УДК 159.93

ОЦЕНКА РЫНОЧНОЙ СТОИМОСТИ КВАРТИР С ПОМОЩЬЮ МЕТОДОВ РЕГРЕССИОННОГО АНАЛИЗА

А.Б. Горобцова

В работе рассматриваются способы формирования стоимости строящейся недвижимости, влияния различных характеристик на цену квартир и построение моделей оценки стоимости.

In the paper ways of forming the value of real estate under construction, the influence of various characteristics on the price of apartments and the construction of models for valuation are considered.

КЛЮЧЕВЫЕ СЛОВА
Линейная регрессия, метод наименьших квадратов, оценка стоимости недвижимости.

ДЛЯ ЦИТАТЫ
А.Б. Горобцова. Оценка рыночной стоимости квартир с помощью методов регрессионного анализа // Моделирование и анализ данных. 2019. №2. С.63-72.

A.B. Gorobtsova. Assessment of the market value of apartment using regression analysis**.** Modelirovaniye i analiz dannykh=Modelling and data analysis *(Russia)*. 2019, no.2, pp.63-72.

1. ВВЕДЕНИЕ

Оценка стоимости недвижимости является актуальной темой, так как рынок недвижимости активно развивается. В Москве около тысячи строящихся и готовых для заселения жилых комплексов. Целью работы является определение реальной стоимости жилья. Большинство публикаций об оценке посвящено рынку вторичного жилья. В данной статье изучается рынок первичного жилья города Москвы, то есть уже готовые новостройки, а также только строящиеся дома.

В ходе исследования предполагается определить характеристики, которые сильнее всего влияют на формирование стоимости недвижимости, построить модель зависимости стоимости от этих характеристик. С помощью метода наименьших квадратов найти оценки неизвестных параметров в линейной модели регрессии. На основе построенной модели провести анализ недооцененных квартир, т.е. квартир, стоимость которых значительно ниже прогнозируемого значения.

2. СБОР ДАННЫХ

Основной задачей данной работы является создание модели стоимости недвижимости. Рассмотрены квартиры в новостройках, готовых и строящихся**.** Полученные данные являются как количественными, так и качественными. Качественные характеристики были переформированы: при наличии признака ставится «1», при отсутствии – «0».

Таблица 1. Характеристики недвижимости.

Параметр	Описание
price	Стоимость жилья, тыс. руб.
mkad	Расстояние до МКАДа, км.
distance_to_metro	Расстояние до ближайшего метро, км.
centre_time	Расстояние от объекта недвижимости до станция метро «Охотный ряд», км.
metro_time	Время поездки от ближайшей к дому станции метро до станции «Охотный ряд», мин.
total_time	Общее время от дома и до станции «Охотный ряд» , мин.
total_space	Общая площадь, кв.м.
living_space	Жилая площадь, кв.м.
kitchen_space	Площадь кухни, кв.м.
Количество комнат number_rooms	Количество комнат, шт.
bathroom	Тип санузла
ecology	Данный параметр является рейтинговым от 1 до 4, где «1» - плохая экология, «4» - очень хорошая.
year	Год сдачи
ipoteka	Возможность ипотеки.
reliability	Надежность застройщика.
shops	Наличие торговых центров.
elevator	Количество лифтов, шт.
number_flats	Количество квартир для продажи, шт.
finishing	Наличие отделки.
class	Класс недвижимости «1» - эконом, «2» - комфорт, «3» - бизнес.
tipe_house	Тип дома, где «1» - монолитный, «0» - панельный.
kindergarten	Наличие детсада
school	Наличие школы
hospital	Наличие поликлиники
balcony	Наличие балкона

В работе для сбора данных был использован сайт *cian.ru*, где для большинства характеристик, представленных в табл. 1, доступна выгрузка в формате Excel. Однако часть необходимых данных приходится выгружать вручную. В итоге была получена таблица с данными более чем трех с половиной тысяч квартир в новостройках.

