STA 642 - Final Project

Jae Hyun Lee, jl914

24 April, 2020

Abstract

In this paper, forecasting market price of individual apartments in Gangnam district in Seoul at certain time point is conducted. In the course of constructing model, Dynamic Linear Regression which is specific example of Dynamic Linear Model(DLM) is used. The main advantage of this model is that it can reflect time varying changes in data with the easiest interpretable model form, linear regression. In this analysis, actual apartment transaction data of Gangnam district in Seoul from 2006 to 2019 is used and next three months data is used for validation. The performance of Dynamic Linear Regression is compared with Ordinary Least Square regression model(OLS) in various criterions and it shows much better performance than OLS.

Keywords: Forecasting Apartment Price, Dynamic Linear Regression, Dynamic Linear Model, Time Varying Coefficient

1. Introduction

Background

In Korea, housing price is considered as one of the most important factor in economy for individuals and government. The proportion that housing price takes in personal asset in Korea is over 50%. As concentration of population becomes more intense, combining with increasing house debt, the housing price in Korea is dramatically increasing in a few years. Especially, 'Gangnam' district has the most expensive average housing price and it is a district which shows this situation most obviously. Explicitly, the housing price is affected by wide range of factors such as time, location, and infrastructure of districts. However, these relationships are not constant in time because economic situation and people's preference are changed by time. Therefore, the main purpose of this project is developing linear model which can reflect changes in relationships between factors and the price of apartment by using DLM so that it can predict the market price of individual apartments of certain time point precisely.

Previous studies

Since housing is commonly important factor for people's lives over the world, there are plenty of previous works whose purposes are predicting future housing price in various aspects. Kok et al.[1] tried to estimate values of real estates in specific time point which is the same goal of this project by using multifamiliy assets data about California, Florida, and Texas from 2011 to 2016. He constructed forecasting model using tree based ensemble method, 'Random Forest', to predict exact price of individual real estates. In addition, Xiaochen et al.[2] applied machine learning method called 'Long Term Short memory' to bulid model which has similar purpose using housing price data in Beijing, Shanghai, and Guangzhou. These two previous studies shows much better performance in prediction compared to OLS model and AutoRegressive Integrated Moving Average (ARIMA) model. However, these two methods have severe disadvantage compared to other linear model, especially LSTM, that they lose interpretabilty of model at the expense of high prediction performance. On the other hand, studies from other aspect are also conducted. Jihye Choi et al.[3] conducted spatial analysis to predict Jinju's transaction price of land by using regression-kriging model which is widely used in spatial regression. But this studies did not consider time varying changes aspect. Lastly, Heonsu Park et al.[4] uses VAR model to predict housing price index for each district in Seoul with macroeconomic variables like interest rate and price index of Cheonse(unusual housing form which is only exist in Korea). The limitation in this study is that it cannot predict the price of indiviual apartment. Thus, model used in this project approches the problem of forecasting housing price differently from previous studies in that it can consider time varying property of housing price keeping interpretability and it is also able to predict individual price of apartment.

