УДК [338.5+347.214.2]:519.2

ПРОГНОЗИРОВАНИЕ ЦЕН НА НЕДВИЖИМОСТЬ С ПОМОЩЬЮ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Суровцев А.И. [1], студент гр.353505, Хорошко К.Н. [2], студент гр.353505

Белорусский государственный университет информатики и радиоэлектроники[1]
г. Минск, Республика Беларусь

Примичева З.Н. - канд. физ.-мат. наук

Аннотация. В данной работе рассмотрены основные этапы написания программы для прогнозирования цен на недвижимость методом множественной линейной регрессии, анализ работы модели, применимость в реальной жизни.

Ключевые слова. Множественная линейная регрессия, регрессионная модель, прогнозирование, ошибка прогноза.

Введение

В современном мире прогнозирование стоимости недвижимости имеет важное значение как для частных лиц, так и для государственных органов, поскольку позволяет снизить финансовые риски, связанные с покупкой или продажей недвижимости, предоставляет информацию о возможных перспективах на рынке при разработке стратегий в жилищном строительстве.

Экономико-математическое моделирование используется для анализа и прогнозирования динамики цен на жилье, а также для выявления факторов, влияющих на эту динамику. С этой целью проводится анализ рыночных данных и разрабатываются прогностические модели, которые помогают предсказать будущее состояние рынка. В современном мире, где экономические процессы усложняются из-за глобализации, стандартные методы моделирования могут оказаться недостаточными, требуя создания специализированных моделей. Цены на жилье являются основным индикатором рыночной активности. Формирование цен зависит от множества факторов, включая как качественные характеристики объектов недвижимости, так и общую динамику рынка.

Реализация метода множественной регрессии для прогнозирования цен на недвижимость

Линейной множественной регрессией называется модель, выражающая линейную зависимость среднего значения зависимой переменной y от нескольких независимых переменных $x_1,\ldots,\ x_m$, следующего вида:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m, \qquad (1)$$

где y – зависимая переменная (прогнозируемая цена квартиры), $x_1,\ldots,\ x_m$ – независимые переменные (значения факторов, влияющих на цену квартиры), $b_0,\ b_1,\ldots,\ b_m$ – параметры уравнения множественной регрессии.

Соответствующая регрессионная модель для вычисления фактической цены имеет вид:

$$p = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m + e, \qquad (2)$$

где e − ошибка модели, являющаяся случайной величиной регрессионной зависимости.

Пусть имеется n наблюдений зависимой переменной и соответствующих им значений независимых переменных:

$$x_{i1}, x_{i2}, \ldots x_{im}, y_i,\ i = 1,2, \ldots n,$$

тогда модель множественной линейной регрессии можно представить в виде:

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_mx_{im}, \qquad (3)$$

$$p_i = b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_mx_{im} + e_i. \qquad (4)$$

Обозначим $Y = \begin{bmatrix} y_1 & y_2 & \vdots & y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} & 1 & x_{21} & x_{22} & \ldots & x_{2m} & \vdots & \vdots & \vdots & \ddots & \vdots & 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$, $B = \begin{bmatrix} b_0 & b_1 & \vdots & b_m \end{bmatrix}$, $E = \begin{bmatrix} e_1 & e_2 & \vdots & e_n \end{bmatrix}$, $P = \begin{bmatrix} p_1 & p_2 & \vdots & p_n \end{bmatrix}$,

где Y -мерный вектор-столбец наблюдений зависимой переменной, X матрица размерности $n \times (m+1)$, i-я строка которой представляет i-ое наблюдение вектора значений независимых переменных $x_1, \ldots, x_m$ (значения факторов влияющих на цену квартиры под номером i), единица соответствует переменной при свободном члене b_0, B вектор-столбец размерности $(m+1)$ параметров уравнения множественной регрессии, P вектор-столбец размерности n, который содержит фактические стоимости квартир, E вектор-столбец размерности n, который содержит отклонения значений p_i фактической цены от значений y_i, получаемых по уравнению регрессии. Тогда модель (1) примет вид:

$$Y = XB. \qquad (5)$$

Матрица E вычисляется по следующей формуле:

$$E = P - XB. \qquad (6)$$

Будем искать коэффициенты $b_i, i = 0,1, \ldots, m$, так, чтобы функция потерь была минимальной при этих коэффициентах. Для их нахождения составим сумму квадратов всех отклонений:

$$\sum_{i=1}^{n} e_i^2 = E^T E. \qquad (7)$$

Пусть $Q = E^T E$. Тогда в силу (6) получим:

$$Q = E^T E = (P - XB)^T (P - XB). \qquad (8)$$

Из формулы (8) следует, Q квадратичная функция относительно $b_i, i = 0,1, \ldots, m$. Поскольку в силу (7) Q сумма квадратов чисел e_i^2, то функция Q является неотрицательной.