После формирования такой таблицы выяснилось, что имеются пропущенные значения. Для определения пропущенных характеристик была подобрана информация из других источников, в частности на официальном сайте жилого комплекса. При невозможности получения данных из других источников в случае количественных характеристик значения можно заменить средним, если же характеристика качественна, то данное наблюдение исключается из анализа. Для анализа стоимости квартир в работе построены модели линейной регрессии с помощью метода наименьших квадратов. Задача заключается в нахождении коэффициентов линейной зависимости, при которых функция потерь принимает наименьшее значение, т.е. при данных коэффициентах сумма квадратов отклонений экспериментальных данных от найденной прямой будет наименьшей. Более подробно изучить линейную регрессию и метод наименьших квадратов можно в специальной литературе [1, 2].

3. ОБЗОР ЛИТЕРАТУРЫ

Целью большинства работ, связанных с моделями стоимости недвижимости, является выявление основных факторов формирования цены, таких работ существует немало [3–6]. Помимо параметров, представленных в таблице 1, в работах встречается: степень износа инфраструктуры, доля благоустройства территории, доля промышленных объектов в общей площади территории, доля озеленения территории общего пользования, количество развлекательных элементов и т.д. Но такие характеристики требуют более подробного изучения и их выяснение – трудоемкий процесс. Некоторые характеристики при анализе нашей выборки квартир можно исключить, так как они являются одинаковыми для всех квартир. Такими характеристиками оказались «наличие школы», «наличие детсада» «наличие поликлиники» «наличие торговых центров», «возможность ипотеки».

Влияния наличия школ на формирование цен на жилье были проведены исследования, в том числе, о взаимосвязи стоимости квартир и характеристики близлежащих школ. В работе [7] было рассмотрено вторичное жилье в городе Перми. Для выявления зависимости были проанализированы результаты ЕГЭ. Они показали, что при улучшении балла ЕГЭ на одно стандартное отклонение увеличивается стоимость квартиры на 30 тысяч рублей, а также в работе выявлена отрицательная зависимость между ценой квартиры и уровнем преступности среди учеников. Покупатели, приобретающие жилье с большим количеством комнат, готовы доплатить за школу с более высокими показателями, так как обычно большая квартира покупается при наличии детей.

Также важно учитывать такую переменную как «экология», так как она может существенно влиять на цену квартиры. В работе [8] сделан акцент при оценке стоимости квартир в городе Москве на экологические факторы, а именно влиянии на стоимость содержание в воздухе оксида углерода, оксида азота и двуокиси азота. При использовании линейной модели регрессии наблюдается зависимость: чем меньше концентрация угарного газа и больше расстояние до промышленного предприятия, тем выше цена квартиры. В работе показано, что диоксид азота и оксид азота не оказывают существенное влияние на формирование стоимости квартиры, а расстояние до ближайшего промышленного предприятия и концентрация угарного газа значимо связаны с ценой недвижимости в городе Москве. В таблице 1 экология рассматривается как рейтинговая переменная. «1» — плохая экология, «4» — очень хорошая. Данные были взяты с сайта [9], где указаны различные характеристики округов, а также дана общая оценка по многим показателям каждого района.

В работе [10] рассмотрена модель ценообразования методом географически взвешенной регрессии в городе Саратове на рынке вторичного жилья. Были изучены несколько различных пространственных эконометрических моделей с постоянными и переменными коэффициентами. Для изначальных данных были добавлены также координаты объектов. Эмпирическим методом выявилась зависимость от расположения цены за квадратный метр. Чем ближе к центру — тем дороже. Также было выяснено, что в центре дополнительный метр кухни стоит дороже, чем метр жилой площади, в отличие от окраин, где эти цены примерно равны.

Существуют работы анализирующие влияние транспортной доступности на цену жилья, как в России, так и за рубежом. Было выявлено, что данный показатель играет важную роль при оценке стоимости квартир. В работе [6] сделан акцент не только на близость остановок рядом с домом, но и на число маршрутов. Кроме того в работе [6] отмечено, что существенное влияние на формирование стоимости квартиры оказывает расположение квартиры на первом этаже. Построенная модель показала, что при увеличении площади квартиры на $1м^2$ цена повышается на 1,5%. Показано, что расстояние до остановки не является суще-

ственным фактором при оценке стоимости, а вот появление каждого нового маршрута общественного транспорта приводит к увеличение цены. Также на ее формирование влияют время, за которое можно добраться до центра города, уровень загрязненности воздуха, наличие дошкольных образовательных учреждений и торговых центров.

