Comparative modelling analysis in the applications of parametric and nonparametric approaches

M A Denisov¹, O A Ikonnikov¹², S N Ezhemanskaya¹ and L N Korpacheva¹

¹Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russia
²Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarskiy Rabochnyi 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. Nonparametric 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 modelling 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 modelling 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 nonparametric. 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 nonparametric model and compare obtained results. Priori information includes a sample of the characteristics of one-room apartments in Krasnoyarsk. Previously, nonparametric 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, m² | Kitchen area, m² | 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     | 3                | 5      | 1.900              |
| 4  | 1        | 26             | 9                | 1     | 2                | 1      | 1.700              |
| 5  | 1        | 30             | 6                | 1     | 1                | 6      | 1.650              |

Modern information technologies allow us to build models of both linear and nonlinear 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:

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

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

When priori information about the object of study is not enough, nonparametric modelling 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 nonparametric Nadaraya-Watson estimate for the multidimensional attribute space, the mathematical description of which is presented in general form below:

$$
\hat{x} = \sum_{i=1}^{n} x_i \prod_{j=1}^{m} \Phi\left(\frac{u_j - \hat{u}_{ji}}{c_s}\right) / \prod_{j=1}^{m} \Phi\left(\frac{u_j - \hat{u}_j}{c_s}\right)
$$

where $\{x_i, u_i, i=1,n, j=1,m\}$ 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:

$$
\hat{x}_i = \sum_{i=1}^{n} (\hat{b}_0 + \hat{b}_1 u_{i1} + \hat{b}_2 u_{i2} + \hat{b}_3 u_{i3} + \hat{b}_4 u_{i4} + \hat{b}_5 u_{i5} + \hat{b}_6 u_{i6}),
$$

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 $\hat{B}$ it is used the method based on minimizing the criterion of least squares (least squares method). Mathematical interpretation of criterion looks as follows:

$$
F = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \rightarrow \min_{\hat{B}},
$$
where $x_i$ is object output (apartment price) and $\hat{x}_i$ is output of parametric model (3).

The nonparametric model, considering multidimensional attribute space of the sample, looks as follows:

$$
\hat{x}_2 = \frac{\sum_{i=1}^{n} x_i \prod_{j=1}^{2} \Phi^{(q)}(u_{j} - u_{j_i}) \prod_{j=3}^{6} \Phi^{(s)}(u_{j} - u_{j_i})}{\prod_{j=1}^{n} \prod_{j=3}^{6} \prod_{j=1}^{2} \Phi^{(q)}(u_{j} - u_{j_i}) \prod_{j=3}^{6} \Phi^{(s)}(u_{j} - u_{j_i})},
$$

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

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

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

$$
\Phi^{(s)}(u) = \begin{cases} 
1, & u = u_{j_i}, \\
0, & u \neq u_{j_i},
\end{cases}
$$

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

$$
\delta_i = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - \hat{x}_{j_i}}{x_i} \right| \times 100\%,
$$

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

4. Computational experiment

At the first stage parametric modelling 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 – $\hat{x}_{j_i}, i = 1, n$, that are illustrated below combines with the genuine apartments prices:

![Figure 1](image.png)

**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 $\delta_1 = 13.6\%$ 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 nonparametric modelling of the data by formula (5). The results are presented in the figure below:
[5] Nadaraya E A 1964 On estimating regression Theory of Probability and Its Applications 9 157-9
[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)