Найдем такую матрицу коэффициентов B, при которой функция Q принимает локальный минимум. Исходя из формулы (8), вычислим частные производные Q по всем $b_i, i = 0,1, \ldots, m$, матрицы B:

$$\frac{\partial Q}{\partial b_i} = \frac{\partial (P^T P - P^T XB - B^T X^T P + B^T X^T XB)}{\partial b_i}, \qquad i = 0,1, \ldots, m. \qquad (9)$$

Заметим, что

$$\frac{\partial P^T P}{\partial b_i} = 0, \qquad i = 0,1, \ldots, m, \qquad (10)$$

так как $P^T P$ не зависит от b_i.

Рассмотрим теперь

$$\frac{\partial A}{\partial b_i} = \frac{\partial P^T XB}{\partial b_i}, \qquad (11)$$

где

$$A = P^T XB = \begin{bmatrix} p_1 & p_2 & \vdots & p_n \end{bmatrix}^T \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 & b_1 & \vdots & b_m \end{bmatrix}.$$

Тогда

$$A = \sum_{j=1}^{n} p_j b_0 + \sum_{i=1}^{m} \sum_{j=1}^{n} p_j x_{ji} b_i. \qquad (12)$$

Следовательно,

$$\frac{\partial A}{\partial b_0} = \sum_{j=1}^{n} p_j, \qquad \frac{\partial A}{\partial b_i} = \sum_{j=1}^{n} p_j x_{ji}, \qquad i = 1, \dots, m\,. \tag{13}$$

Составим матрицу столбец из элементов $\frac{\partial A}{\partial b_i}$, $i = 0,1, \dots, m$,

$$\left(\sum_{j=1}^{n} p_j \sum_{j=1}^{n} p_j x_{j1} \sum_{j=1}^{n} p_j x_{j2} \vdots \sum_{j=1}^{n} p_j x_{jm} \right).$$

Отсюда и в силу формулы (11):

$$\left(\frac{\partial P^T XB}{\partial b_i}\right) = X^T P. \tag{14}$$

Аналогично рассуждая, находим:

$$\left(\frac{\partial B^T X^T P}{\partial b_i}\right) = X^T P, \tag{15}$$

$$\left(\frac{\partial B^T X^T XB}{\partial b_i}\right) = (X^T X + (X^T X)^T)B = 2X^T XB. \tag{16}$$

Таким образом, подставляя (10), (14), (15), (16) в (9), имеем

$$\left(\frac{\partial Q}{\partial b_i}\right) = -2X^T P + 2(X^T X)B. \tag{17}$$

Для нахождения экстремума функции Q, составим уравнение:

$$\left(\frac{\partial Q}{\partial b_i}\right) = O. \tag{18}$$

Отсюда и в силу (17) получим матричное уравнение:

$$(X^T X)B = X^T P. \tag{19}$$

Пользуясь свойством обратных матриц, умножим левую и правую части равенства (19) на $(X^T X)^{-1}$ слева:

$$B = (X^T X)^{-1} X^T P. \tag{20}$$

Формула (20) задает матрицу B коэффициентов b_i, $i = 0,1, \dots, m$ уравнения множественной линейной регрессии.

Теперь можно вычислить отклонения значений p_i фактической цены от значений y_i, получаемых из уравнения регрессии, которые будем называть значениями ошибки прогноза. Найдем среднюю ошибку аппроксимации

$$W = \frac{1}{n} \times \sum_{i=1}^{n} |p_i - x_i B|. \tag{21}$$

Анализ работы

Для прогнозирования цен на недвижимость выберем 9 основных факторов, влияющих на цену жилых помещений, значения которых являются независимыми переменными в модели множественной линейной регрессии, эти факторы представлены в таблице 1.

Таблица 1 - Характеристика факторов жилой недвижимости в коде программы

№ п/п	Название фактора (обозначение в модели множественной линейной регрессии)	Описание фактора	Единицы измерения фактора
1.	House_type (x_1)	Тип дома	Мультиномиальная переменная (string) : P – панельный (4) K – кирпичный (2) KB – кирпично-блочный (5) M – монолитный (1) SB – силикатно-блочный (3)
2.	Square (x_2)	Площадь квартиры	Числовой (double)
3.	Year (x_3)	Год постройки	Числовой (int)
4.	Number_of_rooms (x_4)	Количество комнат	Числовой (int)
5.	Floor (x_5)	Этаж квартиры	Числовой (int)
6.	Number_of_storeys (x_6)	Этажность дома	Числовой (int)
7.	Remont (x_7)	Наличие ремонта в квартире	Бинарная переменная (bool)
8.	Center (x_8)	Находится ли дом в центре Минска	Бинарная переменная (bool)
9.	Minutes_to_metro (x_9)	Расстояние от метро	Минуты (int)

Для наглядности действия модели посчитаем ошибку прогноза для собранных нами данных и составим график для этих значений, как показано на рисунке 1.