4. АНАЛИЗ ДАННЫХ

На рис. 1 представлено распределение цены в зависимости от количества комнат, по графику можно заметить, что разброс цен достаточно большой, особенно у трехкомнатных квартир.

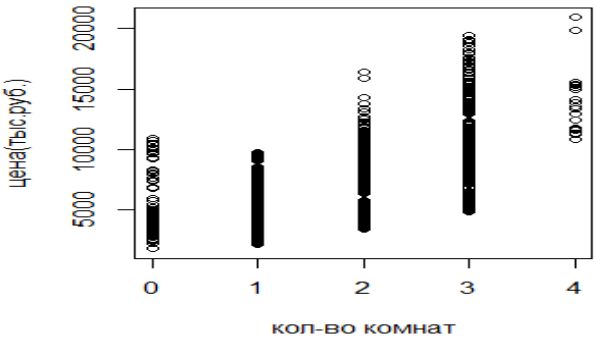

Рис. 1

На рис. 2 можно увидеть, что на рынке в основном представлены квартиры стоимостью от 4 до 10 миллионов руб.

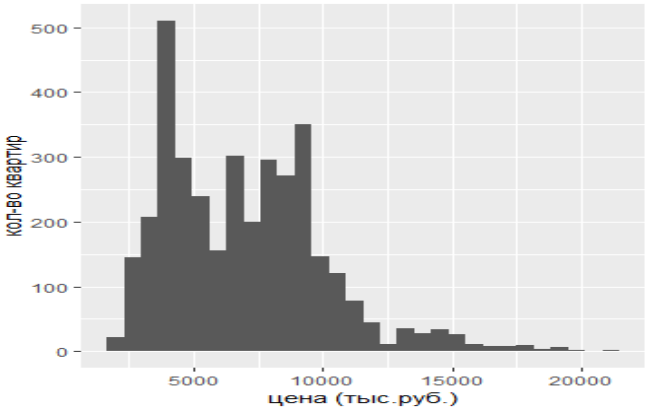

Рис. 2

Расчеты в работе проведены в программной среде R. R — язык программирования и среда для статистической обработки данных и работы с графикой. R широко используется как статистическое программное обеспечение для анализа данных и в последнее время стал стандартом для статистических программ.

Корреляционный анализ показал, что часть характеристик является сильно зависимыми, в особенности жилая и общая площади квартиры. Это вполне ожидаемо, поэтому было принято решение при построении регрессии оставить только общую площадь. Площадь кухни слабо связана с другими площадями, что не удивительно: сейчас существует очень много планировок, зачастую они индивидуальны, также довольно распространенное явление – студии, где отдельное пространство для кухни и вовсе не предусмотрено.

Таблица 2. Корреляционный анализ величин площадей

	total_space	living_space	kitchen_space
total_space	1,0	0,7578	0,4004
living_space	0,7578	1,0	0,0803
kitchen_space	0,4004	0,0803	1,0

Аналогичный анализ для характеристик расстояния и времени позволили исключить из рассмотрения все переменные кроме «время на метро» и «общее время».

Далее модель стоимости квартиры строится отдельно по числу комнат, так как каждая квартира относительно этого параметра будет иметь свои особенности. Первыми были рассмотрены однокомнатные квартиры. Однокомнатные квартиры преимущественно продаются с отделкой и совмещенным санузлом в монолитных домах комфорт-класса. В таблице 3 представлены описательные статистики количественных переменных: выборочное среднее, среднее квадратическое отклонение, максимальное и минимальное значение, медиана.

Таблица 3. Описательные статистики.

