Modeling real estate value by means of parametric and nonparametric approaches

M A Denisov[1], O A Ikonnikov[1,2], S N Ezhemanskaya[1] and L N Korpacheva[1]

[1]Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russia

[2]Siberian State University of Science and Technology, 31, Krasnoyarskiy Rabochiy Ave., Krasnoyarsk, 660037, Russia

E-mail: max_denisov00@mail.ru

Abstract. The paper considers the problem of estimating a sample that contains data characterizing one-room apartments in Krasnoyarsk. Two approaches are described: parametric and nonparametric. A linear regression model is used as a parametric structure. Non-parametric methods are implemented using the Nadaraya-Watson kernel estimate. The obtained results are compared as a result of which a conclusion is formulated about which model better approximates the original sample.

1. Introduction

Real estate plays a big part in human life and appears to be as a valuable economic resource. In this regard, the problem of modeling real estate value is quite relevant. On the basis of constructed models it is possible, for instance, to solve problems of forecasting its price [1, 2], which is high-demand in a rapidly developing and changing economy.

The problems of estimating the value of real estate are often solved using parametric regression models [3, 4]. Parametric modeling is used when priori data contains information about the structure of the object under study. Such tasks are also called identification in the narrow sense. However, there is another approach which is called non-parametric. It refers to identification in a broad sense. In this case, the amount of information given priori is relatively small. The structure of the object accurate to a parameter vector is unknown as well as the type of distribution.

The authors of this paper propose to construct a parametric and non-parametric model and compare obtained results. Priori information includes a sample of the characteristics of one-room apartments in Krasnoyarsk. Previously, non-parametric estimation using Nadaraya-Watson kernel function [5] for the above-mentioned data has already been carried out in [6]. However, the comparison with the parametric approach was not implemented. That is the novelty of this study.

2. Problem statement

There is a sample of observations that describes one-room apartments of the Krasnoyarsk city including several characteristics. The total area of the apartment, the area of the kitchen, as well as the price are quantitative attributes. Area, floor, walling material, type of layout are qualitative attributes. The data is as follows:

Table 1. Initial sample.

№	District	Total area, m^2	Kitchen area, m^2	Floor	Walling material	Layout	Price, million rub.
1	1	33	10	1	1	4	1.900
2	1	42	9	1	1	4	2.450
3	1	32	6	1	1	1	1.900
4	1	26	9	1	3	5	1.900
5	1	30	6	1	2	1	1.700
…	…	…	…	…	…	…	…
1359	8	31	6	2	1	6	1.650

Modern information technologies allow us to build models of both linear and non-linear regression. But the methods for constructing linear models are much simpler and more reliable. They impose less stringent requirements on the amount of initial information and are better adapted to consider possible dependencies between parameters. In this regard, it was decided to use a linear parametric model of multifactor regression, the structure of which is presented in general form below:

$$\hat{x} = \hat{b}_0 + \sum_{i=1}^{n}\sum_{j=1}^{m}\hat{b}_j u_{ji} , \tag{1}$$

where n is sample size, m is dimensions number of the problem, $\{\hat{b}_j, j=\overline{0,m}\} = \vec{b}$ is vector of parameters that need to be evaluated and $\{\hat{u}_{ji}, j=\overline{1,m}, i=\overline{1,n}\} = \vec{U}$ is vector of independent factor attributes.

When priori information about the object of study is not enough, non-parametric modeling methods are used (black-box problems). There are large number of algorithms for implementing this approach [7]. In this paper, the authors propose to use the non-parametric Nadaraya-Watson estimate for the multidimensional attribute space, the mathematical description of which is presented in general form below:

$$\hat{x} = \sum_{i=1}^{n} x_i \prod_{j=1}^{m} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right) \Big/ \prod_{j=1}^{m} \Phi\left(\frac{u_j - u_{ji}}{c_s}\right) \tag{2}$$

where $\{x_i, u_i, i=\overline{1,n}, j=\overline{1,m}\}$ initial sample, c_s is kernel bandwidth and $\Phi(\bullet)$ is bell-shaped kernel function.

3. Solution methods

Computational experiment involves constructing two types of models using data in table 1. The first model is parametric. As it was mentioned above the mathematical structure (1) is used. Its mathematical description looks as follows:

$$\hat{x}_1 = \sum_{i=1}^{n}(\hat{b}_0 + \hat{b}_1 u_{1i} + \hat{b}_2 u_{2i} + \hat{b}_3 u_{3i} + \hat{b}_4 u_{4i} + \hat{b}_5 u_{5i} + \hat{b}_6 u_{6i}) , \tag{3}$$

where u_1 is total area, u_2 is kitchen area, u_3 is district, u_4 is floor, u_5 is walling material of an apartment and u_6 is layout.

To find parameters vector $\vec{b}$ it is used the method based on minimizing the criterion of least squares (least squares method). Mathematical interpretation of criterion looks as follows:

$$F = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_{1i})^2 \to \min_{\vec{b}} , \tag{4}$$

where x_i is object output (apartment price) and $\hat{x}_1$ is output of parametric model (3).

