(t+1) = \mathbf{B_0} + \sum_{\mathbf{r}=1}^{p} \mathbf{B_r}\mathbf{Y}(t-r+1) + \eta(t+1)$$

where $\mathbf{B_0} = (b_1, \cdots, b_m)$ and $\mathbf{B_r} = (b_{ijr})_{i,j=1}^{m}$ are an m-dimensional vector and matrix with constant parameters, and $\eta'(t) = (\eta_1(t), \cdots, \eta_m(t))$ is a vector of m normally identically and independently distributed random variables with mean zero and covariance matrix $\mathbf{V}$.

For each variable $1 \leq i \leq m$, we have that

$$\begin{aligned} Y_i(t+1) &= b_i + \sum_{j=1}^{p} b_{ij1} Y_j(t) + \sum_{j=1}^{p} b_{ij2} Y_j(t-1) + \cdots \\ &\quad + \sum_{j=1}^{p} b_{ijp} Y_j(t-p+1) \;\; + \;\; \eta_i(t+1) \\ &= c_{0i}(1) + \sum_{j=1}^{p} c_{ij1}(1) Y_j(t) + \cdots + \sum_{j=1}^{p} c_{ijp}(1) Y_j(t-p+1) \\ &\quad + \sum_{j=1}^{m} d_{ij}(1) \eta_j(t+1). \end{aligned} \tag{1}$$

where

$$\begin{aligned} c_{0i}(1) &= b_i \\ c_{ijl}(1) &= b_{ijl} \end{aligned}$$

and

$$d_{ij}(1) = \begin{cases} 1, & for \quad j = i; \\ 0, & \text{otherwise.} \end{cases}$$

In a similar way, we have that

$$Y_i(t+2) = b_i + \sum_{j=1}^{p} b_{ij1} Y_j(t+1) + \cdots + \sum_{j=1}^{p} b_{ijp} Y_j(t-p+2) \;\; + \;\; \eta_i(t+2).$$

Substituting $Y_i(t+1)$ for its value according to (1) we obtain

$$\begin{aligned} Y_i(t+2) &= c_{0i}(2) + \sum_{j=1}^{p} c_{ij1}(2) Y_j(t) + \cdots + \sum_{j=1}^{p} c_{ijp}(2) Y_j(t-p+1) \\ &\quad + \sum_{j=1}^{m} d_{ij}(2) \eta_j(t+1) + \sum_{j=1}^{m} d_{ij}(1) \eta_j(t+2) \end{aligned}$$

where

$$c_{0i}(2) = b_i + \sum_{j=1}^{m} b_{ij1} b_j$$

$$c_{ijl}(2) = \sum_{r=1}^{m} b_{ir1} c_{rjl}(1)$$

and

$$d_{ij}(2) = \sum_{r=1}^{m} b_{ir1} d_{rj}(1)$$

In order to simplify notation, let us call $\mathbf{C_0}(k)$ the vector column $\mathbf{C_0}(k)' = (c_{0i}(k))_{i=1}^{m}$, and $\mathbf{C_l}(k)$ and $\mathbf{D}(k)$ the matrix $(c_{ijl}(k))_{i,j=1}^{m}$ and $(d_{ij}(k))_{i,j=1}^{m}$, respectively.

Calculating in a similar way $Y_i(t+3), \cdots, Y_i(t+k)$, we arrive at the expression

$$\begin{aligned} Y_i(t+k) &= c_{0i}(k) + \sum_{l=1}^{p} \sum_{j=1}^{p} c_{ijl}(k) Y_j(t-l+1) \\ &\quad + \sum_{r=1}^{k} \sum_{j=1}^{m} d_{ij}(k-r+1) \eta_j(t+r) \end{aligned}$$

where

$$\mathbf{C_0}(k) = \mathbf{B_0} + \sum_{i=1}^{p} \mathbf{B_i} \mathbf{C_0}(k-i)$$

$$\mathbf{C_l}(k) = \sum_{i=1}^{p} \mathbf{B_i} \mathbf{C_l}(k-i)$$

$$\mathbf{D}(k) = \sum_{i=1}^{p} \mathbf{B_i} \mathbf{D}(k-i)$$

with the initial conditions

$$\mathbf{C_0}(j) = \mathbf{0} \qquad \text{for } j = 0, -1, \cdots, -p+1$$

$$\mathbf{C_l}(j) = \begin{cases} \mathbf{I_m}, & if \quad j = 1 - l; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \qquad \begin{array}{l} \text{for} \quad 1 \leq l \leq p \text{ , and} \\ j = 0, -1, \cdots, -p+1 \end{array}$$

