ORIGINAL ARTICLE

Statistical Analysis with Hedonic Pricing Models: Estimation of Real-Estate Values and Anticipation of Housing Price in Beijing, China

Haoxuan Lyu[*], **Zeyuan Liu**

The Australian National University, Canberra 2600, Australia. E-mail: u5954654@anu.edu.au

Abstract: On the basis of big data regarding the sold house price in Beijing from 2011 to 2017, this research aims to exhibit a comprehensive and rational deducing progress in character with the justified and formulated variables and visible results to be instrumental in the prediction of housing price in Beijing, and it attributes the dominated factors amongst all dependent variables. In this paper, the hedonic price model is the primary analyzing tool to cope with the intrinsic factors in real-estate estimation and prediction.

Keywords: Hedonic Price Models (HPMs); House Price Prediction; Real-Estate Estimation

1. Introduction

Effective price is a potent guide for decision or policy makers of a corporation to narrow down the divergence between customers' expectations and the company's lucrativeness. It is also why research topics about effective price have been the core interests in the area of econometrics and its related fields. Hedonic price models are one of the most applicable and efficacious measurements of estimating the value and obtaining an effective price. The intrinsic characteristics of hedonic price models are to qualify the definition of convoluted dependent variables that might make an impact on the study object to much the evidently distinct extent. Naturally, it explains the properties of product parameters and their probable combinations[1]. This is also why it has been prevailingly utilized in the hospitality and real estate industry ever since the theoretical basis has been built by [2] and [3]. Before delving into any analyzing particle, this research and its comparative parameters should be defined within qualified and quantified boundaries with some reasonable justifications as followings.

At this point in research, the dependent variables should be defined and justified within merely the internal factors of house, and the research objects themselves are not compassing external factors, such as monetary policy regulation on commercial housing price, building outdoor environment and international & domestic economy marketing performance.

The time scope of this project mainly focuses on the sold houses during the past ten years. Any time beyond this range has been so peculiarly lower than people's expectations in accordance with the viewpoint of nowadays' buyer. Basically, because of the exceptional development of Beijing in its internationalized process, its house price has burgeoned drastically as a corollary of China's economy steady-state growth.

Every recommendation and prediction made at the end of this paper is only meaningfully significant for the next two or three years in case the accuracy cannot remain a delicate level once updated dependents are taken into considera-

doi: 10.18686/fm.v5i3.2499

tion. In other words, the formulaic conclusion drawn in this paper is not an axiomatic equation; instead a mutable function commensurable with disparate circumstances.

2. Literature review

The study in the area of Hedonics is scintillating in the statistical literature for the past twenty years, and it is compactly affiliated with the forecast of real-estate and property. As a standardized econometric tool for estimation, the hedonic pricing model has been the fundamental methodology to determine the inherent dependent variables of the study object[4]. For instance, the hedonic pricing model provides many perspective judgements about the characteristics of research objects[3,5]. Apart from that, there is more available information on the returns to housing investment due to the work done by Case and Shiller after introducing the hedonic pricing model[6,7]. The development of hedonic studies had been instantiated, in further, by the stimulation of the complicated statistical situation in various industries. In a conclusion of this model, it makes an assumption that different attribution of different dependent variables services the price of study object as a function listed as below to varying degrees[8,9].

$$P = f(x_i) \tag{2.1}$$

$$P(RE) = f(S) + f(L) + \varepsilon \tag{2.2}$$

Where P(RE) is a function of housing price affected by several contributors. There are two functional forms of attributes, the structural and locational ones.