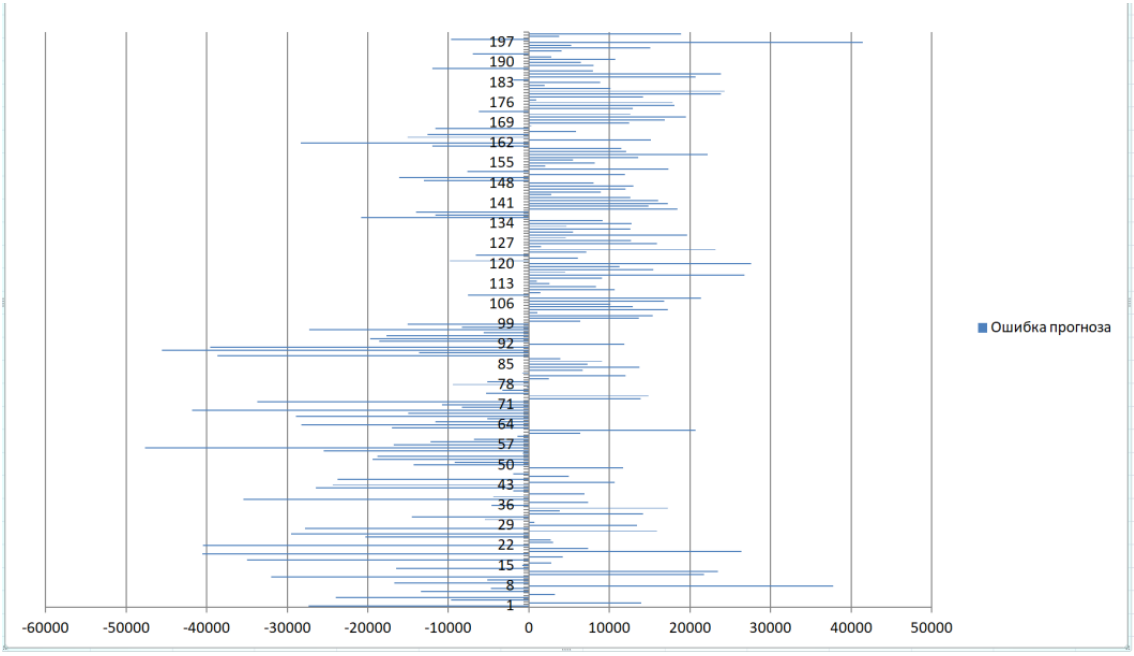

Рисунок 1 – Разность между прогнозируемой ценой и реальной

Значение средней ошибки аппроксимации в нашей модели составляет 13571 долларов.

Для реализации модели множественной регрессии на языке программирования C++ была написана программа, реализующая внутри себя программный код, позволяющий пользователю по заданным характеристикам квартиры, которые представлены в таблице 1, спрогнозировать цену. В программе реализована структура “House”, которая определяет структуру данных для описания домов. Каждый объект “House” содержит информацию о типе дома, количестве комнат, площади, годе постройки, наличии ремонта, этаже, количестве этажей, времени до метро, расположении и цене. Для

работы с матрицами была выбрана библиотека в языке C++ - “Eigen”, реализующая внутри себя различные операции над матрицами. Подсчет матрицы B, коэффициентов множественной линейной регрессии был реализован следующим образом:

```
MatrixXd B = (data_matrix.transpose() * data_matrix).inverse() * data_matrix.transpose() * price_matrix,
```

где data_matrix – матрица, соответствующая матрице X, а price_matrix – соответствует матрице P, описанным выше.

Пользователь может ввести данные квартиры, для которой он хочет спрогнозировать цену. Для подсчета прогнозируемой цены, программа перемножает согласно формуле (5) матрицу X на матрицу коэффициентов B и выводит окончательную цену на квартиру. Также в программе реализована функция для подсчета матрицы отклонений E и функция для подсчета ошибки аппроксимации. В целом, программа предоставляет возможности для прогнозирования цен на квартиры в Минске, для заданных характеристик квартиры и реализует модель множественной линейной регрессии.