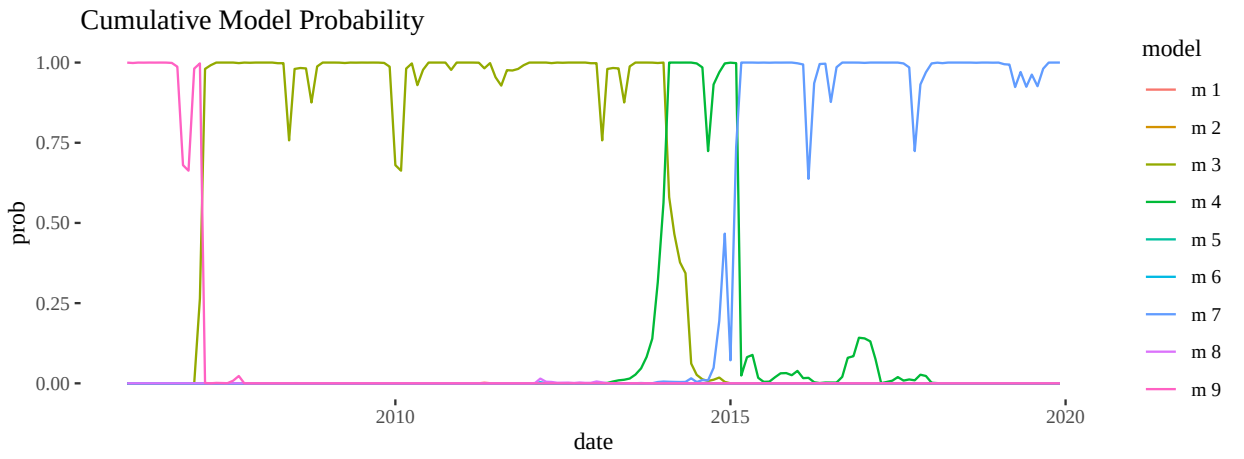

Figure 1: Model probability using different predictor sets

2. Apartment transaction data

Data source

In this analysis, real apartment transaction data of Gangnam district in Seoul from 2006 to 2019 is used and next three months transaction data is also used as test data. Data can be obtained from two sources: https://rt.molit.go.kr/ and http://www.k-apt.go.kr/. At the first data source, it is possible to download publically available real transaction price data of apartments in Korea. The transaction data of Gangnam district includes 69523 apartment transaction records with their `Address`, `Size`, `Transaction date`, `Price` and etc which is mainly about price and specific information of individual apartments. On the other hand, it is also possible to get data about environmental detail imformations of 173 apartment complexes such as the number of subway or education facility nearby apartment complexes from second data source. From this data, `Size`, `Age`,`Floor`, `Number of parking lot`, `Number of subway line`, `Number of education facility`, and `Number of convenience facility` are selected as candidate predictor to construct forecasting model. After merging data and select variables, I could get 51309 apartment transaction records with 11 variables. In addition, `Price`, `Size`, `Age`, `Floor`, and `Parking lot` are log transformed to satisfy normality assumption and to make better prediction performance.

Predictor variable selection

From selected predictor variables, all combinations of predictor variables are considered as candidate models. Based on model I constructed, some specified candidate models are selected which showed better prediction performance on test data than others. After preselection, marginal likelihood value and cumulative probability of each models are used as criterion to select model. As a result, model7 which uses `Size`, `Age`, `Floor`, `Parking lot`, `Subway`, and `Edu` as predictor variables is selected.

Discount factor selection

To reflect stochastic variation in β_t and ν_t, discount factors should be selected. I set values from 0.8 to 0.99 as candidates for each discount factors, and AIC of models are calculated for each combinations of discount factors (b,d). From the results in Figure 2, it is able to find that combination of d = 0.99 and b = 0.99 shows the lowest AIC. Thus (0.99,0.99) is selected optimal value of discount factors and this indicates that the model is very stable.

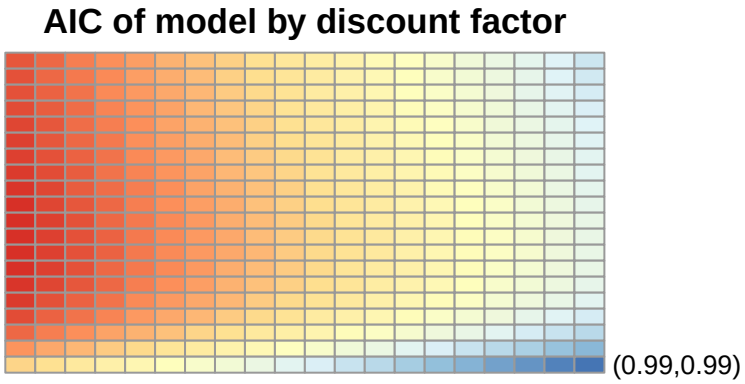

Figure 2: Heatmap for discount factors

3. Model

Model structure

Dynamic linear regression is fundamentally based on DLM. Only few parameters and settings are changed.