3. Methodology

3.1 Data source

The raw data was collected by Lianjia.com with its open-source database in Beijing from 2011 to 2017. Lianjia.com is a domestic leader in China and a highly vertical real-estate service platform for the entire industrial chain in the property business, which integrates house source information search, product development, big data processing and service standard establishment. Considering what a renowned company and a house sold platform Lianjia.com is, driving data from this company is virtually incontrovertible concerning the authenticity of data. Besides, averting any present interests' conflict so as to choose data from slight behind nowadays' trend shall guarantee the objectivity of data to the greatest extent. Most notably, the topic of this paper is to exhibit the rational process of data analysis rather than to focus on the conclusion of prediction and recommendation. After all, the dependent variables contributing to the study object remain versatile swiftly from year to year. Nonetheless, the analyzing method to cope with these similar problems should be fixed within a relatively immutable frame, which is also what this paper pursues. More concretely, the table below makes a demonstration of dependent variables this paper is taking into consideration.

Variables	**Description**
totalPrice	The total price of house (in 10000RMBs)
square	The square of house (in m^2)
livingRoom	The number of living rooms
drawingRoom	The number of drawing rooms
kitchen	The number of kitchens
bathRoom	The number of bathrooms
floor	The height of the house
constructionTime	The time of construction
buildingType	Tower = 1
	Bungalow = 2
	Combination of plate and tower = 3

Table 1. (continued).

Variables	Description
buildingType	Plate = 4
renovationCondition	Other = 1
	Rough = 2
	Simplicity = 3
	Hardcover = 4
buildingStructure	Unknow = 1
	Mixed = 2
	Brick and wood = 3
	Brick and concrete = 4
	Steel = 5
	Steel-concrete composite = 6.
ladderRatio	The proportion between the number of residents on the same floor and the number of elevators of ladder. It describes how many ladders a resident has on average.
elevator	Have elevators = 1
	Otherwise = 0
fiveYearsProperty	The owner has the property for less than 5 years = 1
	Otherwise = 0
subway	The house has subway stops within 1 kilometre = 1
	Otherwise = 0

Table 1. Variable definition and descriptive

3.2 Research method

The raw data is sorted and qualified through the excel and python respectively and by sifting out outliers and incomplete data on account of the succinct observation, such as house price extremely lower than normal market expectation. The major data process is carried by R-code to find out what the interaction is amongst all dependent variables and how much the contribution each dependent variable makes to the research topic. As for how to winnow out those irrelevant dependent factors from the raw data, the paper is using some published fairly-established opinions for reference, which treats real estate as a heterogeneous good with various attributions ranging from locational and structural to spatial attributions[10,11].

Here are some formulaic-theoretical subgrades for later analysis.

$$\mathrm{Cov}\,(X,Y) = \frac{\Sigma(X_i - \overline{X})((Y_j - \overline{Y}))}{n} \quad (3.2.1)$$

$$\ln P = \alpha_0 + \beta_K X_{aiK} + \beta_K X_{njK} + \beta_K X_{stK} + \varepsilon \quad (3.2.2)$$

$$H_0: \frac{\sigma^2_{addition}}{\sigma^2_{error}} = 1 \qquad H_1: \frac{\sigma^2_{addition}}{\sigma^2_{error}} > 1 \quad (3.2.3.1\ \&\ 3.2.3.2)$$

$$F_{obs} = \frac{\frac{SS_{addition}}{df_{addition}}}{\frac{SS_{error}}{df_{error}}} \sim F_{df_{addition},df_{error}} \quad (3.2.4)$$

4. Discussion and analysis

This section exhibits the detailed procedure of data processing with the in-depth analysis of their corresponding statistical significance by applying data collected in the year of 2011 via R-code. Next section will list the rest results and sketchy explanation in the same statistical principle.