$$\mathbf{D}(j) = \mathbf{0} \qquad \text{for } j = 0, -1, \cdots, -p+1.$$

The above expressions can be used to obtain forecasts of the untransformed level time series. The naive forecast of $X_i(t+k)$ is

$$\widehat{X}_i^*(t+k) = \exp[\mathbf{E}_t(Y_i(t+k))] = \prod_{l=1}^{p}\prod_{j=1}^{p} X_j(t-l+1)^{c_{ijl}(k)} \exp(c_{0i}(k))$$

This expression gives us the exponential of the k-step ahead forecast of the log-transformed series $Y_i(t+k)$ according to the specified VAR(p) model. However, the unbiased forecast of $X_i(t+k)$ is

$$\widehat{X}_i(t+k) = E[\exp(Y_i(t+k))] = \widehat{X}_i^*(t+k) \exp(e_i(k)/2)$$

where

$$e_i(k) = e_i(k-1) + (d_{i1}(k), \cdots, d_{im}(k))\mathbf{V}(d_{i1}(k), \cdots, d_{im}(k))'$$

with the initial condition $e_i(0) = 0$. Notice that in practice one needs to estimate $\mathbf{V}$ and all the other parameters.

2.2 An example

In order to provide some intuition for the expressions in the previous subsection, consider the particular example in which $m = 2$ and $p = 1$.

Let $(X_1(t), X_2(t))$ be a vector time series and that $(Y_1(t), Y_2(t))$ with $Y_i(t) = \log X_i(t)$ obeys

$$\begin{pmatrix} Y_1(t+1) \\ Y_2(t+1) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix} + \begin{pmatrix} \eta_1(t+1) \\ \eta_2(t+1) \end{pmatrix}$$

or equivalently, using the notation of the previous section

$$\mathbf{Y}(t+1) = \mathbf{B_0} + \mathbf{B_1}\mathbf{Y}(t) + \eta(t+1)$$

where $\eta(t) = (\eta_1(t), \eta_2(t))'$ is a vector of two normally identically and independently distributed random variables with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}$$

Notice that we have dropped the third subscript to the elements of matrix $\mathbf{B_1}$ since $p = 1$.

It is now easy to verify that

$$\begin{aligned} Y_1(t+2) \quad = \quad & b_1 + b_{11}b_1 + b_{12}b_2 + (b_{11}^2 + b_{12}b_{21})Y_1(t) + (b_{11}b_{12} + b_{12}b_{22})Y_2(t) + \\ & b_{11}\eta_1(t+1) + b_{12}\eta_2(t+1) + \eta_1(t+2) \end{aligned}$$

and

$$\begin{aligned} Y_2(t+2) &= b_2 + b_{21}b_1 + b_{22}b_2 + (b_{21}b_{11} + b_{22}b_{21})Y_1(t) + (b_{12}b_{21} + b_{22}^2)Y_2(t) + \\ & \quad b_{21}\eta_1(t+1) + b_{22}\eta_2(t+1) + \eta_2(t+2) \end{aligned}$$

and in general, expressing $Y_i(t+3), \cdots, Y_i(t+k)$ in terms of $Y_1(t)$, $Y_2(t)$ and the errors $\eta_i(t+j)$ we get

$$Y_1(t+k) = c_{01}(k) + c_{11}(k)Y_1(t) + c_{12}(k)Y_2(t)$$

$$+d_{11}(k)\eta_1(t+1) + d_{11}(k-1)\eta_1(t+2) + \cdots + d_{11}(1)\eta_1(t+k)$$

$$+d_{12}(k)\eta_2(t+1) + d_{12}(k-1)\eta_2(t+2) + \cdots + d_{12}(1)\eta_2(t+k)$$

and

$$Y_2(t+k) = c_{02}(k) + c_{21}(k)Y_1(t) + c_{22}(k)Y_2(t)$$

$$+d_{21}(k)\eta_1(t+1) + d_{21}(k-1)\eta_1(t+2) + \cdots + d_{21}(1)\eta_1(t+k)$$

$$+d_{22}(k)\eta_2(t+1) + d_{22}(k-1)\eta_2(t+2) + \cdots + d_{22}(1)\eta_2(t+k)$$

