Multiple Linear Regression Analysis of Real Estate Price

Wang Aiyin , Xu Yanmei

Department of Mathematics, Xinjiang University of Finance and Economics, Urumqi, 830012, China
*corresponding author's email: 727668279@qq.com

***Abstract*—Considering the influence of residents' income, consumption level and national macro-control on real estate price, this paper selects seven indexes of population, GDP per capita, average income of urban residents, price level, real estate land purchase fee, loan interest rate and tax, quantifies the relationship between real estate price and various influencing factors, and evaluates the influencing factors of real estate price.**

Keywords- Real estate prices; Multiple linear regression: Multicollinearity: heteroscedasticity; Autocorrelation

I. INTRODUCTION

The real estate industry is the " backbone" of our national economy, and its status in the national economy is gradually rising. In recent years, the high real estate prices have become the focus of more and more social groups. From 2000 to 2015, the average price of real estate sales in China showed an increasing trend. As shown in figure 1, the growth rate of real estate sales exceeded the economic growth rate, making it difficult for ordinary income families to buy housing.

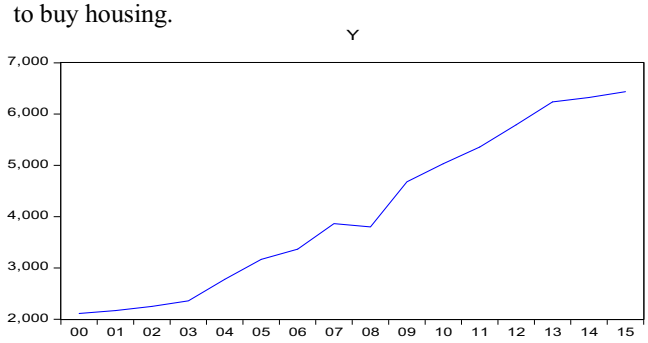

Figure 1 real estate prices 2000 - 2015

China 's rapid rise in urban housing prices has already exceeded the ordinary people 's economic capacity, the real estate market differentiation is serious, first-line cities and part of the second-line cities in short supply, soaring housing prices, and inventory pressure is mainly concentrated in the three or four line cities. In order to prevent the future may be due to the economic imbalance of real estate prices, in order to ensure the healthy and stable development of China 's real estate market, the real estate as a huge industry research object has important theoretical significance and practical value.

II. ESTABLISH PRELIMINARY REGRESSION MODEL

Data on average price and influencing factors of China 's housing sales from 2000 to 2015 are shown in table 1:

Table 1 national housing sales average price and influencing factors

Time	Average house sales price (yuan / m2)	Per capita GDP (yuan)	Price level (CPI)	Interest rate (%)	Income (yuan)	Land price (billion yuan)	Population (thousands)	Taxation (billion yuan)
2000	2112	7942	100.4	5.58	9333	733.99	126743	12581.51
2001	2170	8717	100.7	5.58	10834	1038.77	127627	15301.38
2002	2250	9506	99.2	5.04	12373	1445.81	128453	17636.45
2003	2359	10666	101.2	5.04	13969	2055.17	129227	20017.31
2004	2778	12487	103.9	5.22	15920	2574.47	129988	24165.68
2005	3168	14368	101.8	5.22	18200	2904.37	130756	28778.54
2006	3367	16738	101.5	5.49	20856	3814.49	131448	34804.35
2007	3864	20205	104.8	6.135	24721	4873.25	132129	45621.97
2008	3800	24121	105.9	5.73	28898	5995.62	132802	54223.79
2009	4681	26222	99.3	5.73	32244	6023.71	133450	59521.59
2010	5032	30876	103.3	5.225	36539	9999.92	134091	73210.79
2011	5357	36403	105.4	5.85	41799	11527.25	134735	89738.39
2012	5791	40007	102.6	5.725	46769	12100.15	135404	100614.28
2013	6237	43852	102.6	5.725	51483	13501.73	136072	110530.7
2014	6324	47203	102.0	5.6	56360	17458.53	136782	119175.31
2015	6437	49992	101.4	4.85	62029	17675.44	137462	124922.2

Data sources: China statistical yearbook

978-1-5386-6580-0/18/$31.00 ©2018 IEEE
DOI 10.1109/ICRIS.2018.00145

According to the data shown in table 1, 16 years of relevant data were analyzed in time series, and multiple linear regression models were set up:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + u$$

Among them: Y – housing sales average, X_2 — per capita GDP, X_3 — price level, X_4 — interest rate, X_5 — income, X_6 — land price, X_7 — population, X_8 — tax; $\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ are unknown, u is residual, and E(u)=0 is independent of the other seven independent variables.