4.1 Summary of data

4.1.1 Digital analysis

This section presents the prosperities of variables via the table of succinct figures and picturable method, which defines the size of data set in a general way.

totalPrice		**square**		**livingRoom**		**drawingRoom**	
Min.	4.0	Min.	20.00	Min.	0.000	Min.	0.000
1st Qu.	113.0	1st Qu.	58.49	1st Qu.	2.000	1st Qu.	1.000
Median	155.0	Median	75.61	Median	2.000	Median	1.000
Mean	181.9	Mean	84.85	Mean	2.043	Mean	1.216
3rd Qu.	216.0	3rd Qu.	102.23	3rd Qu.	3.000	3rd Qu.	2.000
Max.	1420.0	Max.	497.65	Max.	6.000	Max.	3.000
kitchen		**bathRoom**		**buildingType**		**constructionTime**	
Min.	0.0000	Min.	0.000	Min.	1.000	Min.	1954
1st Qu.	1.0000	1st Qu.	1.000	1st Qu.	1.000	1st Qu.	1994
Median	1.0000	Median	1.000	Median	4.000	Median	2001
Mean	0.9849	Mean	1.196	Mean	3.059	Mean	1999
3rd Qu.	1.0000	3rd Qu.	1.000	3rd Qu.	4.000	3rd Qu.	2004
Max.	3.0000	Max.	5.000	Max.	4.000	Max.	2015
renovationCondition		**buildingStructure**		**ladderRatio**		**elevator**	
Min.	1.000	Min.	2.000	Min.	0.0140	Min.	0.0000
1st Qu.	1.000	1st Qu.	2.000	1st Qu.	0.2500	1st Qu.	0.0000
Median	1.000	Median	6.000	Median	0.3330	Median	1.0000
Mean	1.014	Mean	4.299	Mean	0.3843	Mean	0.5423
3rd Qu.	1.000	3rd Qu.	6.000	3rd Qu.	0.5000	3rd Qu.	1.0000
Max.	4.000	Max.	6.000	Max.	2.0000	Max.	1.0000
fiveYearsProperty		**subway**					
Min.	0.0000	Min.	0.0000				
1st Qu.	0.0000	1st Qu.	0.0000				
Median	0.0000	Median	1.0000				
Mean	0.4798	Mean	0.6174				
3rd Qu.	1.0000	3rd Qu.	1.0000				
Max.	1.0000	Max.	1.0000				

Table 2. Summary of variables with their statistical prosperities

The graphic presentation exhibits the distribution and location of statistical points more vividly and directly.