where

$$\begin{pmatrix} c_{01}(k) \\ c_{02}(k) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} c_{01}(k-1) \\ c_{02}(k-1) \end{pmatrix}$$

$$\begin{pmatrix} c_{11}(k) & c_{12}(k) \\ c_{21}(k) & c_{22}(k) \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} c_{11}(k-1) & c_{12}(k-1) \\ c_{21}(k-1) & c_{22}(k-1) \end{pmatrix}$$

$$\begin{pmatrix} d_{11}(k) & d_{12}(k) \\ d_{21}(k) & d_{22}(k) \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} d_{11}(k-1) & d_{12}(k-1) \\ d_{21}(k-1) & d_{22}(k-1) \end{pmatrix}$$

with the initial conditions

$$\begin{pmatrix} c_{01}(0) \\ c_{02}(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} c_{11}(0) & c_{12}(0) \\ c_{21}(0) & c_{22}(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} d_{11}(0) & d_{12}(0) \\ d_{21}(0) & d_{22}(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The naive forecast of $X_i(t+k)$ is $\widehat{X}_i^*(t+k) = \exp[\mathrm{E}_t(\widehat{Y}_i(t+k)]$, which is

$$\widehat{X}_1^*(t+k) = X_1(t)^{c_{11}(k)} X_2(t)^{c_{12}(k)} \exp(c_{01}(k))$$

$$\widehat{X}_2^*(t+k) = X_1(t)^{c_{21}(k)} X_2(t)^{c_{22}(k)} \exp(c_{02}(k)),$$

while for the unbiased forecast of $X_i(t+k)$ we obtain

$$\widehat{X}_1(t+k) = \widehat{X}_1^*(t+k)\exp(e_1(k)/2)$$

and

$$\widehat{X}_2(t+k) = \widehat{X}_2^*(t+k)\exp(e_2(k)/2)$$

where

$$e_i(k) = e_i(k-1) + (d_{i1}(k), d_{i2}(k))\mathbf{V}(d_{i1}(k), d_{i2}(k))' =$$
$$e_i(k-1) + d_{i1}^2(k)\sigma_{11}^2 + 2d_{i1}(k)d_{i2}(k)\sigma_{12}^2 + d_{i2}^2(k)\sigma_{22}^2$$

3 An Application

In this section we apply the expressions obtained in the previous section to a two-dimensional time series $(X_1(t), X_2(t))'$. X_1 is the real GNP of the US and X_2 is the real gross domestic investment series. The data are given in Pindyck and Rubinfeld (1991, chapter 12). Quarterly observations are available from the first quarter of 1947 until the first quarter of 1988. We will use observations until the fourth quarter of 1980 to estimate our VAR model for the log-transformed data, and we leave the remaining 29 data points to evaluate our naive and unbiased forecasts. We find that a VAR model of order 3 fits the data well. Since the logged series appear to be cointegrated according to several of the currently available tests, the cointegrating relationship between both series is imposed in the VAR, that is, we obtain a VAR model with (nonlinear) parameter restrictions.

The estimated model for the log-transformed series $(Y_1(t), Y_2(t))'$ with $Y_i(t) = \log X_i(t)$ is

$$\Delta_1 Y_1(t) = \underset{(0.024)}{-0.075} + \underset{(0.089)}{0.430}\,\Delta_1 Y_1(t-1) + \underset{(0.095)}{0.257}\,\Delta_1 Y_1(t-2) - \underset{(0.013)}{0.043}\,Z(t-1) + \eta_1(t)$$

$$\Delta_1 Y_2(t) = \underset{(0.117)}{-0.593} + \underset{(0.436)}{1.842}\,\Delta_1 Y_1(t-1) + \underset{(0.090)}{0.216}\,\Delta_1 Y_2(t-2) - \underset{(0.064)}{0.325}\,Z(t-1) + \eta_2(t)$$

, where standard errors are given in parentheses. The cointegrating relationship is

$$Z(t) = Y_1(t) - Y_2(t)$$

and the estimated covariance matrix

$$\operatorname{cov}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}0.0001041 & 0.0004112\\ 0.0004112 & 0.0025960\end{pmatrix}.$$