The non-parametric model, considering multidimensional attribute space of the sample, looks as follows:

$$\hat{x}_2 = \sum_{i=1}^{n} x_i \prod_{j=1}^{2} \Phi^{(q)}\left(\frac{u_j - u_{ji}}{c_s}\right) \prod_{j=3}^{6} \Phi^{(n)}(u_j - u_{ji}) \Big/ \sum_{i=1}^{n} \prod_{j=1}^{2} \Phi^{(q)}\left(\frac{u_j - u_{ji}}{c_s}\right) \prod_{j=3}^{6} \Phi^{(n)}(u_j - u_{ji}), \quad (5)$$

where bandwidth parameter c_s is configured by cross-validation method; for quantitative characteristics it is used the parabolic kernel function:

$$\Phi^{(q)} = \begin{cases} (0.75 \cdot (1 - |z|)^2, & |z| < 1, \\ 0, & |z| \geq 1, \end{cases} \quad (6)$$

where $z = (u_j - u_{ji})/c_s$ but for nominal characteristics it is used following kernel:

$$\Phi^{(n)} = \begin{cases} 1, & u = u_i, \\ 0, & u \neq u_i. \end{cases} \quad (7)$$

Model accuracy of the obtained results is calculated using relative error value:

$$\delta_k = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - \hat{x}_{ki}}{\hat{x}_{ki}} \right| \cdot 100\%, \quad (8)$$

where k=1 is for parametric model (3) and k=2 is for nonparametric model (5).

4. Computational experiment

At the first stage parametric modeling of the sample of observations using the least squares method is implemented. To do this, the system of equations in accordance with the necessary condition of a minimum existence for the least squares criterion (4) using the values of all attributes presented in table 1 is to be solved. As a result, we obtain the estimates cost values of apartments – $\hat{x}_{1i}, i = \overline{1, n}$, that are illustrated below combines with the genuine apartments prices:

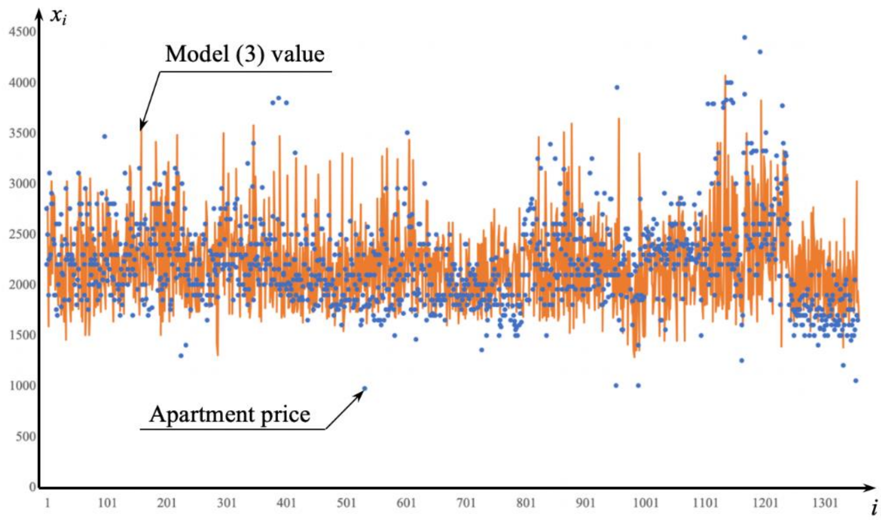

Figure 1. Genuine and parametric model (3) values of apartment prices (i – apartment number).

In figure 1 it can be seen that the model approximates the points of the object quite accurately. The value of simulation accuracy is calculated using formula (8) and is equal to δ_1 = 13.02% which also confirms the conclusion made above. The estimates values of the parametric model coefficients are as follows: $\hat{b}_0 = 153.7,\ \hat{b}_1 = 85.6,\ \hat{b}_2 = 47.7,\ \hat{b}_3 = 37.7,\ \hat{b}_4 = -193.6,\ \hat{b}_5 = -1.1,\ \hat{b}_6 = -11.3$.

Next, it is conducted non-parametric modeling of the data by formula (5). The results are presented in the figure below:

Figure 2. Genuine and nonparametric model (5) values of apartment prices (i – apartment number).

The accuracy value of the computational modeling experiment using the formula (8) is equal to δ_2 = 7.28%. In figure 2 it can be seen that for some residential objects the model value was not calculated (the place on the graph where model values (5) equal zero). This is due to the fact that the attribute u_1 (the total area) imposes some restrictions on the points entering the kernel function due to a small sample size. However, the accuracy of approximation (8) is slightly higher than in parametric modeling.

5. Conclusion

In this paper it is considered two approaches to modeling the cost of real estate. For the parametric method the linear regression structure which accurately approximated the sample of observations has been chosen. However, the non-parametric model has coped with the task more precisely. Surely, the difference is not so great but it should be kept in mind that the object structure was not set as priori information. Having such a research objective, it is more typical to use non-parametric modeling methods. Authors would like to add that the linear regression model is not always able to describe the data accurately, and there is always a chance that some non-linear dependencies will not be taken into account.

References

[1] Beracha E and Wintoki M B 2013 *Journal of Real Estate Research* **35** pp 283-312

[2] Pace R K, Barry R, Gilley O W and Sirmans C F 2000 *International Journal of Forecasting* vol. **16** pp 229-46

[3] D S Parygin, Malikov V P, Golubev A V, Sadovnikova N P, Petrova T M and Finogeev A G 2018 *J. Phys.: Conf. Ser.* **1015** 032102

[4] Manjula R, Shubham J, Sharad S and Pranav R K 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **263** 042098

[5] Nadaraya E A 1964 On estimating regression *Theory of Probability and Its Applications* vol 9

pp 157-9
[6] Denisov M A, Korneeva A A and Ikonnikov O A 2017 *Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis (Krasnoyarsk)* (Novosibirsk: NSTU publisher) pp 28-34
[7] Hollander M and Wolfe D A 1999 *Nonparametric statistical methods* (New York: Wiley–Interscience)