ISSN: 1978 - 8282

AN EXPERIMENTAL STUDY ON BANK FORECASTING USING REGRESSION DYNAMIC LINIER MODEL

Wiwik Anggraeni[1]
Danang Febrian[2]

e-mail : anggraeni@gmail.com, danang_febrian@yahoo.co.id

Diterima : 20 Mei 2011/ Disetujui : 24 Juni 2011

ABSTRACT

Nowadays, forecasting is developed more rapidly because of more systematicaly decision making process in companies. One of the good forecasting characteristics is accuration, that is obtaining error as small as possible. Many current forecasting methods use large historical data for obtaining minimal error. Besides, they do not pay attention to the influenced factors. In this final project, one of the forecasting methods will be proposed. This method is called Regression Dynamic Linear Model (RDLM). This method is an expansion from Dynamic Linear Model (DLM) method, which model a data based on variables that influence it. In RDLM, variables that influence a data is called regression variables. If a data has more than one regression variables, then there will be so many RDLM candidate models. This will make things difficult to determine the most optimal model. Because of that, one of the Bayesian Model Averaging (BMA) methods will be applied in order to determine the most optimal model from a set of RDLM candidate models. This method is called Akaike Information Criteria (AIC). Using this AIC method, model choosing process will be easier, and the optimal RDLM model can be used to forecast the data. BMA-Akaike Information Criteria (AIC) method is able to determine RDLM models optimally. The optimal RDLM model has high accuracy for forecasting. That can be concluded from the error estimation results, that MAPE value is 0.62897% and U value is 0.20262.

Keyword : Forecasting, Regression variables, RDLM, BMA, AIC

1. **Information System Department, Institut Teknologi Sepuluh Nopember**
   Kampus Keputih, Sukolilo, Surabaya 60111, Indonesia
2. **Information System Department, Institut Teknologi Sepuluh Nopember**
   Kampus Keputih, Sukolilo, Surabaya 60111, Indonesia

INTRODUCTION

Nowadays, forecasting has developed more rapidly because of the more systematically decision making in a organization or company. One of the good forecasting characteristic is from accuration, and should get error that is as minimal as possible. Usually, forecasting just estimates based on historical data only without considering external factors that might influence the data. Because of that, in this paper will be proposed a method that takes all external factors into consideration, this method is called Regression Dynamic Linear Models (RDLM), with Bayesian Model Averaging (BMA) applied in order to choose the most optimal model. By using this method, the forecast results will have high accuracy. (Mubwandarikwa et al., 2005).

The Method

There are four steps to forecast a data using RDLM method, i.e. : forming candidate models, choosing optimal model, forecasting using optimal model, and measuring accuracy of optimal model.

Dynamic Linear Models (DLM)

Dynamic Linear Model is an extension of state-space modeling on prediction and dynamic system control (Aplevich, 1999). State-space model of time series contains data generating process with state (usually shown by vector of parameter) that can change over time. This state is only observed indirectly, as far as values of time series that are obtained as function of state in correspond period. DLM base model at all time t is described by evolution / system and observation equation. The equation forms are as follow :

o Observation equation :

$$Y_t = F_t\theta_t + v_t,$$

where $v_t \sim N[0, V_t]$ (1)

o System equation :

$$\theta_t = G_t\theta_{t-1} + \omega_t,$$

where $\omega_t \sim N[0, W_t]$ (2)

o Initial information :

$$\theta_0 \sim N[m_0, C_0] \qquad (3)$$

DLM can be explained alternatively with 4 sets as follow :

$$M_t(j_t) = \{F_t, G_t, V_t, W_t\}_j \quad j = 1,2,... \quad (4)$$

Where at time t, :

- o θ_t is state vector at time t.
- o F_t is known regression variable vector.
- o v_t is observation noise that has Gaussian distribution with zero mean and known variance V_t, where it represents estimation and error trial of changing observation of Y_t.
- o G_t is state evolution matrix, it describes deterministic mapping of state vector between time t – 1 and t.
- o ω_t is evolution noise that has Gaussian distribution with zero mean and variance matrix Wt, where it represents changing in state vector.