This bivariate error correction model can be expressed as a VAR(3) model with parameter restrictions as follows:

$$\begin{pmatrix} Y_1(t) \\ Y_2(t) \end{pmatrix} = \begin{pmatrix} -0.075 \\ -0.593 \end{pmatrix} + \begin{pmatrix} 1.473 & -0.043 \\ 2.167 & 0.675 \end{pmatrix} \begin{pmatrix} Y_1(t-1) \\ Y_2(t-1) \end{pmatrix}$$

$$+ \begin{pmatrix} -0.173 & 0 \\ -1.842 & 0.216 \end{pmatrix} \begin{pmatrix} Y_1(t-2) \\ Y_2(t-2) \end{pmatrix}$$

$$+ \begin{pmatrix} -0.257 & 0 \\ 0 & -0.216 \end{pmatrix} \begin{pmatrix} Y_1(t-3) \\ Y_2(t-3) \end{pmatrix} + \begin{pmatrix} \eta_1(t) \\ \eta_2(t) \end{pmatrix}.$$

A summary of the errors obtained for the naive and the appropriate unbiased forecasts of X_1 and X_2 is presented in Table 1. When we compare the naive forecasts with the unbiased forecasts for the restricted VAR(3) model, we rapidly notice the better performance of the unbiased forecasts with respect to the naive forecasts. Not only the mean absolute error, mean percentage absolute error, and mean squared error are smaller for the unbiased forecast in both series, but it also appears for example that for GNP, the unbiased forecasts outperform the naive forecasts by 18 times to 11. Using a binomial distribution of parameters $n = 29$ and $p = 0.5$ this is significant at the 7% level. For the investments series the score is 20 to 9 which is significant even at the 2% level. In addition, and as expected, the outperformance of the unbiased forecasts is obvious especially for the long-term. For the GNP series all of the last 17 forecasts are better for the unbiased than for the naive forecasts. For the investment series we find that 16 of the last 17 forecasts are better using the unbiased method. This means that the farther the time horizon the better is the performance of the unbiased forecast with respect to the naive. In Table 2 we report the same statistics as in Table 1 but using only the forecasts from 10 to 29 periods ahead (20 forecasts). Hence, for the longer horizons one clearly needs the use of the unbiased forecasts.

Table 1. Evaluation of 29 quarters out-of-sample forecasting performance for naive and unbiased forecasts from a VAR(3) model for log-transformed series for the untransformed level time series.

	log-model X_1		*log*-model X_2	
	naive	unbiased	naive	unbiased
ME	-9.720	-15.1593	4.906	0.557
MAE	93.474	89.587	55.652	53.659
MAPE	2.72%	2.62%	10.20%	9.96%
MSE	12248	11873	4475	4395
RMSE	110.7	109.0	66.9	66.3

The forecast errors are defined as the true value minus the forecasted value. Forecast evaluation criteria are mean error, ME, mean absolute error, MAE, mean average percentage error MAPE and (root) mean squared error, (R)MSE.

Table 2. Evaluation of 20 quarters out-of-sample forecasting performance from 10 to 29 periods ahead for naive and unbiased forecasts from a VAR(3) model for log-transformed series for the untransformed level time series.

	log-model X_1		*log*-model X_2	
	naive	unbiased	naive	unbiased
ME	84.328	74.508	42.928	37.442
MAE	75.529	69.099	47.385	43.633
MAPE	2.04%	1.87%	7.37%	6.82%
MSE	2511	1979	833	679
RMSE	50.1	44.5	28.9	26.0

These statistics are defined in Table 1

4 Concluding remarks

In this paper we presented explicit expressions for forecasts for the levels of a vector time series when a VAR model was used for the log-transformed data. We showed that exponentials of the forecasts for logged data are biased, as could also be observed from our empirical forecasts from a bivariate cointegrated vector autoregressive time series model containing US GNP and investment. Our results can be practically relevant in case one aims to forecast the levels of a vector time series. Because multi-step forecasts are linked with impulse-response functions, see Lutkepohl (1991), an extension of our results to these functions seems also relevant. Finally, our expressions can be useful to properly evaluate forecasts from VAR models for logged data versus such models for untransformed data. In fact, it may sometimes be unclear from the outset whether taking logs amounts to the best empirical strategy.

References

[1] Granger, C.W.J. and P. Newbold (1976), Forecasting Transformed Series, *Journal of The Royal Statistical Society* B, 38, 189-203.

[2] Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford: Oxford University Press.

[3] Lutkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Berlin: Springer Verlag.

[4] Pindyck R.S. and D.L. Rubinfeld (1991). *Econometric models and economic forecasts*. McGraw-Hill International Editions, Economic Series.