The structure of model is as follow:

$$Y_t = X_t\beta_t + \epsilon_t$$
$$\beta_t = \beta_{t-1} + w_t$$

If we call the number of transaction occurred at time t as O_t and the number of predictor variable in X_t as p, then we can describe notation and dimension as follow table

Table 1: Description of notation

Notation	Description	Dimension
Y_t	Price of apartment	$O_t \times 1$
X_t	Predictor variables	$O_t \times p$
β_t	Linear regression coefficient	$p \times 1$
ϵ_t	Measurement error	$O_t \times 1$
w_t	Innovation error	$p \times 1$

First equation in the model represents linear regression of Y_t given predictor dataset X_t. In this analysis, six predictor variables are used: `Size`, `Age`, `Floor`, `Parking lot`, `Subway`, `Edu`. i.e. X_t is consist of seven predictors including intercept term. Moreover, since OLS assumes that error terms are independent and identically distributed, this model also follows this assumption. Therefore, distribution of measurement error is as follow:

$$\epsilon_t \sim N(0, \nu_t I_{O_t})$$

At second equation, β_t has random walk process with innovation error w_t. This is specific example of latent state process in DLM which has I(identity matrix) as evolution matrix G. In addition, this model assumes voliatility in measurement error which means that ν_t varies according to the time. Thus, distribution of innovation error is as follow:

$$w_t \sim N(0, \frac{\nu_t}{s_t} W_t) \quad where \;\; W_t = (\frac{1-d}{d}) \times C_t$$

Forward Filtering and Backward Smoothing

In this model, forward filtering is relatively different from standard DLM. First, Y_t is not a scalar, but a vector. Therefore, it couldn't use same algorithm of univariate DLM because q_t which represent forecast error is no longer scalar. Moreover, it cannot apply multivariate DLM filtering methods because the parameter regarding variance is a scalar, not a covariance matrix. Thus, instead of standard algorithm of DLM, I decided to use bayesian update algorithm which is used for getting posterior distribution of β and σ^2 in linear regression.

First step is specifying prior for ν_t and β_t. For priors, conjugate priors are used. Prior parameters m_0 is mean, and C_0 is covariance of roughly estimated coefficient at each time point. Moreover, n_0 is 100 which is nearly average of observation at each time and S_0 is 0.09 which is also nearly average of estimated variance.

$$\nu_0 \mid D_0 \sim IG(\frac{n_0}{2}, \frac{n_0 s_0}{2})$$
$$\beta_0 \mid \nu_0, D_0 \sim N(m_0, \frac{\nu_0}{s_0} C_0)$$

Second step is evolving parameters according to model

$$\nu_{t-1} \to \nu_t : \nu_t \mid D_{t-1} \sim IG(\frac{bn_{t-1}}{2}, \frac{bn_{t-1}s_{t-1}}{2})$$
$$\beta_{t-1} \to \beta_t : \beta_t \mid D_{t-1} \sim N(a_t, \frac{\nu_t}{s_{t-1}} R_t)$$
$$where\ a_t = m_{t-1}, R_1 = C_{t-1}/d$$

Final step in forward filtering is updating evolved distribution of ν_t and β_t based on new observation Y_t and corresponding X_t.

$$\nu_t \mid D_t \sim IG(\frac{n_t}{2}, \frac{n_t s_t}{2})$$
$$\beta_t \mid \nu_t, D_t \sim N(m_t, \frac{\nu_t}{s_t} C_t)$$

where

$$T_t = (X_t' X_t + s_{t-1} R_t^{-1})^{-1}$$
$$m_t = T_t (X_t' Y_t + s_{t-1} R_t^{-1} a_t)$$
$$n_t = bn_{t-1} + O_t,$$
$$s_t = \frac{1}{n_t}(bn_{t-1}s_{t-1} + Y_t' Y_t + s_{t-1} a_t' R_t^{-1} a_t - m_t' T_t^{-1} m_t)$$
$$C_t = T_t / s_t$$

Compared to forward filtering, backward smoothing is identical with algorithm of standard univariate DLM. Backward smoothing starts with the known distribution $N(\beta_T \mid m_t, \frac{\nu_T}{s_T} C_T)$. If we set up $a_T(0) = m_T, R_T(0) = C_T, n_T(0) = n_T, s_T(0) = s_T$, for $t = T-1, T-2, \cdots 1$, estimations for mean and variance of β_t are modified as follow:

$$a_T(t-T) = m_t - d \times (a_{t+1} - a_T(t-T+1)) = (1-d)m_t + d a_T(t-T+1)$$
$$R_T(t-T) = C_t - d^2 \times (R_{t+1} - R_T(t-T+1)) = (1-d)C_t + d^2 R_T(t-T+1)$$
$$n_T(t-T) = (1-b)n_t + b n_T(t-T+1)$$
$$s_T(t-T) = [(1-b)/s_t + b/s_T(t-T+1)]^{-1}$$