Regression Dynamic Linear Models (RDLM)

Regression Dynamic Linear Model (RDLM) is an extension of DLM, which RDLM considers regression variables (regressor) in modeling process. For example, for time series data that has regressors X1, X2, then it will have several possible models, that are M1(,X1), M2(,X2) and M3(,X1,X2). For time t, t = 1,2,… Regression Dynamic Linear Model (RDLM), ,(j = 1,2,...,k), represents a base time series model with 4 observations, which can be identified by 4 sets, where :

- o $F_j = (X_1,...,Xp)_j$ is regression vector *(1 x p)*, X_{ij} is ith variable regression (i =1,2,...,p) which for X_1 has value of 1.
- o G_j is system evolution matrix *(p x p)* with the value of $G_j = I(n)$ identity matrix.
- o V_j is observation variance of .
- o is system evolution variance matrix *(p x p)* which is estimated using discount factors, for ith time :

$$W_{jt} = \frac{1-\delta}{\delta} C_t \quad (5)$$

Discount factors are determined by checking off model to determine the optimal values. Optimal value for trend component $\delta_T = 0.9$, seasonality $\delta_S = 0.95$, variance $\delta_W = 0.99$ and regression $\delta_R = 0.98$ (Mubwandarikwa et al., 2005).

RDLM Sequential Updating

Estimation of state variables (θ) can not be done directly at all times, but by using information from data which update from time t-1 to t is performed using Kalman Filter. For further information, see West and Harrison (1997).

Take as example D_t describes all information from past times until time t and Y_t is data at time t.

Assume that :

$$\theta_{t-1} \mid D_{t-1} \sim N(m_{t-1}, C_{t-1}) \qquad (6)$$

Equation (2) and (6) have Gaussian distribution, so linear combinations of both of them can be formed and produce prior distribution that is :

$$\theta_t \mid D_{t-1} \sim N(G_t m_{t-1}, G_t C_{t-1} G_t' + W_t) \quad (7)$$

Then from equation (1) and (7), forecast distribution can be obtained, that is :

$$Y_t \mid D_{t-1} \sim N(F_t G_t m_{t-1}, F_t R_t F_t' + V_t) \quad (8)$$

where

$$R_t = G_t C_{t-1} G_t' + W_t$$

From forecast distribution at equation (8), forecast result for Y_t can be obtained using :

$$\hat{y} = E(Y_t \mid D_{t-1}) = F_t G_t m_{t-1} \qquad (9)$$

By using Kalman Filter, posterior distribution can be obtained :

$$\theta_t \mid D_t \sim N(m_t, C_t) \qquad (10)$$

where

$$m_t = G_t m_{t-1} + A_t e_t \qquad C_t = R_t - A_t B_t A_t'$$

with $A_t = R_t F_t' B_t^{-1}$ $\qquad e_t = y_t - \hat{y}_t$ $\qquad B_t = F_t R_t F_t' + V_t$

All the steps above solve recursive update of RDLM and can be summaried as following :

1. determining model by choosing $[F, G, V, W]_t$.
2. setting initial values of m_0, c_0.
3. forecasting y_{t+1} using equation (9) .

4. observing y_{t+1} and updating using equation (10).
5. back to (c), then substituting t+1 with t.

Bayesian Model Averaging of RDLM

In RDLM method, there are many candidate models. For determining the most optimal model, one of BMA method is used, that is Akaike Information Criteria (AIC).

Akaike Information Criteria (AIC)

Akaike Information Criteria (AIC) by Akaike (1974) originates from maximum (log-)likelihood estimate (MLE) from error variance of Gaussian Linear regression model. Maximum (log-) likelihood model can be used to estimate parameter value in classic linier regression model. AIC suggests that from a class of candidate models, choose model that minimize :

$$AIC = -2\ln Lj + 2p \quad (11)$$

Where for j[th] model :

- o L_j is likelihood.
- o p is number of parameters in model.