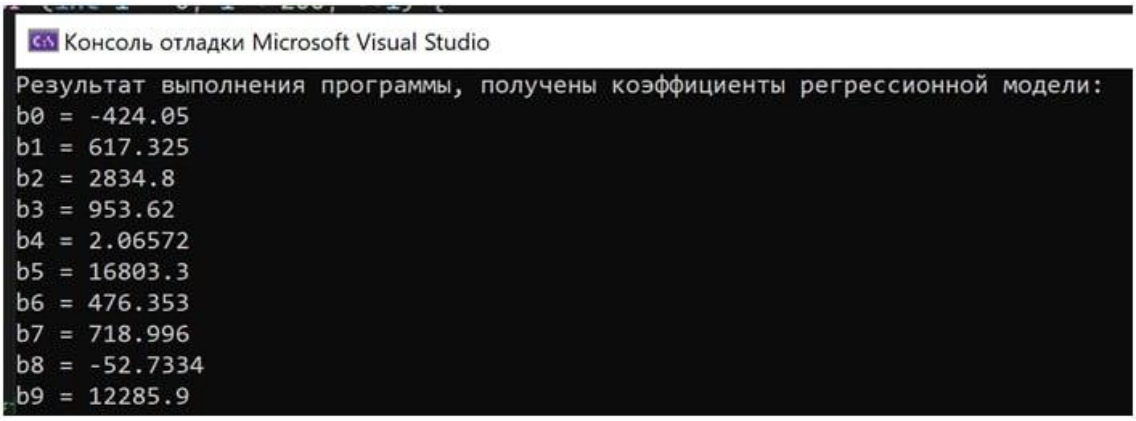

Рисунок 2 – Значения матрицы B

Заключение

Таким образом, модель множественной линейной регрессии для прогнозирования цен на недвижимость с учетом 9 основных факторов показала хорошие результаты, хотя количество используемых данных было всего 200 квартир, а также существовали квартиры, цена которых была сильно завышена или занижена по сравнению с другими квартирами с такими же характеристиками. Стоит отметить, что ошибка аппроксимации достаточно маленькая в отношении стоимости самой квартиры к средней фактической стоимости квартиры для данного метода, которую можно вычислить по формуле:

$$k = \frac{W}{\frac{1}{n} \times \sum_{i=1}^{n} y_i} \times 100\%. \qquad (22)$$

Значение k составило 13.5%. Из полученного результата можно сделать вывод о том, что в среднем вычисленная стоимость квартиры с помощью построенной модели множественной линейной регрессии на основе предоставленных данных будет неверной на ± 13.5%, хотя теоретически могут быть точные совпадения. Анализируя полученный результат, можно сделать вывод о преимуществах и недостатках модели множественной линейной регрессии.

Преимущества метода множественной линейной регрессии:

1. Лёгкость алгоритмизации: метод линейной регрессии легко алгоритмизируется и реализуется в программном коде, что делает его привлекательным для применения в различных прикладных задачах.

2. Интерпретируемость: коэффициенты модели множественной линейной регрессии имеют прямую интерпретацию, что позволяет понять, как каждый фактор влияет на прогнозируемую переменную.

3. Эффективность: при правильном выборе факторов и предварительной обработке данных модель множественной линейной регрессии может давать точные прогнозы и эффективно описывать зависимость между переменными.

Недостатки метода множественной линейной регрессии:

1. Линейность: метод работает только с линейными зависимостями между переменными, что ограничивает его применимость в случае нелинейных взаимосвязей.

2. Чувствительность к выбросам: наличие выбросов или аномальных значений в данных может исказить результаты модели, особенно при использовании наименьших квадратов для оценки коэффициентов.

3. Мультиколлинеарность: проблема мультиколлинеарности (высокой корреляции между факторами) может привести к нестабильности оценок коэффициентов модели.

В целом, несмотря на свои ограничения, метод множественной линейной регрессии является простым и эффективным инструментом для анализа и прогнозирования, особенно в случаях, когда данные легко интерпретируются и предполагается линейная зависимость между переменными.

Список использованных источников:

1.Источник содержащий подробные данные о ценах и изменениях цен на недвижимость в Минске. [Электронный ресурс]. — Режим доступа: https://realt.by/

2. Практическая статистика для специалистов Data Science 50 важнейших понятий / Пер. с англ. / П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил [Электронный ресурс] — Режим доступа: https://batrachos.com/sites/default/files/pictures/Books/Bruce_Bruce_2018_Practical%20Statistics%20for%20Data%20Scientists.pdf

3. Расчет коэффициентов множественной линейной регрессии матричным способом. [Электронный ресурс] — Режим доступа: https://univer-nn.ru/ekonometrika/raschet-koefficientov-mnozhestvennoj-linejnoj-regressii-matrichnym-sposobom/

UDC [338.5+347.214.2]:519.2

FORECASTING REAL ESTATE PRICES USING MULTIPLE LINEAR REGRESSION

Surautsau A.I. [1], *Kharoshka K.N.* [2]

Belarusian State University of Informatics and Radioelectronics[1], *Minsk, Republic of Belarus*

Prymichova Z.N. - PhD, Associate Professor

Annotation. This paper discusses the main stages of writing a program for predicting real estate prices using the multiple linear regression method, analyzing the operation of models, and applicability in real life.

Keywords. Multiple linear regression, regression model, forecasting, forecast error real estate.