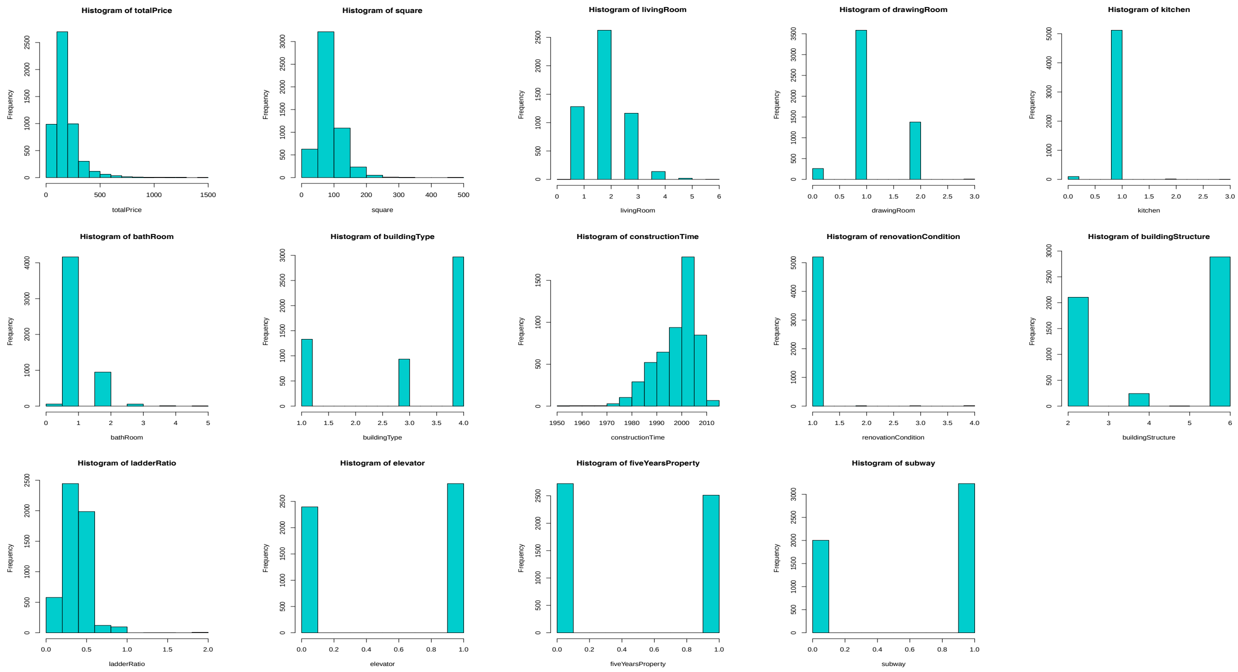

Figure 1. Graphic summary of variables with their statistical prosperities.

4.1.2 Covariance analysis amongst variables

By applying the Equation 3.2.1 computing similarity between the [12, 13], and the Pearson correlation coefficient, it is to estimate the correlation amongst all variables. The ideal case is that none of the covariances is higher than 0.8 in that the problem of multicollinearity can be neglected during the next-stage analysis. **Table 3** depicts the results of co-variance analysis.

	totalPrice	*square*	*livingRoom*	*drawingRoom*	*kitchen*	*bathRoom*
totalPrice	1.0000000	0.54924719	0.4236551	0.3526009	0.11530578	0.4351027
square	0.5492472	1.00000000	0.7124847	0.6065995	0.08375687	0.7250360
livingRoom	0.4236551	0.71248470	1.0000000	0.4834582	0.14821643	0.5536612
drawingRoom	0.3526009	0.60659954	0.4834582	1.0000000	0.22380701	0.5822231
kitchen	0.1153058	0.08375687	0.1482164	0.2238070	1.00000000	0.2336145
bathRoom	0.4351027	0.72503598	0.5536612	0.5822231	0.23361454	1.00000000

Table 3. Summary of covariance

4.2 Process of fitting model

Presenting as the examination of the residual plot, the regression models fitting between the original dependent variable, total price and the rest independent variables are listed below.

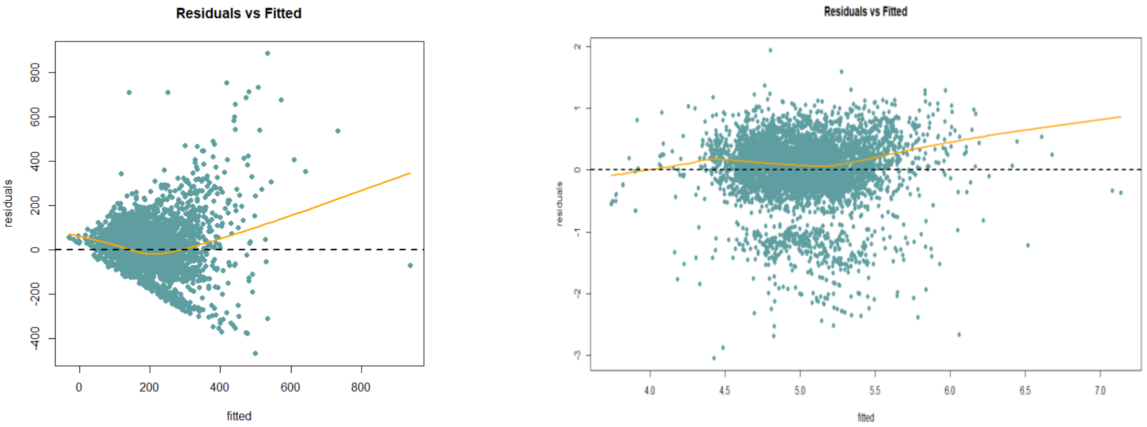

Figure 2. Regression fitting lines before and after logarithms transformation.