Since the sample data is time series data, the stability of the sample data is checked. This sequence may have trend entries, so select the unit root (ADF) check and run the results as shown in table 2:

Table 2 ADF test results

	ADF test	5%	10%	p
Y	-4.7273（2）	-3.8753	-3.3883	0.0146
x2	-4.9483（2）	-3.8753	-3.3883	0.0106
x3	-4.5014（1）	-3.8753	-3.3883	0.0202
x4	-7.0689（2）	-3.8289	-3.3629	0.0005
x5	-5.2464（2）	-3.8753	-3.3883	0.0071
x6	-6.3188（2）	-3.9333	-3.4200	0.0023
x7	-4.5746（2）	-3.8289	-3.3629	0.0161
x8	-3.3949（1）	-1.9740	-1.6029	0.0009

As can be seen from table 2, when the significance level is 5 % and 10 %, the test value of each index is less than the critical value of 5 % and 10 %, the original hypothesis h0 is rejected: there is a unit root, so the selected time series data is stable.

Make the linear diagram of each explanatory variable and the explained variable of the sequence, as shown in figure 2:

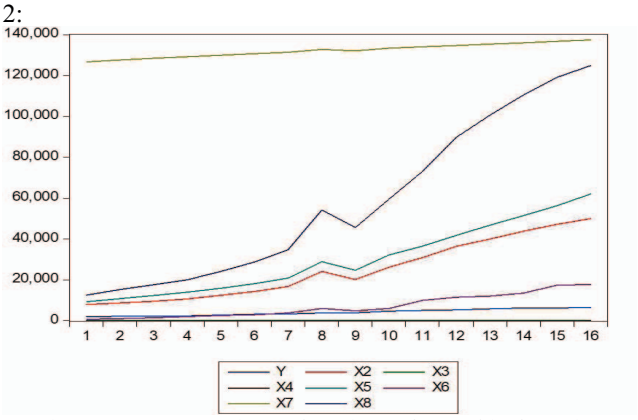

Figure 2 average price of housing sales and influencing factors

As can be seen from fig. 2, the average price of housing sales and the change direction of each influencing factor are basically the same from 2000 to 2015, and have certain correlation with each other. Using OLS estimation method to estimate parameters, a multiple linear regression model is established. The regression results are shown in table 3:

Table 3 regression results

Dependent Variable: Y
Method: Least Squares
Date: 04/09/17 Time: 22:10
Sample: 2000 2015
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-24391.41	8955.678	-2.723569	0.0261
X2	0.256975	0.182943	1.404668	0.1977
X3	-64.38251	29.55412	-2.178461	0.0610
X4	159.4830	182.1028	0.875786	0.4067
X5	-0.125582	0.069789	-1.799456	0.1096
X6	-0.005766	0.072162	-0.079907	0.9383
X7	0.247525	0.081443	3.039238	0.0161
X8	-0.019454	0.050967	-0.381707	0.7126

R-squared	0.994610	Mean dependent var	4107.938
Adjusted R-squared	0.989893	S.D. dependent var	1592.674
S.E. of regression	160.1139	Akaike info criterion	13.29650
Sum squared resid	205091.6	Schwarz criterion	13.68279
Log likelihood	-98.37200	Hannan-Quinn criter.	13.31628
F-statistic	210.8830	Durbin-Watson stat	2.661202
Prob(F-statistic)	0.000000		

As can be seen from table 3, the regression equation of the estimation model can be written as:

$$\hat{Y} = -24391.41 + 0.256975X_2 - 64.38251X_3 + 159.4830X_4 - 0.125582X_5 - 0.005766X_6 + 0.247525X_7 - 0.019454X_8$$

III. MODEL TEST

1、Goodness of fit. It can be seen from table 3: R^2 =0.9946, the modified determinable coefficient $\bar{R}^2$=0.9899, the model fitting is good.

2、F test. For H0: β2= β3= β4= β5= β6= β7= β8=0, at a given significance level α=0.05, $F_\alpha(7,9)$=3.68, F=210.8830>F(7,9)=3.68,rejecting the original hypothesis H_0, the model as a whole is significant, that is, all variables together have a significant impact on the average price of housing sales.

3、t test. For H0: β2= β3= β4= β5= β6= β7= β8=0, at a given significance level α=0.05, check the t distribution table t0.025(9)=2.262. From the operation results, only the value of t statistic corresponding to $\hat{\beta}$ 7 is greater than t0.025(9) =2.262, which indicates that the explanation

variable " population" has a significant effect on the average price of housing sales under the condition of the significance level of 0.05. The rest of the explanatory variables t statistic value is less than 2.262, in the case of other variables unchanged, the explanatory variables have no significant effect on the interpreted variable " housing sales average price" and the symbol of X_5、X_6、X_8 contrary to the expected symbol, so it is likely to exist serious multicollinearity.

