Advances in Economics, Business and Management Research (AEBMR), volume 60
4th Annual International Conference on Management, Economics and Social Development (ICMESD 2018)

Research on the Influence Factors of Housing Price Based on Semi-Parametric Functional Linear Model

Peng JIN[1,a,*],Qing-Guo TANG[1,b]

[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing 21094, China

[a]18539920797@163.com, [b]tangqguo@163.com

Keywords: Semi-parametric functional linear model, Functional principal component analysis, Housing price, Influence factors.

Abstract. In this paper, using the theory and method of functional principal component analysis and the optimization theory and method, we develop a new estimation method for estimating the unknown parameters and function in semi-parametric functional linear model. We use the model and its estimation method to study urban housing data. We establish the quantitative relationship between urban housing prices and its influencing factors. By using the research results, we give the important factors that affect housing prices.

Introduction

As a basic industry, the real estate industry plays an important role in the national economy and personal property. Statistics from the National Bureau of Statistics shows that in 2016, the real estate industry directly accounted for 6.5% of GDP, and the GDP growth rate was about 0.22 percentage points. The narrow property industry chain accounted for about 12% of GDP. According to CHFS data, Chinese households account for 69% of total assets. However, with the rapid development of the real estate industry, many problems have emerged. The most significant problem is that with the development of the commercialization of land and housing, the price of commercial housing has risen rapidly. In 2008, the house prices in Beijing and Shanghai were 13,302 yuan/square meter and 15,163 yuan/square meter respectively; By 2017, the prices in Beijing and Shanghai rose to 54,704 yuan/square meter and 53,343 yuan/square meter respectively.

Functional data analysis(FDA) has been receiving more and more attention in the past two decades. Function linear models have been extensively studied and applied. See Ramsay and Silverman[1][2], Yao et al.[3], Hall and Horowitz[4], Cram beset et al.[5], Reiss and Ogden[6].In order to improve the predictive power and explanatory power of the regression model, some additional real-valued independent variables were introduced into the function linear model, resulting in some new functional models. For example, Shin[7] proposed a semi-parametric function linear model; Lian[8] proposed a new semi-function linear model; Tang and Kong[9] proposed a semi-function linear semi-parametric model.

In this paper, a new method of estimation, which is different from Shin[7], is developed using function principal component analysis to estimate the unknown parameters and functions in a semi-parametric function linear model. At the same time, this paper will use this model and its estimation method for the first time to study the quantitative correlation between the sales price of urban commercial housing and the important factors affecting the sales price. To the best of our knowledge, this is the first time that the theory and methods of functional data analysis have been applied to real estate data analysis.

Models and Estimation Methods

Let Y be a real-valued random variable defined on a probability space(Ω, F, P). Let Z be a d-dimensional vector of random variables with finite second moments and let$\{X(t): t \in T\}$be a zero-mean and second-order (i.e., $EX(t)^2 < \infty$ for all$t \in T$)stochastic process defined on(Ω, F, P)

with sample paths in $L^2(T)$, the set of all square integrable functions on T , where T is a bounded closed interval. Let <•,•> and $\|\bullet\|$ represent, respectively, the $L_2(T)$ inner product and norm. Considering the following partial functional linear regression model:

$$\text{Eq.1}\; Y = \int_T \gamma(t)X(t)dt + Z^T\beta_0 + \varepsilon \tag{1}$$

where $\gamma(t)$ is a square integrable function on T, Z is a d-dimensional vector of random variables with finite second moments, β_0 is a d × 1 coefficient vector to be estimated and ε is a random error with mean zero and is independent of (X,Z) .Denote the covariance function of the process $X(t)$ by $K(s,t) = cov(X(s), X(t))$ .We suppose that $K(s,t)$ is positive definite, in which case it admits a spectral decomposition in terms of strictly positive eigenvalues λ_j .

$$\text{Eq.2}\quad K(s,t) = \sum_{j=1}^{\infty}\lambda_j\phi_j(s)\phi_j(t), \quad s,t\in T \tag{2}$$

where (λ_j,ϕ_j) are (eigenvalue, eigenfunction) pairs for the linear operator with kernel *K*, the eigenvalues are ordered so that $\lambda_1 > \lambda_2 > \cdots$ and the functions $\phi_1,\phi_2,\cdots$ form an orthonormal basis for $L_2(T)$. This leads to the Karhunen-Loèvere presentation

$$\text{Eq.3}\quad X(t) = \sum_{j=1}^{\infty}\xi_j\phi_j(t) \tag{3}$$

where the $\xi_j = \int_T X(t)\phi_j(t)dt$ are uncorrelated random variables with mean 0 and variance $E\xi_j^2 = \lambda_j$. Let $\gamma(t) = \sum_{j=1}^{\infty}\gamma_j\phi_j(t)$, then model (1) can be written as

$$\text{Eq.4}\quad Y = \sum_{j=1}^{\infty}\gamma_j\xi_j + Z^T\beta_0 + \varepsilon \tag{4}$$