пар-р./хар-ка	**mean**	**sd**	**median**	**min**	**max**
metro_time	28,36	3,98	30	20	39
total_time	63,22	24,74	66	25	96
total_space	36,61	4,44	36,6	18	54,93
kitchen_space	12,94	3,9	12,5	2,1	22,4
floor	9,62	6,65	9	1	37
floors	16,91	7,64	17	5	39
class	1,77	0,66	2	1	3
ecology	1,66	0,65	2	1	3
number_flats	460,58	260,72	247	6	1166
price	5160,2	1857,8	4551	2126	9747
year	2019,5	1,11	2020	2018	2022
elevator	2,03	0,56	2	1	6

Рассмотрим линейную модель регрессии:

$$price_i = k_0 + k_1 * metro\_time_i + k_2 * total\_time_i + k_3 * total\_space_i + k_4 * kitchen\_space_i + k_5 * floor_i + k_6 * floors_i + k_7 * tipe\_house_i + k_8 * finishing_i + k_9 * class_i + k_{10} * reability_i + k_{11} * ecology_i + k_{12} * number\_flats_i + k_{13} * balcony_i + k_{14} * batchroom_i + k_{15} * year_i + k_{16} * elevator_i + \varepsilon_i$$

$$i = 1, \dots, N,$$

где N – число квартир, ε_i — независимые и одинаково распределённые случайные величины.

С помощью метода наименьших квадратов были найдены оценки неизвестных параметров $k_0, \dots, k_{16}$, которые представлены во втором столбце таблицы 4. В третьем столбце представлены средние квадратические отклонения соответствующих оценок. В четвертом столбце приведено p-value для критерия Фишера [7] при проверке гипотезы $H_0: k_j = 0$, $j = 1, \dots, 16$. Если значение p-value больше или равно заданного уровня значимости, то эта нулевая гипотеза принимается, иначе – отвергается. Система R предлагает удобный визуальный способ отображения данных: в последнем столбце отмечены те оценки, для которых гипотеза отвергается на уровне значимости 0.01 (***), 0.05 (**) и 0.1 (*). Пустая ячейка говорит о том, что гипотеза принимается для всех трёх уровней значимости и соответствующую характеристику возможно следует исключить из модели.

Таблица 4. Результаты оценивания параметров модели 1

	Estimate	**Std. error**	**Pr(>|t|)**	
(Intercept)	-2,44E+05	67850	0,000343	***
metro_time	1,24E+02	54,6	0,023326	*
total_time	-3,64E+01	11,89	0,002269	**
total_space	1,13E+02	5,9	<2e-16	***
kitchen_space	1,11E+01	7,891	0,159392	
floor	3,01E+01	2,988	<2e-16	***
floors	1,50E+01	3,237	4,24E-06	***
tipe_house	-2,08E+02	341,8	1,51E-09	***
finishing	-5,29E+02	455,3	0,245828	
class	2,15E+03	3,46E-02	7,15E-10	***
reliability	-2,84E+02	2226	0,898389	
ecology	-2,39E+00	528,6	0,9964	
number_flats	-5,35E-01	2,717	0,843887	
balcony	5,32E+01	64,88	0,412299	

bathroom	-8,40E+01	116,5	0,471277	
year	1,22E+02	33,32	0,000273	***
elevator	-3,54E+02	47,37	1,60E-13	***
AIC	18475,5		**Adjusted R-squared**	0.9179

Скорректированный коэффициент детерминации $R^2 = 0{,}9179$ (чем ближе R^2 к 1, тем лучше построенная нами модель) оказался высоким. Также было вычислено значение информационного критерия Акаике (AIC) [11], который используется при выборе одной из нескольких моделей регрессии. Меньшее значение этого критерия говорит о том, что данная модель лучше других.

После исключения статистически незначимых характеристик была построена новая модель регрессии с меньшим количеством переменных. Исключение переменных проводилось по одному и в зависимости от поведения скорректированного коэффициента детерминации были приняты решения о том, исключать переменную или нет. В итоге была получена следующая модель.

Таблица 5. Результаты оценивания параметров модели 2

	Estimate	**Std. error**	**Pr(>|t|)**	
(Intercept)	-2,37E+05	6,52E+04	0,000282	***
metro_time	7,89E+01	1,11E+01	1,80E-12	***
total_time	-2,60E+01	1,31E+00	<2e-16	***
total_space	1,19E+02	4,13E+00	<2e-16	***
floor	3,06E+01	2,96E+00	<2e-16	***
floors	1,48E+01	3,08E+00	1,82E+06	***
tipe_house	-1,98E+03	8,43E+01	<2e-16	***
class	1,18E+03	7,97E+02	<2e-16	***
reliability	-1,92E+03	1,92E+01	<2e-16	***
ecology	-4,24E+02	5,20E+01	8,08E-16	***
number_flats	-2,64E+00	1,52E-01	<2e-16	**
year	1,19E+02	3,23E+01	0,00023	***
elevator	-3,41E+02	3,47E+01	<2e-16	***
AIC	18469,8		**Adjusted R-squared**	0,9179

В новой модели коэффициент детерминации не изменился, а критерий Акаике оказался меньше, что указывает на то, что выбор нужно сделать в пользу второй модели, тем более что в ней меньше регрессоров, что делает данную модель проще для анализа. По модели 2 определены прогнозные цены и получен следующий график, изображенный на рис. 3.