4. Result

Dynamic linear regression is compared with ordinary linear regression(OLS). The predictor variables in both models are identical which are: `Size`, `Age`, `Floor`, `Parking lot`, `Subway`, and `Edu`. The structure of candidate models are as follow:

Table 2: Model description

Model1	Model2
$Y_t = X_t\beta + \epsilon_t$	$Y_t = X_t\beta_t + \epsilon_t$ $\beta_t = \beta_{t-1} + w_t$

The comparison between two models is conducted based on these criterions

Table 3: Criterion description

Criterion	Formula	Description
AIC	$\sum(y_i - \hat{y}_i)^2 + 2p$	Goodness of fit
MSE	$\sum \frac{1}{n_{test}}(y_{i,test} - \hat{y}_{i,test})^2$	Prediction accuracy for future data
Coverage	$Pr(y_{i,test} \in CI(y_{i,test}))$	Proportion that CI includes actual future data

The comparison between two models is as follow:

Table 4: Comparison result

	AIC	MSE	Coverage
OLS	5302.87	0.38	58.53 %
DLM	2990.06	0.08	100 %

The difference in performance of these two models is very obvious. AIC of OLS is 5302 whereas AIC of DLM is 2990 which is so much lower value than OLS that AIC of OLS is larger than AIC of DLM as much as twice. This finding indicates that DLM predicts the price of apartment for training data from 2006 to 2019 more precisely than OLS. The difference of two models' performance becomes much bigger when we focus on forecasting future apartment price in test dataset. Mean sqaured error(MSE) of DLM for future dataset is 0.075 whereas MSE of OLS is 0.383 which is larger than DLM as much as five times. In addition, prediction interval of OLS for future individual apartment price produced from simulation only included 58.53% of actual data. On the other hand, DLM succeed to include all data. This bigger difference in prediction performance is explained when we investigate marginal likelihood values for both models by time. At Figure 5, we can notice that marginal likelihood of OLS dropped dramatically from 2014 and this indicates that OLS becomes very inaccurate after this time point. This time point is that the average price of apartment in Seoul increase dramatically. From this situation, it can be inferred that OLS failed to reflect this dramatic changes in apartment price meanwhile DLM succeed to capture it. Therefore, DLM is a more appropriate model to forecast future price of apartments.

The reason of this difference of two model in performance is obvious when we investigate some time series plots of regression coefficients. First, according to left-upper plot of Figure 4, observed coefficient of `Intercept` constantly has increased from 2006 which is about -1 to 2019 which is about 3. However, OLS failed to reflect this change and underestimate intercept than observed intercept. In additon, coefficients of `Size`, `Subway`, `Age` are overestimated than actual coffecients. Especially, most of observed coefficients of `Age` are