By (2), we have

$$\text{Eq.5}\quad \gamma_j = E\left\{\left[Y - Z^T\beta_0\right]\xi_j\right\}/\lambda_j \tag{5}$$

Let $(X_i(t), Z_i, Y_i), i = 1,2,\cdots n,$ be independent realizations of $(X(t),Z,Y)$ generated by the model (1). Empirical versions of *K* and of its spectral decomposition are

$$\text{Eq.6}\quad \hat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n}X_i(s)X_i(t) = \sum_{j=1}^{\infty}\hat{\lambda}_j\hat{\phi}_j(s)\hat{\phi}_j(t) \tag{6}$$

Analogously to the case of *K*, $(\hat{\lambda}_j,\hat{\phi}_j)$ are (eigenvalue, eigenfunction) pairs for the linear operator with kernel $\hat{K}$, ordered such that $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge 0$. We take $(\hat{\lambda}_j,\hat{\phi}_j)$ to be the estimator of (λ_j,ϕ_j) and take

$$\text{Eq.7}\quad \tilde{\gamma}_j = \frac{1}{n\hat{\lambda}_j}\sum_{i=1}^{n}(\mathrm{Y}_i - \mathrm{Z}_i^T\beta_0)\hat{\xi}_{ij} \tag{7}$$

be the estimator of γ_j

we use $\sum_{j=1}^{m}\tilde{\gamma}_j\hat{\xi}_j$ to approximate $\sum_{j=1}^{\infty}\gamma_j\xi_j$ in (2). Combining (2) and (7),we then solve the following minimization problem

$$\text{Eq.8}\quad \min_{\beta}\sum_{i=1}^{n}\left\{Y_i-\sum_{j=1}^{m}\frac{\hat{\xi}_{ij}}{n\hat{\lambda}_j}\sum_{l=1}^{n}(Y_l-Z_l^{T}\beta)\hat{\xi}_{lj}-Z_i^{T}\beta\right\}^2 \qquad (8)$$

to obtain the estimator of β_0 . Define $\tilde{\xi}_{li}=\sum_{j=1}^{m}\frac{\hat{\xi}_{lj}\hat{\xi}_{ij}}{\hat{\lambda}_j}, \breve{Y}_i=Y_i-\frac{1}{n}\sum_{l=1}^{n}Y_l\tilde{\xi}_{li}$ and $\breve{Z}_i=Z_i-\frac{1}{n}\sum_{l=1}^{n}Z_l\tilde{\xi}_{li}$ .Then (8) can be written as

$$\text{Eq.9}\quad \min_{\beta}\sum_{i=1}^{n}(\breve{Y}_i-\breve{Z}_i^{T}\beta)^2 \qquad (9)$$

Let $\breve{Y}=(\breve{Y}_1,\breve{Y}_2,\cdots,\breve{Y}_n)^T$ and $\breve{Z}=(\breve{Z}_1,\breve{Z}_2,\cdots,\breve{Z}_n)^T$. Then the estimator $\hat{\beta}$ of β_0 is given by

$$\text{Eq.10}\quad \hat{\beta}=(\breve{Z}^{T}\breve{Z})^{-1}\breve{Z}^{T}\breve{Y} \qquad (10)$$

The estimator of $\gamma(t)$ is given by $\hat{\gamma}(t)=\sum_{j=1}^{m}\hat{\gamma}_j\hat{\phi}_j(t)$ with

$$\text{Eq.11}\quad \hat{\gamma}_j=\frac{1}{n\hat{\lambda}_j}\sum_{i=1}^{n}(Y_i-Z_i^{T}\hat{\beta})\hat{\xi}_{ij} \qquad (11)$$

To implement our estimation method, we need to know how to choose m. The value for m can be selected by leave-one-curve-out cross-validation of the prediction error. Define CV function as

$$\text{Eq.12}\quad CV(m)=\sum_{i=1}^{n}(Y_i-\sum_{j=1}^{m}\hat{\gamma}_j^{-i}\hat{\xi}_{ij}-Z_i^{T}\hat{\beta}^{-i})^2 \qquad (12)$$

where $\hat{\gamma}_j^{-i},\ j=1,\cdots,m$ and $\hat{\beta}^{-i}$ are computed after removing (X_i,Z_i,Y_i) . As an alternative to cross-validation, the m that minimizes the CV function is what we want to select.