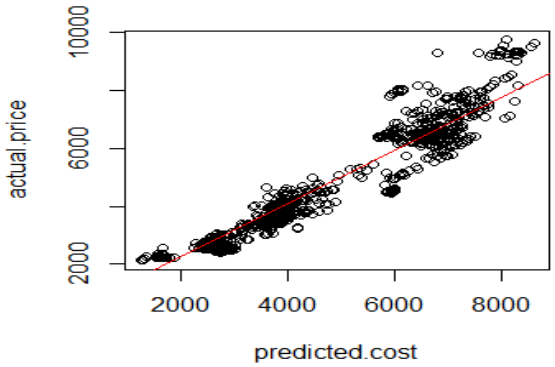

Рис.3

Квартиры, которые находятся выше линии регрессии, переоценены, то есть предсказанная цена оказалась ниже рыночной, квартиры под линией регрессии соответственно недооценены. Для покупателей наибольший интерес представляют вторые. В таблице 6 приведены квартиры, предсказанные цены которых наиболее значительно отличаются от рыночной цены (отношение предсказанной и рыночной цен).

Таблица 6. Недооцененные квартиры для модели 2

Номер наблюдения	786	777	789	787	795
Отношение цен	1,338	1,337	1,335	1,327	1,322

Так, квартира 786 недооценена более чем на 33%. Такая разница между ценами может возникать по разным причинам, как под воздействием человеческого фактора, то есть ошибки или продавец может быть заинтересован в быстрой продажи квартиры, так и от того, что есть еще какие-то характеристики, которые не учтены в данной модели, но влияют на цену этой квартиры. Для выяснения причин такое разницы в ценах требуется более подробно изучать описание квартиры, отзывы, посмотреть квартиру и т.п.

Была также рассмотрена модель 3, где были взяты логарифмы от цен на квартиру, что достаточно часто позволяет построить более точную модель линейной регрессии при анализе цен на недвижимость.

Таблица 7. Результаты оценивания параметров модели 3

	Estimate	**Std. error**	**Pr(>|t|)**	
(Intercept)	2,77E+01	1,00E+01	0,00601	**
metro_time	9,56E-03	1,75E-03	5,88E-08	***
total_time	-6,49E-03	2,53E-04	<2e-16	***
total_space	2,52E-02	6,39E-04	<2e-16	***
floor	4,19E-03	4,59E-04	<2e-16	***
floors	4,43E-03	4,75E-04	<2e-16	***
tipe_house	-3,84E-01	1,31E-02	<2e-16	***

class	1,91E-01	2,24E-02	<2e-16	***
reliability	-4,83E-01	2,95E-02	<2e-16	***
ecology	-7,50E-02	8,16E-03	<2e-16	***
number_flats	-5,71E-04	2,39E-05	<2e-16	***
balcony	2,11E-02	9,90E-03	0,0331	*
year	-9,46E-03	4,97E-03	0,05735	
elevator	-7,63E-02	5,36E-03	<2e-16	***
AIC	-2508		**Adjusted R-squared**	0,9524

Скорректированный коэффициент детерминации в этом случае был выше, чем в предыдущих двух моделях. Данная модель похожа на вторую, только здесь можно исключить еще и характеристику «наличие балкона».

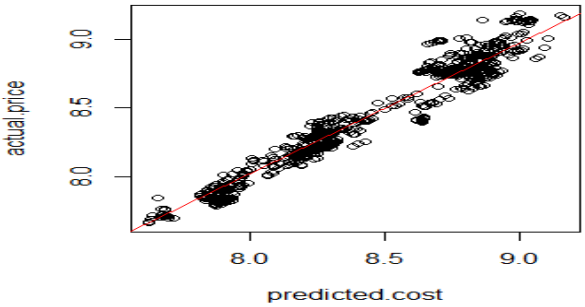

Рис. 4

График получился похожим на тот, который был построен по второй модели.