This method chooses model that gives best estimates asymptotically (Akaike, 1974) in explanation of Kullback-Leibler. Akaike weight can be estimated by defining:

$$\Delta_j = AIC_j - \min(AIC) \quad (12)$$

where *minAIC* is the smallest value of AIC in a set of models. Likelihood L_j from every model $M_j(j)$ conditional on data and set of models. Then Akaike weight w_j can be estimated using equation :

$$w_j = \frac{e^{\frac{-\Delta_j}{2}}}{\sum_j^k e^{\frac{-\Delta_j}{2}}} \quad (13)$$

where k is number of possible models in consideration and the rest of defined models component. (Turkheimer et al., 2003)

Error Estimation

For knowing the accuration of forecasting model, it can be seen from error estimation result. According to Makridakis *et al.*, 1997, several methods in forecast error estimation that can be used are as following :

o *Mean Absolute Percentage Error* (MAPE)

MAPE is differences between real data and forecast result that is divided with forecast result then is absoluted and the result is on percent value. A model has excellent performance if MAPE value lies under 10%, and good performance if MAPE value lies between 10% and 20% (Zainun dan Majid, 2003).

$$MAPE = \frac{1}{n}\sum_{1}^{t}\left|\frac{Y_t - \hat{Y}_t}{Y_t} * 100\right| \quad (14)$$

o *Theil's U statistic*

U statistic is performance comparation between a forecasting model with naïve forecasting, that predicts future value is equivalent with real value one time before. Comparation takes correspond ratio with RMSE (root mean squared errors), that is square root of average squared differences between prediction and observation. As the main rule, forecasting method that has Theil's U value larger than 1 is not effective.

$$U = \sqrt{\frac{\sum_{1}^{t}(\hat{Y}_t - Y_t)^2}{\sum_{1}^{t}(Y_t - Y_{t-1})^2}} \quad (15)$$

where for all methods,

Y = data, $\hat{Y}$ = forecasting result.

Implementation and Analysis

Several trial test that have been done are choosing optimal model, forecasting optimal model, testing AIC performance and comparing DLM with RDLM.

To do the trial tests, world commodity price index data is used. This data contains many kinds world commodity including food, gas, agriculture, and many kinds of metal. Several variables used are :

1. Rice price index (D).
2. Fertilizer price index (X1).
3. Agriculture tools price index (X2).
4. Refined fuel oil price index (X3).

This data is from 1980 until 2001. The target forecast data is the first variable that is rice price index, with regression variables fertilizer price index, agriculture tools price index, and refined fuel oil price index. Plot of rice price index is shown on figure 1.

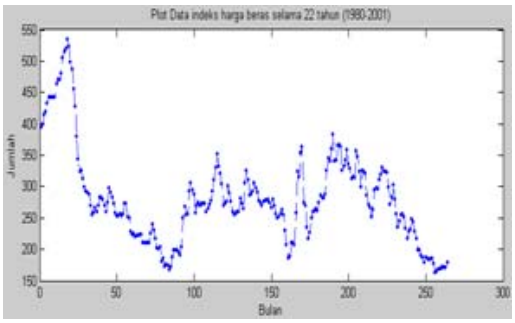

Figure 1 Data Plot

Model Choosing

Since data is influenced by 3 variables, then there are 7 RDLM candidate models, that are : M1(D,X1),M2(D,X2),M3(D,X3),M4(D,X1,X1), M5(D,X1,X3),M6(D,X2,X3), M7(D,X1,X2,X3).

After implementing AIC, then weight of every model is obtained as following:

Table 1 AIC Weights

Model	AIC Weight
M1	4,8344e-007
M2	1,7704e-050
M3	2,0559e-009
M4	2,2081e-017
M5	**2,0597e+006**
M6	6,5219e-116
M7	3,0446e-069

From the table above, it can be seen that M5 has the largest weight, so M5 is the most optimal model.

Optimal Model Forecasting

From the previous section, M5 has been chosen as the most optimal model which will be used for forecasting. The forecasting result of M5 is shown on figure 2.

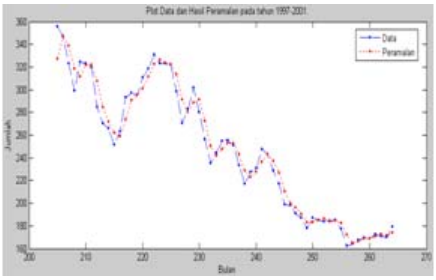

Figure 2 RDLM Forecasting

From the forecasting result, then the accuration is calculated and shown on the following table

Table 2 Error Estimation Result

Method	Result
MAPE (%)	0.62897
Theil's U	0.20262

From those calculation, it can be seen that RDLM model has excellent performance in forecasting. This is because its MAPE value lies under 10%, that is 0.62897%. From Theil's U point of view, this model is effective since its U value is under 1.