4、Multiple collinearity test. First calculate the correlation coefficient matrix between the explanatory variables as shown in table 4:

Table 4 correlation coefficient matrix

Variable	x2	x3	x4	x5	x6	x7	x8
x2	1.000000	0.266445	0.138054	0.998454	0.990608	0.972230	0.999583
x3	0.266445	1.000000	0.434570	0.244778	0.249334	0.343617	0.263631
x4	0.138054	0.434570	1.000000	0.103340	0.648040	0.156306	0.141653
x5	0.998454	0.244778	0.103340	1.000000	0.991129	0.973711	0.997332
x6	0.990608	0.249334	0.648040	0.991129	1.000000	0.851831	0.991039
x7	0.972230	0.343617	0.156306	0.973711	0.851831	1.000000	0.967620
x8	0.999583	0.263631	0.141653	0.997332	0.991039	0.967620	1.000000

As can be seen from table 4, the explanatory variables X_2 and X_5、X_6、X_7、X_8, X_5 and X_6, X_6 and X_8, X_7 and X_8, the correlation coefficient between them is high, there are multiple collinearity.

For each variable as explained variables for the rest of the variables to do auxiliary regression, regression can be determined coefficient and variance expansion factor values, the results are shown in table 5:

Table 5 R2 values for auxiliary regression

Explained variable	Coefficient of determination R2	Variance inflation factor VIF=I/(1-R2)
x2	0.9998	4166.6667
x3	0.2665	1.3632
x4	0.5775	2.3666
x5	0.9988	847.4576
x6	0.9902	102.3541
x7	0.9768	43.1406
x8	0.9996	2380.9524

It can be seen from table 5 that the decisive coefficients of the auxiliary regression are very high, except for X_3、X_4, the variance expansion factor VIF ≧ 10 of the other variables can be deduced from experience that there are serious multicollinearity among the model variables.

IV. MODIFIED REGRESSION MODEL

After the model is diagnosed as having serious multicollinearity, remedial measures need to be taken to reduce multicollinearity in the model. In this paper, we use the method of eliminating variables, remove the unnecessary variables that cause multiple collinearity to test the model in turn, and get the final regression model. the operation results are shown in table 6

Table 6 correction model regression results

Dependent Variable: Y
Method: Least Squares
Date: 04/09/17 Time: 22:14
Sample: 2000 2015
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-31808.85	6473.810	-4.913467	0.0005
X3	-65.14160	29.65057	-2.196976	0.0504
X4	502.9260	162.1980	3.100691	0.0101
X6	0.108519	0.029809	3.640460	0.0039
X7	0.295098	0.053379	5.528318	0.0002

R-squared	0.989119	Mean dependent var	4107.938
Adjusted R-squared	0.985162	S.D. dependent var	1592.674
S.E. of regression	194.0033	Akaike info criterion	13.62393
Sum squared resid	414010.3	Schwarz criterion	13.86537
Log likelihood	-103.9915	Hannan-Quinn criter.	13.63630
F-statistic	249.9856	Durbin-Watson stat	1.684545
Prob(F-statistic)	0.000000		

As can be seen from table 6, the regression equation of the modified model can be written as:

$$\hat{Y} = -31808.85 - 65.14160X_3 + 502.9260X_4 + 0.108519X_6 + 0.295098X_7$$

(6473.810) (29.65057) (162.1980) (0.029809) (0.295098)

t = (-4.913467) (-2.196976) (3.100691) (3.64046) (5.528318)

R^2=0.989119 F=249.9856 n=16 DW=1.68

V. STATISTICAL TEST OF MODIFIED MODEL

1、Goodness of fit. It can be seen from table 6: R^2=0.989119, the modified determinable coefficient $\bar{R}^2$=0.985162, the model fitting is good.

2、F test. F test value is 249.9856, and the model is significant as a whole.

3、t test. At a given significance level α=0.05 or α=0.1, t0.025(9) =2.262、t0.05(9) =1.833, The estimated values

of all coefficients are significant, i.e. each explanatory variable has a significant influence on the explained variable.

4、Heteroscedasticity test. White test results are shown in table 7:

Table 7 White test

Heteroskedasticity Test: White			
F-statistic	0.854344	Prob. F(12,3)	0.6384
Obs*R-squared	12.37794	Prob. Chi-Square(12)	0.4158
Scaled explained SS	3.205145	Prob. Chi-Square(12)	0.9939

As can be seen from table 7, nR^2=12.37794, at a given significance level α =0.05, $\chi^2_{0.05}(4)$ =9.48773, because nR^2=12.37794> $\chi^2_{0.05}(4)$ =9.48773, reject the original assumption that H0: the model has variance. So the model does not have variance.