The y-axis in figures above represents Real Value – Fitted Value (= Residual). And also, the x-axis in figures above denotes as the predicted Y by substituting real X (= Fitted Value). Apparently, it has a quadratic trend in Pic 4.2.1. Hence, the residuals are not independently distributed in the original model, which violates the assumption of the multi-linear regression model. In other words, it is necessary to apply logarithms transformation to Y to generate Pic 4.2.2. In plot 4.2.2, the residuals are identically and independently distributed roughly. Therefore, the assumption is not violated. Fitting line in the second picture is approaching to the line of y = 0, which proves a progressive improvement.

4.3 Overall test

The fitting-goodness of a linear regression model can be examined by the F-test of overall significance by using Equation 3.2.3.1 & 3.2.3.2 and 3.2.4. As a result, corresponding F-value of the test is 149.5 and its p-value is perfectly convergent to zero less than 0.05. Therefore, the model has its statistical significance, dependent and independent equipping with a linear relationship, passing the F-test. Taking all dependent variables into consideration as a fundamental model, the table below concludes the details about dependent variables with their coefficients and other characteristics.

	Estimate	*Std. Error*	*t value*	*Pr (>|t|)*	*Significance* [a]
Intercept	23.6983968	2.3988710	9.751	< 2e-16	***
square	0.0045270	0.0004119	10.889	< 2e-16	***
livingRoom	0.0970412	0.0152448	6.347	2.39e-10	***
drawingRoom	0.1017275	0.0189720	5.401	6.92e-08	***
kitchen	0.4631596	0.0553361	8.296	< 2e-16	***
bathRoom	-0.0631619	0.0248823	-2.428	0.0152	*
buildingType	-0.0103583	0.0086155	-1.202	0.2293	
constructionTime	-0.0101515	0.0012070	-8.410	< 2e-16	***
renovationCondition	0.2182232	0.0420448	5.190	2.18e-07	***
buildingStructure	0.0865553	0.0063488	13.633	< 2e-16	***
ladderRatio	0.2229062	0.0529977	4.206	2.64e-05	***
elevator	0.0645851	0.0284432	2.271	0.0232	*
fiveYearsProperty	0.1529519	0.0163096	9.378	< 2e-16	***
subway	-0.0010779	0.0152536	-0.071	0.9437	

Table 4. Coefficients of variables

[a] Significance: 0 -- '***'; 0.001 -- '**'; 0.01 -- '*'; 0.05 -- '.'; 0.1 -- '_'; 1 -- '_'.

F- statistic = 149.5 on 13 and 5221 degree of freedom; P – value < 2.2e-16

Besides, from this table above, it can be observed that the variables, except “building type” and “subway”, are significantly important at the significance level of 0.05 if treated each dependent variable individually.

Hence, the aggregated relationship is shown below:

l*og(totalPrice) ~ (square + livingRoom + drawingRoom + kitchen + bathRoom + constructionTime + renovationCondition + buildingStructure + ladderRatio + elevator + fiveYearsProperty) + (buildingType + subway)*

4.4 Process of improving model

Applying nested F-test is to check whether the co-impact of these two insignificant variables can be neglected or not. Here is the result of nested F-test.

	RES.DF	RSS	DF	SUM OF	SQ	F~PR(>F)
[A]**MODEL_1**	5223	1501.3				
[B] **MODEL_2**	5221	1500.9	2	0.41685	0.725	0.4844

Table 5. Simulation results about nested F-test ANOVA in R

[a] Model_1: Nested model deleting two dependent variables.

[b] Model_2: Originally full model.

Reading from the table above, P = 0.4844;

The improvement has been made toward this modelling under such two assumptions:

H0: The two models before and after improving, are the same. In other words, $\beta_{buildingType} = \beta_{subway}= 0$.