Application

Selection of Data and Variables

We have collected real estate data from 90 four-tier cities. These data are mainly from the statistical yearbooks of various cities, real estate market reports, and statistical bulletins on national economic and social development. The dependent variable is house price; we use Y to indicated. We selected the following housing price influencing factors as independent variables: urban livability index(Z_1), urban comprehensive competitiveness (Z_2), urban development index (Z_3), urban population (Z_4), and urban GDP (Z_5), interest rate (Z_6). Among them, the average annual income of residents has been selected from 2000 to 2016.Since the average annual income of residents is divided into three types: the average annual income of urban residents, the average annual income of rural residents, and the average annual income of the residents of the city. For the convenience of calculation, the average annual income of urban residents is used in this paper. Other data selected 2016 data. Fig 1 shows the box plot of the statistics. The box plot shows the difference in the value of each variable. Each box plot represents 25/50/75 percentage point. It can be seen that the difference between house prices is the largest and the difference between interest rates is different. This is also in line with our expectations.

Data Outlier Test

The accuracy of statistical data is an important topic in the field of statistical research, it is also a problem that is generally concerned with statistical work. In the actual statistical work, we often encounter the following situations: The value of a single variable is too large or too small, Obvious deviations from most observations will affect the accuracy of statistical results.

Since the variables we select contain qualitative data and quantitative data, quantitative data (such as city GDP and residents' annual income, etc.) come from official statistical bulletins, and there are generally no abnormal values. Qualitative data, such as the city's overall competitiveness and livability index, are often derived from expert scores and have a certain degree of subjectivity, therefore have abnormal values. For this reason, we have introduced cook distances for identifying outliers. The Cook distance is a common distance in statistical analysis and is used to diagnose abnormal data in various regression analyses. Larger Cook distances indicate that there is a fundamental change in the coefficient after removing cases from regression statistics and calculations.

Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis.

For the algebraic expression, first define

$$\text{Eq.13}\quad \underset{n\times 1}{y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \qquad (13)$$

Where ε is the error term, $b=(b_0,b_1,...,b_{p-1})^T$ is the coefficient matrix, p is the number of covariates or predictors for each observation, and X is the design matrix including a constant. The least squares estimator then is $\overset{\$}{b}=(X^TX)^{-1}X^Ty$, and consequently the fitted(predicted) values for the mean of y are

$$\text{Eq.14}\,\overset{\$}{y}=X\overset{\$}{b}=X(X^TX)^{-1}X^Ty=Hy \qquad (14)$$

Where $H=X(X^TX)^{-1}X^T$ is the projection matrix (hat matrix). The i-th diagonal element of *H*, given by $h_i=X_i^T(X^TX)^{-1}x_i$, is known as the leverage of the i-th observation. Similarly, the i-th element of the residual vector $e=y-\overset{\$}{y}=(I-H)y$ is denoted by e_i.

Cook's distance D_i of observation $i(\forall i=1,2,...,n)$ is defined as the sum of all the changes in the regression model when observation *i* is removed from it.

$$\text{Eq.15}\quad D_i=\frac{\sum_{j=1}^{n}(\overset{\$}{y}_j-\overset{\$}{y}_{j(i)})^2}{ps^2} \qquad (15)$$

Where $\overset{\$}{y}_{j(i)}$ is the fitted response value obtained when excluding *i*, and $s^2=(n-p)^{-1}e^Te$ is the mean squared error of the regression model.

We sorted Pinyin by 90 fourth-tier cities and calculated the cook distance of the above-mentioned qualitative indicators using R statistical software.

It can be clearly seen from Fig. 2 that the distance of the habitability index of cooks in each city fluctuates between 0 and 0.1, and neither exceeds 0.4, so there is no abnormal value. Similarly, in Figure 4, there is no abnormal value for the development index of each city. In Figure 3, it can be seen that most cities have a city-community competitiveness of 0-0.1, and very few cities have a cook distance greater than 0.4, such as City 1, which is an outlier.