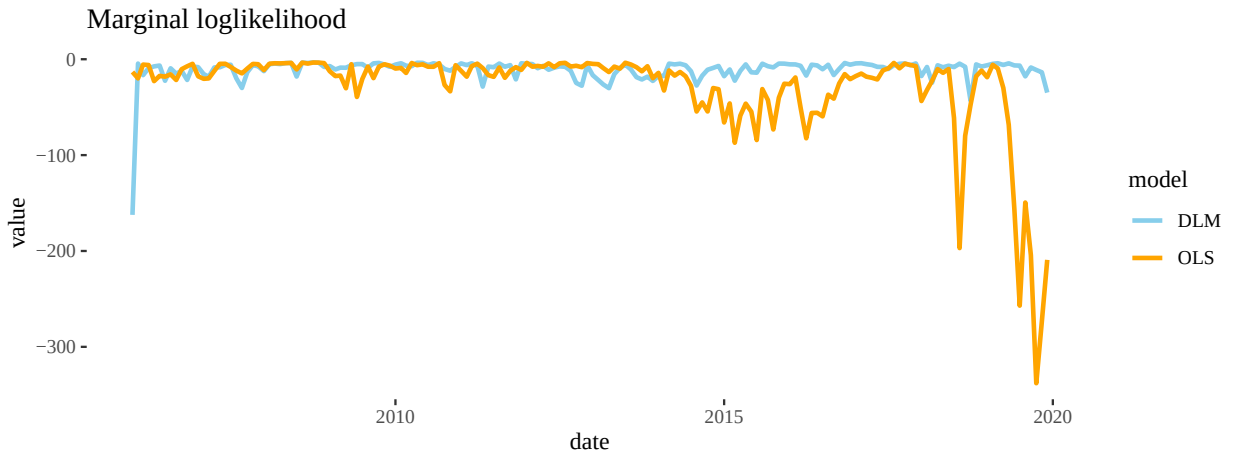

Figure 3: Marginal likelihood for models

Figure 4: Estimate of coefficient for models and observed coefficient

less than OLS estimate. Furthermore, obvious discrepancy between observed coefficients and OLS estimates in intercept and size coefficent starts from 2014 which is the time point that OLS becomes very inaccurate. This indicates that inaccurate estimates in intercept and size make OLS inaccurate and it also indicates that these two factors are playing a important role in determining the price of apartments.

5. Conclusion

Summary

The price of apartment is determined by various factors such as time, size of apartment, the number of education facility and their convenience of transportation. However, the relationships between price and these factors are not fixed, but flexible. Vanilla OLS cannot reflect these time varying relationship and changes of price itself. As an alternative, in this project, dynamic linear regression model is suggested to accomplish more accurate prediction on the price of individual apartments. Consequently, the prediction performance is highly improved by using dynamic linear regression instead of OLS with same predictor dataset. Even though, dynamic linear model showed better performance in prediction for training dataset which is transaction data from 2006 to 2019, the point dynamic linear regression has big advantage over OLS is prediction on recent data. In both aspect of point estimate and interval estimate, dynamic linear regression showed much better performance for recent data.

Limitation

Although the model I used in this project showed improved performance over OLS, there are several limitation in this project. First, this model fails to include autoregressive structure in regression coefficients. At the beginnig, VAR coefficients matrix was used as evoultion matrix G to contruct DLM so that it can use previous estimations of coeffcients to make estimated current coefficients more precise. However, when VAR coefficients matrix G satisfy stationary condtition of VAR model, it turns out that its inverse matrix has eigenvalues larger than 1 which cause backward smoothing to be amplified and estimations to be very inaccurate. I have tried to select discount factor for state which is lower than the lowest value among G's eigenvalues, but this selected discount factor is usually so low that it makes estimations of coefficients be very volatile. As a result, it failed to construct autoregressive structure in this model. Secondly, data from second data source is relatively imprecise because there were several duplicated apartment complexes which had different information and this might be caused by the fact that this data is manually written. Therefore, I used average value of each information for duplicated apartment complexes. But this might potentially cause inaccuracy in model I used.

Further study

There are several points that can be improved from this project. Firstly, as written at above limitation, it can include VAR coefficient as evolution matrix. Another point that can be improved is including macroeconomic factors and other source from social media using dynamic latent factor model. Understandably, the price of apartment is largely affected by macroeconomic factor because it is considered as asset. In addition, currently, certain districts in Seoul become popular and this makes the price of this distict higher. In my personal aspect, this popularity can be examined by the number of posting with Hashtag of districts for each month in social media called "Instagram". Lastly, I want to analyze not only Gangnam district, but also all apartment transaction data in Seoul. One of the assumed way to utilize them is using hierachical structure on regression coefficients for each district. The main reason I could not use VAR model to include other district in Seoul is that they does not have exchangable structure. The requirement in exchangable DLM is that they should have common predictor matrix. However, in this case, predictor matrix are not idential over all district. Therefore, if I apply hierachical structure in regression coefficients, I will be able to include all information of Seoul and share them for each districts.

6. Reference

[1] N.Kok, E.L.Koponen, and C.Barbosa. "Big Data in Real Estate? From Manual Appraisal to Automated Valuation." *The Journal of Portfolio Management Special Real Estate*, Vol.43, No.6, (2017) pp. 202-211

[2] X.Chen, L.Wei, and J.Xu. (2017) "House Price Prediction Using LSTM."

[3] J.Choi, H.G.Jin, and Y.Kim. "Spatial analysis for a real transaction price of land." *The Korean Journal of Applied Statistics* , Vol.31, No.2, (2018) pp. 217-228

[4] H.S.Park & J.A.An. "The Sources of Regional Real Estate Price Fluctuations." *Korea Real Estate Review*, Vol.19, No.1, (2009) pp. 27-49.

[5] R.Prado & M.West (2010). "Time series: Modeling, computation, and inference." 10.1201/9781439882757.