H1: The two models before and after improving, are significantly different. In other words, at least one of the parameters in β does not equal zero.

As the p-value is greater than 0.05, we cannot reject the null (H0) but reject H1. So, the model_2 after appending two dependent variables is not a significant improvement. In a word, $\beta_{buildingType} = \beta_{subway}= 0$.

	Estimate	*Std. Error*	*t value*	*Pr (>|t|)*	*Significance* [a]
Intercept	23.6983968	2.3988710	9.879	< 2e-16	***
square	0.0045270	0.0004119	10.990	< 2e-16	***
livingRoom	0.0970412	0.0152448	6.366	2.11e-10	***
drawingRoom	0.1017275	0.0189720	5.362	8.59e-08	***
kitchen	0.4631596	0.0553361	8.370	< 2e-16	***
bathRoom	-0.0631619	0.0248823	-2.538	0.01116	*
constructionTime	-0.0102846	0.0012018	-8.558	< 2e-16	***
renovationCondition	0.2171200	0.0420324	5.166	2.49e-07	***
buildingStructure	0.0869723	0.0063387	13.721	< 2e-16	***
ladderRatio	0.2007084	0.0496517	4.042	5.37e-05	***
elevator	0.0799994	0.0253784	3.152	0.00163	**
fiveYearsProperty	0.1543686	0.0162652	9.491	< 2e-16	***

Table 6. Coefficients of variables after adjustment

[a] Significance: 0 -- ‘***’; 0.001 -- ‘**’; 0.01 -- ‘*’; 0.05 -- ‘.’; 0.1 -- ‘_’; 1 -- ‘_’.

F- statistic = 176.6 on 11 and 5223 degree of freedom; P – value < 2.2e-16

Then the final model is followings:

log(totalPrice) ~ 0.0045270square + 0.0970412livingRoom + 0.1017275drawingRoom + 0.4631596kitchen + -0.0631619bathRoom + -0.0102846constructionTime + 0.2171200renovationCondition + 0.0869723buildingStructure + 0.2007084ladderRatio + 0.0799994elevator +0.1543686fiveYearsProperty + 23.6983968

5. Result exhibitions of seven-year data

The data processing of other years, instead of 2011, is similar to what has been exhibited in 2011 with distinct results due to the very various data sets. In this section, there are going to present a summary of models in the rest years.

2012:

log(totalPrice) ~ 0.0073992square + 0.0738736livingRoom + 0.0577131drawingRoom + 0.1295710kitchen + -0.0547004bathRoom + -0.0099324constructionTime + 0.0419271buildingStructure + 0.0476998ladderRatio + 0.1112577elevator + 0.0761803fiveYearsProperty + -0.0062817buildingType + 23.8837765

2013:

log(totalPrice) ~ 0.0065911square + 0.071924livingRoom + 0.0659200drawingRoom + 0.1918934kitchen + -0.0756613bathRoom + -0.0110727constructionTime + 0.0382882buildingStructure + 0.0963845ladderRatio + 0.1347686elevator + 0.1046632fiveYearsProperty + 26.4327825

2014:

log(totalPrice) ~ 0.0065326square + 0.0958597livingRoom + 0.0653894drawingRoom + 0.1275215kitchen + -0.0309858bathRoom + -0.0088322constructionTime + -0.0121253renovationCondition + 0.0275171buildingStructure + 0.1453614elevator + 0.0657031fiveYearsProperty + 0.0097885buildingType + 22.0769707

2015:

log(totalPrice) ~ 6.662e-03square + 8.814e-02livingRoom + 8.120e-02drawingRoom + 1.947e-01kitchen + -2.680e-02bathRoom + -1.281e-02constructionTime + 1.949e-02renovationCondition + 3.537e-02buildingStructure + 7.324e-02ladderRatio + 1.571e-01elevator + 5.382e-02fiveYearsProperty + 7.412e-03buildingType + 2.984e+01