Таблица 8. Недооцененные квартиры для модели 3

Номер наблюдения	786	777	789	787	795
Отношение цен	1,271	1,270	1,268	1,254	1,254

Видно, что номера самых недооцененных квартир остались теми же, что и для модели 2.

Аналогично были исследованы двухкомнатные и трехкомнатные квартиры; модели оказались схожими и результаты недооцененности этих квартир тоже доходили до 30%.

Однако для четырехкомнатных квартир ситуация оказалась иной. Недооцененных четырехкомнатных квартир практически нет. Конечная модель стоимости четырехкомнатных квартир выглядит следующим образом:

$$price = k_0 + k_1 * metro\_time + k_2 * total\_time + k_3 * total\_space + k_4 * kitchen\_space + k_5 * floor + k_6 * floors + k_7 * tipe\_house$$

Цены, предсказанная и рыночная, почти совпадают: самая большая разница — 3%. Отдельно были рассмотрены квартиры-студии, так как отдельного пространства для кухни в них не

предусмотрено, соответственно переменная «площадь кухни» равна нулю, но результаты оказались схожими с однокомнатными квартирами.

5. ЗАКЛЮЧЕНИЕ

В ходе работы были рассмотрены более трех тысяч квартир-новостроек в Москве и Подмосковье. В зависимости от количества комнат были построены линейные модели регрессии. Выяснилось, что среди всех выборок, кроме четырехкомнатных квартир, имеются сильно переоцененные и недооцененные квартиры. Отличие рыночной цены от предсказанной может доходить до 30%. При построении моделей исключить большое количество показателей не удалось, модели все равно имеют достаточное большое число регрессоров, это говорит о том, что, вероятно, каждый из этих регрессоров оказывает существенно влияние на формирование цены.

Предложенные модели позволяют выделить из общей масс квартир наиболее интересные кандидаты для более подробного рассмотрения. Например, определить для себя наиболее важные параметры, а потом, относительно них, рассматривать недооцененные квартиры. Данный метод также может быть использован при оценке стоимости недвижимости, например, когда покупатель брал ипотечный кредит, после получения акта приема-передачи объекта недвижимости. Со стороны продавца могут быть рассмотрены переоцененные квартиры, на которые следует обратить внимание в случае, если на них нет покупательского спроса.

ЛИТЕРАТУРА

1 Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.
2 Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981.
3 Вязова Г.А., Попелюк В.С. Прогнозирование стоимости двухкомнатной квартиры на вторичном рынке недвижимости в г. Хабаровска с использованием модели множественной регрессии // Молодой ученый. 2011. №25. С. 87–88.
4 Хлюпина М.А., Исавнин А.Г. Моделирование зависимости и анализ цен на квартиры от ряда факторов на примере города Елабуга // Фундаментальные исследования. 2011. №5. С. 213–217.
5 Березина А. В. Эконометрическая модель стоимости вторичного жилья на примере г. Челябинска// Современные научные исследования и инновации. 2015. №7.
6 Сидоровых А.С. Оценка влияния транспортной доступности на цены недвижимости // Прикладная эконометрика. 2015. №37. С. 43–56.
7 Ожегов Е.М., Косолапов Н.А., Позолотина Ю.А. О взаимосвязи между стоимостью жилья и характеристиками близлежащих школ // Прикладная эконометрика. 2017. №47. С. 28–48.
8 Катышев П.К., Хакимова Ю.А. Экологические факторы и ценообразование на рынке недвижимости (на примере г. Москвы) // Прикладная эконометрика. 2012. №28. С. 113–123.
9 https://www.novostroy-m.ru/analitika/ekologicheskiy_reyting_rayonov_moskvy
10 Балаш В.А., Балаш О.С., Харламов А.В. Экономический анализ геокодированных данных о ценах на жилую недвижимость // Прикладная эконометрика. 2011. №22. С. 62–77.
11 Akaike, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974. P. 716–723 .

Работа поступила 20.02.2019г.