5、Auto correlation test. From table 6, DW = 1.6845, the sample data n = 16, k = 4, and dl = 0.734, du = 1.935, it is not possible to determine whether the model has autocorrelation problems because 0.734<DW<1.935. In this paper, LM test is used to further determine whether the model has autocorrelation. The test results are shown in 8:

Table 8 LM test

Breusch-Godfrey Serial Correlation LM Test:			
F-statistic	0.379910	Prob. F(2,9)	0.6944
Obs*R-squared	1.245628	Prob. Chi-Square(2)	0.5364

As can be seen from table 8, LM=nR^2=1.245628, the p value is 0.5364, at a given significance level a = 0.05, reject the original hypothesis H0: model autocorrelation, do not reject the equipment hypothesis, that is, the resulting modified model does not have autocorrelation problems.

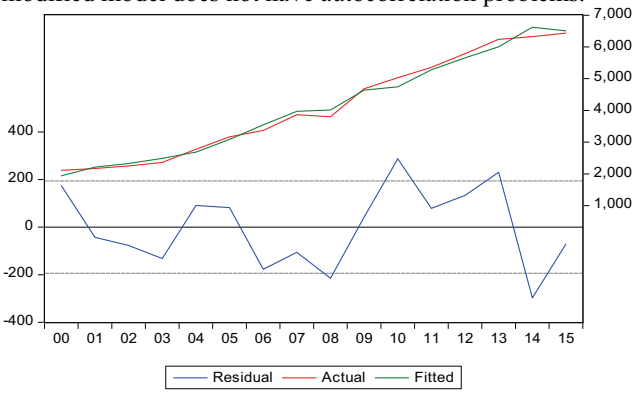

Figure 3 fitting effect

After the initial model is modified, all the explanatory variables of the resulting model are in agreement with expectations. Fig. 3 is a fitting effect diagram of the modified model showing the changing trend of fitting values and real values and the relationship between the two. The coefficient estimates are explained as follows: if the price level increases by 1 %, the average price of China 's housing sales declines by 65.14160 yuan; If the loan interest rate increases by 1 %, the average price of China 's housing sales increases by 502.9260 yuan; If the real estate land purchase fee (land price) every 100 million yuan, the average increase in China 's housing and sales price of 0.108 519 yuan; If China 's population increased by 10,000 people, the average price of housing sales in China increased by 0.295,098 yuan.

VI. POLICY RECOMMENDATIONS

1, Control the number of urban population and land supply. As the population increases, the demand for housing also increases; demand will make housing sales average rising. With the continuous development of society, more and more people yearn for the life of big cities, into a second-tier cities. Therefore, the government should put forward corresponding policies to develop four or five-tier cities, control the scale of development and population in urban areas, effectively control the flow of population in urban areas, so as to suppress the sustained growth of housing prices to make a modest contribution.

Land acquisition cost is a part of the cost of real estate development, is the variable cost of real estate development, directly affect the price of housing, the price changes and land supply has a great relationship. For small land supply and large demand areas, local governments will increase land prices to achieve sustainable urban development, land price increases will affect real estate prices through the transmission mechanism. So the local government can indirectly control the real estate price by controlling the supply of land.

2, Increase the macro-control efforts. Real estate industry is an important pillar of China 's national economy, a certain period of time the growth trend of housing prices will not change, in order to avoid the sustained rapid growth of housing prices and the emergence of real estate bubbles, the government should increase macroeconomic regulation and control efforts. Such as adjusting the loan interest rate, adjusting the money supply, indirect and effective control of real estate prices.

3, Strengthen supervision and enhance market transparency. Establish a sound real estate information system, as far as possible to make the public real estate information reasonable, effective and accurate. The establishment of a reasonable regulatory system to ensure the effective disclosure of real estate information, increase market transparency, so that buyers to make the right investment decisions, but also to guide consumer expectations. The government to strengthen the supervision of the real estate market, curb illegal behavior, effectively protect the rights of both buyers and sellers, safeguard the interests of both buyers and sellers, to create a healthy and good real estate market.

REFERENCES

[1] Wang Qianyi., 'Harbin real estate price impact factor research', Northeast agricultural university, June 2016

[2] Luan Tianyi, ‘regression analysis of the main urban housing price factors in China’, business economy, 2016, no 1, p. 5

[3] Ning Baoquan., ‘multivariant linear regression analysis based on Eviews software’, journal of six water teacher college, 2011, no 3, p. 11

[4] Liu Ximei, ‘prediction and analysis of real estate prices in Chongqing’, Chongqing normal university, June 2014

[5] Hao Danlu., ‘China real estate price impact factor research’, Jilin university, April 2014

[6] An Hui, Wang Ruidong, ‘China 's real estate price factors empirical analysis’, economic and economic weft', 2013, no 3, p. 115

[7] Li Ying, ‘prediction of rural cigarette sales based on time series and multiple linear regression model’, Yunnan university, June 2015

[8] Ding Jie., ‘China 's real estate price impact factors empirical analysis’, central China normal university, May 2014