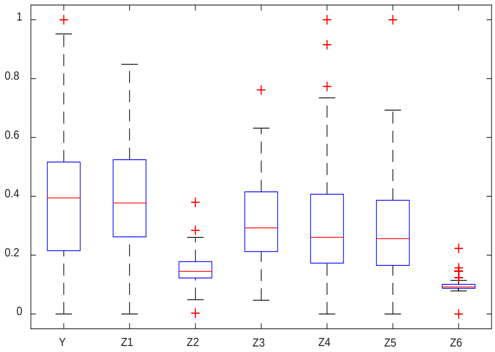

Fig.1 Box-plot of related data

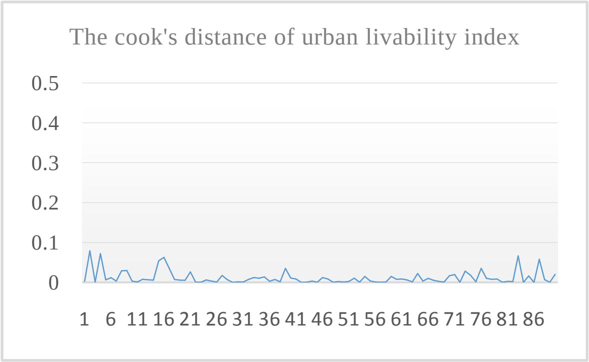

Fig.2 The cook's distance of Z_1

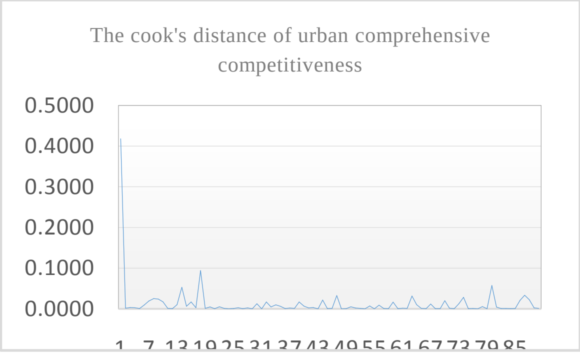

Fig.3 The cook's distance of Z_2

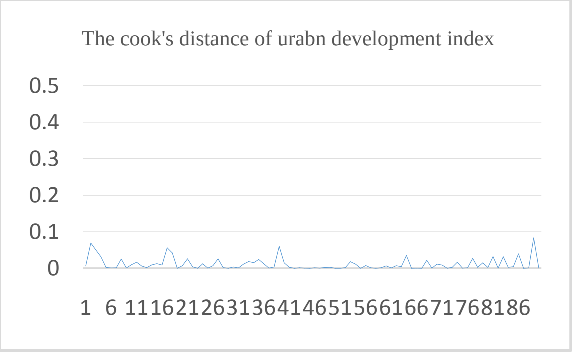

Fig.4 The cook's distance of Z_3

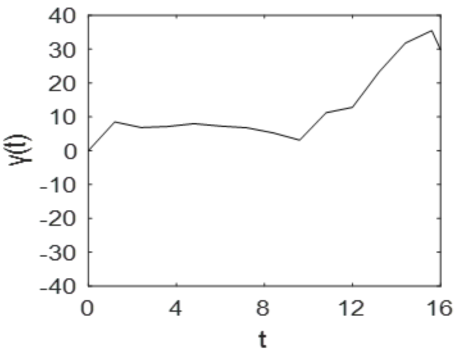

Fig.5 estimated curve of $\gamma(t)$

Application of Models

We will use the model that proposed in the second section and its estimation method to study the influencing factors of house prices. The response variable Y represents the urban housing price. We have chosen a function-type independent variable and six common digital-type independent variables. Since ordinary residents purchase housing for many years of savings, we select the average annual income of residents as a function type variable, and use $X_i(t)$ to represent the average annual income of residents in the t-th year of the i-th city. We provide: t = 0 refers to the average annual income of residents in 2000, t = 1 refers to the average annual income of residents in 2001, and so on, t = 16 refers to the average annual income of residents in 2016.The digital independent variables include urban livability index (Z_1), urban comprehensive competitiveness (Z_2), urban development index (Z_3), urban residents permanent population (Z_4), city GDP (Z_5), and bank deposit interest rate (Z_6). We construct the following semi-parametric function linear model:

$$\text{Eq.16}\quad Y=\int_0^{16}\gamma(t)X(t)dt+\beta_0+\beta_1Z_1+\beta_2Z_2+\beta_3Z_3+\beta_4Z_4+\beta_5Z_5+\beta_6Z_6+\varepsilon \qquad (16)$$

We use the second part of the estimation method to estimate the unknown function $\gamma(t)$ and unknown parameters $\beta_i, i=1,2,...6$ in model (16). Figure 5 shows the estimated curve of the unknown function. Table 1 shows the value of the unknown parameter estimator. From Figure 5 we can see that the estimated curve of the unknown function is relatively stable in the interval [0, 10], but it is increasing in the interval [10, 15], and finally decreases slightly at t = 16.Since the estimation accuracy of the points near the boundary is worse than the estimation accuracy at other internal points, we think that the slight decrease of the estimated curve at t=16 is probably caused by the boundary estimation error. These indicate that the average annual income of residents before 2010 contributes less to housing prices, while the average annual income of residents after 2010 contributes more to housing prices, and this contribution is increasing year by year.