Testing AIC Performance

In order to analyze AIC performance, every model accuration will be compared, then it can be seen whether model that has been chosen by AIC is a model with the smallest error. Accuration of every model is shown on the following table :

Table 3 Errors of Every Model

Model	MAPE (%)	U
M1	0.63678	0.20535
M2	0.64513	0.20637
M3	0.63415	0.20373
M4	0.63892	0.20490
M5	**0.62897**	**0.20262**
M6	0.63965	0.20421
M7	0.63717	0.20406

It can be seen from the table above that M5 has the smallest error, so it can be concluded that AIC method works well in choosing model.

Comparation Between RDLM and DLM Performance

In this section, RDLM and DLM will be compared to prove that RDLM method work better than DLM method. This can be done by comparing DLM model with the most optimal RDLM model that is M5. Forecasting result of both of those models is plotted on figure 3.

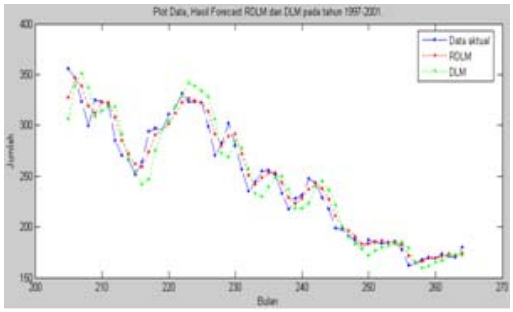

Figure 3 RDLM and DLM Forecasting

From the forecasting result, error estimation will be done and shown on the following table.

Table 4 RDLM and DLM Errors

Method	MAPE (%)	U
RDLM	0.62897	0.20262
DLM	1.2062	0.37716

From the above table, it can be seen that RDLM method has smaller error than DLM model, from the MAPE and Theil's U value.

CONCLUSION

Several conclusions than can be taken about application of BMA-Akaike Information Criteria (AIC) in RDLM (Regression Dynamic Linear Model) forecasting method is as follows :

1. Forecasting using RDLM (Regression Dynamic Linear Model) has high accuration as long as the chosen model is the most optimal model.
2. BMA-Akaike Information Criteria (AIC) method is proven to determine the RDLM models optimally.

3. Forecasting using RDLM method has better result than normal DLM method as long as the RDLM model is the most optimal model.
4. Using rice price index data on 1997 – 2001, RDLM method works 48% better than DLM method judging from MAPE value, and 46% better judging from Theil's U value.

DAFTAR PUSTAKA

1. Akaike, H. (1974), A new look at the Statistical model identication. *IEEE Trans. Auto. Control*, 19, 716-723.
2. Aplevich, J., (1999). *The Essentials of Linear State-Space Systems*. J. Wiley and Sons.
3. Grewal, M. S., Andrews, A. P., (2001). *Kalman Filtering: Theory and Practice Using MATLAB (2nd ed.)*. J. Wiley and Sons.
4. Harvey, A., (1994). *Forecasting, Structural Time-series Models and the Kalman Filter*. Cambridge University Press.
5. Mubwandarikwa, E., Faria A.E. (2006) *The Geometric Combination of Forecasting Models* Department of Statistics, Faculty of Mathematics and Computing, The Open University.
6. Mubwandarikwa, E., Garthwaite, P.H., dan Faria, A.E., (2005). *Bayesian Model Averaging of Dynamic Linear Models*. Department of Statistics, Faculty of Mathematics and Computing, The Open University.
7. Turkheimer, E., Hinz, R. and Cunningham, V., (2003), On the undesirability among kinetic models: from model selection to model averaging. *Journal of Cerebral Blood Flow & Metabolism*, 23, 490-498.
8. Verrall R. J. (1983). *Forecasting The Bayesian* The City University, London
9. West , Mike. (1997). *Bayesian Forecasting*, Institute of Statistics & Decision Sciences Duke University.
10. World Primary Commodity Prices.(2002). Diambil pada tanggal 23 Mei 2008 dari http://www.economicswebinstitute.org.
11. Yelland, Phillip M. & Lee, Eunice. (2003), *Forecasting Product Sales with Dynamic Linear Mixture Models.* Sun Microsystem.
12. Zainun, N. Y., dan Majid, M. Z. A., (2003). *Low Cost House Demand Predictor.* Universitas Teknologi Malaysia.