2016:

log(totalPrice) ~ 6.326e-03square + 7.513e-02livingRoom + 1.019e-01drawingRoom + 2.836e-01kitchen + -4.320e-02bathRoom + -1.410e-02constructionTime + 4.529e-02renovationCondition + 2.892e-02buildingStructure + 1.471e-01 ladderRatio + 1.867e-01 elevator + -2.300e-02 fiveYearsProperty+ 6.107e-03buildingType + 3.260e+01

2017:

log(totalPrice) ~ 6.871e-03square + 7.119e-02livingRoom + 8.931e-02drawingRoom + 3.341e-01kitchen + -2.970e-02bathRoom + -1.243e-02constructionTime + 1.502e-02renovationCondition + 2.012e-02buildingStructure + 1.881e-01elevator + 2.518e-02fiveYearsProperty + 1.004e-02buildingType + 2.963e+01

6. Conclusion

From the results of high-volume data, some dependent variables had been playing an essential role in the house price estimation, such as the square of house, the number of living rooms, the number of drawing rooms, the number of kitchens, the number of bathrooms and the height of the house, and they kept appearing as determining terms in the equation of estimation in every year from 2011 to 2017. In another word, they predominantly influenced the house price in Beijing for the past decades. In the near future, they are still worth paying close attention to. Whereas, some factors might not be as important as the ones aforementioned, like the type of buildings and the condition of renovation. They might make a difference in the house price from time to time during the past years. But they are definitely not the priority to confirm every time when making a marketing prediction. Last, the factor of the subway distribution is never statistical meaningful towards this model, which means that it can be confidently ignored when a market analysis is proposed in the next few years. As the fluctuation of preference of housing purchasing, this model only has its statistical meaning within a reasonable time span not eternally.

This paper highly focuses on the interpretation of data processing, not the results itself. By substituting the different category of data in the same principle, there is always a relative sound conclusion drawn to be a recommendation.

Acknowledgement

We would like to show our gratitude to the Dr. Xuezheng Qin, professor and deputy dean in the School of Economics at Peking University for sharing their pearls of wisdom with us during the course of this research and guiding us to find a more statistical-sense topic. We also want to thank Doctoral Candidate Yanran Chen in the School of Econom-

ics at Peking University for sharing her statistical coding experience.

References

1. Yim ES, Lee S, Kim WG. Determinants of a restaurant average meal price: An application of the hedonic pricing model. International Journal of Hospitality Management 2014; 39: 11–20.
2. Lancaster KJ. A new approach to consumer theory. Mathematical Models in Marketing 1976; 132: 106-107.
3. Rosen S. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 1974; 82(1): 34–55.
4. Malpezzi S. Hedonic pricing models: A selective and applied review. Housing Economics and Public Policy 2008.
5. Triplett JE. Automobiles and hedonic quality measurement. University Chicago Press 1969; 77(3): 408–417.
6. Case KE, Shiller RJ. Forecasting prices and excess returns in the housing market. Real Estate Economics 1990; 18(3): 253–273.
7. Case KE, Shiller RJ. The efficiency of the market for single family homes. Am. Econ. Rev. 1989.
8. Thrane C. Examining the determinants of room rates for hotels in capital cities: The Oslo experience. Journal of Revenue & Pricing Management 2007; 5(4): 315–323.
9. Thrane C. Hedonic price models and sun-and-beach package tours: The Norwegian case. Journal of Travel Research 2005; 43(3): 302–308.
10. Orford S. Valuing locational externalities: A GIS and multilevel modelling approach. Environment & Planning B Planning & Design 2002; 29(1): 105–127.
11. Can A. Specification and estimation of hedonic housing price models. Regional Science & Urban Economics 1992; 22(3): 453–474.
12. Hall MA. Correlation-based feature selection for machine learning. 1999.
13. Katrutsa A, Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Systems with Applications 2017; 76: 1–11.