Table 1 Estimated curve of $\gamma(t)$

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
9.873	0.163	0.299	0.148	0.089	0.369	-0.706

According to the parameter estimation results in Table 1, we can see that the main factors affecting housing prices are urban GDP, urban comprehensive competitiveness, urban livability index and bank deposit interest rate.

Gross domestic product (GDP) refers to the market value of all final products and services produced by a permanent unit of a country or region during a certain period of time. GDP is the core indicator of national economic accounting, and it is also an important index to measure the overall economic status of a region. The GDP level and growth rate of a city are directly related to the level and growth rate of housing prices. When the economic situation is good, it will attract more investment, which will attract more companies to invest, and it will attract more people to employment and entrepreneurship, and it will also motivate people to increase their purchases. People generally think that the better the economic growth, the greater the possibility of rising house prices. While the economy is booming, developers will also be encouraged to increase investment scale, increase housing supply, and the demand for GDP and housing market supply will rise.

Urban comprehensive competitiveness refers to the ability of a city to collect resources and provide products and services within a certain region. The economic, social, technological, and environmental factors of a city largely determine the overall competitiveness of a city. If a city is more competitive than other cities, it means that the city has a better entrepreneurial and employment environment and higher income and welfare levels. Therefore, it will be able to attract more companies and residents to move into the city, leading to more people to buy a house, the city will be more able to accumulate population, resources, capital, housing demand will continue to increase. In addition, the degree of agglomeration also makes the industry upgrade faster, and the unit space creates greater value and higher house prices.

From the table, we can see that the coefficient of Z_6 is -0.7061, which shows that the bank interest rate is inversely proportional to the real estate price, and the real estate price drops when the interest rate rises; When the bank interest rate falls, the real estate price rises. The reasons are as follows: First, when the interest rate rises, the profits of saving and buying bonds will increase. Relatively speaking, real estate investment income is not attractive, and the amount of investment is greatly reduced. This will lead to a lack of support for real estate prices, resulting in falling house prices. Secondly, due to the rise in interest rates, investors' interest expense on capital use will increase, and in the event that commodity prices do not increase, the interest cost of funds cannot be transferred to real estate prices in a reasonable way. Real estate investment income is greatly reduced. At this time, Real estate investment cannot be said to be an ideal investment method.

When the interest rate fell to a certain degree, the demand for funds increased in all aspects, the economy began to revitalize, and more and more real estate funds were invested, thereby driving up the price of housing.

Summary

This paper presents a new estimation method for estimating unknown parameters and functions in a semi-parametric functional linear model. We selected real estate data for 90 four-tier cities in 2016, and the average annual income for urban residents selected the data for the past 17 years to study the relationship between the sales prices of real estate and influencing factors in the process of urbanization. We find that the factors affecting housing prices are mainly the comprehensive competitiveness of cities, city GDP, and urban livability index. The most important factor is the city's GDP, in which interest rates are negatively correlated with housing prices.

Acknowledgement

This research was financially supported by the Humanity and Social Science Research and Planning Foundation of Ministry of Education of China (Grant No. 14YJA910004), National Social Science Foundation (Grant No. 16BTJ019), Jiangsu Provincial Natural Science Foundation of China (Grant No. BK20151481).

References

[1] Ramsay J O, Silverman B W. Applied Functional Data Analysis: Methods and Case Studies [M]. New York: Springer, 2002.

[2] Ramsay J O, Silverman B W. Functional Data Analysis [M]. New York: Springer, 2005.

[3] Yao F, Müller H G, Wang J L. Functional linear regression analysis for longitudinal data [J]. Annals of Statistics, 2005, 33(6).

[4] Hall P, Horowitz J L. Methodology and convergence rates for functional linear regression [J]. Annals of Statistics, 2007, 35(1).

[5] Crambes C, Kneip A, Sarda P. Smoothing splines estimators for functional linear regression [J].Annals of Statistics, 2009, 37(1).

[6] Reiss P T, Ogden R T. Functional generalized linear models with images as predictors [J]. Biometrics, 2010, 66(1).

[7] Shin H. Partial functional linear regression [J]. Journal of Statistical Planning & Inference,2009, 139(10).

[8] Lian H. Functional partial linear model [J].Journal of Nonparametric Statistics, 2011, 23(1).

[9] Tang Q, Kong L. Quantile regression in functional linear semiparametric model [J].Statistics, 2017, 51(6).