Национальный исследовательский университет

“Высшая школа экономики”

Факультет экономических наук

Направление: “Экономика”

Домашняя работа по курсу «Эконометрика – 2»

Изучение рынка вторичной недвижимости города Москва

Работу выполнили

Золотухина Евгения

Рыбин Сергей

Агамалов Юрий

Москва 2024

Введение:

Рынок недвижимости является неотъемлемой частью как экономической деятельности, так и социального благополучия населения в современном мире. Актуальность данного высказывания подтверждается большим количеством исследований, в которых под разным углом рассматриваются факторы, влияющие на ценообразование недвижимости (причем как частных домов, так и квартир) в разных городах по всему миру. Примером таких работ служат Pardoe(2008) [1], в которой рассматривается цены на частные дома в одном из городов штата Орегон (США) по данным за 2005 год, или Vânia Ceccato and Mats Wilhelmsson [2], в которой авторы изучают наличие влияния как уровня преступности в целом, так и различных видов преступлений, в окрестностях на цены на квартиры в Стокгольме (Швеции). Данные работы подтверждают высказанное утверждение о том, что рынок недвижимости и факторы, влияющие на него, действительно волнуют людей по всему миру.

Основная цель данного исследования более конкретна - изучение факторов, влияющих на ценообразование квартир вторичного рынка жилья в городе Москва на 14.04.2024.

Гипотезы:

Гипотеза 1: Чем ближе квартира к центру города, тем она дороже

Обоснование: По данным литературы Ndegwa James(2018) [3] и Bui Toan (2020) [4], квартиры, расположенные ближе к центру города, могут иметь преимущества, такие как легкий доступ к инфраструктуре, транспорту, культурным и развлекательным мероприятиям, что делает их более привлекательными для покупателей и повышает их стоимость. Такие квартиры часто ценятся за удобство расположения и потенциал для инвестиций, что может быть отражено в их цене на рынке недвижимости.

Гипотеза 2: Квартиры, расположенные на более высоких/последних этажах домов, имеют более высокую цену по сравнению с квартирами на первых этажах.

Обоснование: На основе предыдущих исследований [5] и рыночных данных можно предположить, что квартиры на последних этажах оцениваются выше из-за лучших видов, меньшего уровня шума от улицы. В то же время,

квартиры на первом этаже могут стоить дешевле из-за повышенных рисков влажности, доступности для взломщиков и отсутствия видов.

Описание данных:

Данные были собраны авторами работы 14.04.2024 с сайта “Циан” [6] - интернет-сервиса для размещения объявлений о недвижимости. Для сбора данных с сайта был использован парсер с гитхаба [7].

В первую очередь данные были очищены от переменных по типу “Контактный номер телефона”. После этого этапа были оставлены следующие переменные:

floor - этаж, на котором находится квартира

floors_count - общее количество этажей в доме

rooms_count - количество комнат в квартире

total_meters - общая площадь квартиры, м²

price - общая стоимость квартиры, рубли

year_of_construction - год постройки дома

kitchen_meters - площадь кухни, м²

underground - ближайшая станция метро

Затем данные были очищены от пропусков путем исключения их из датасета. После чего была введена бинарная переменная “**is_CAO**”, которая отражает расположена ли квартира в центре Москвы (где 1 - квартиры, расположенные на кольцевой линии метро и внутри неё, 0 - остальные). В самом конце предварительной работы с данными была введена бинарная переменная “**house_category”** - категория дома в зависимости от года постройки дома (где 1-монолитный дом, 0 - остальные).

Таким образом, после предварительной обработки было выделено 9 независимых переменных (**floor, floors_count, rooms_count, total_meters, price, year_of_construction, kitchen_meters, is_CAO, house_category**) и 554 наблюдения.

Предполагаемые методы:

В рамках данного исследования основной целью является оценка факторов, влияющих на цены на вторичные квартиры в Москве. Будет использована линейная модель множественной регрессии, где зависимая переменная - цена квартиры, а независимые переменные включают: общую площадь квартиры, общую этажность дома, этаж, на котором находится квартира, количество комнат, площадь кухни, категория дома (монолитный или иной), а также расположение квартиры (находится квартира в ЦАО или нет).

Основной проблемой в данной модели может быть эндогенность, вызванная пропуском переменной, а именно качества дома. Пропущенная переменная скоррелирована с категорией дома.

Таким образом, для решения проблемы эндогенности была подобрана инструментальная переменная для категории дома - год постройки дома. Предполагается, что год постройки дома коррелирует с категорией дома, то есть инструмент релевантен (разные исторические периоды характеризуются преобладанием определенных строительных технологий и стандартов), но не влияет напрямую на цену квартиры (не скоррелирован с ошибкой, то есть инструмент - валидный). В качестве способа оценки регрессии для борьбы с эндогенностью планируем использовать оценку инструментальных переменных (IV) для множественной регрессии.

Графики и дескриптивные статистики:

В Табл.1 приведены основные описательные статистики: минимум, максимум, медиана, среднее значение, стандартное отклонение для переменных модели. Можно сделать вывод о том, что данные адекватны: в них нет ошибок, контринтуитивных значений и аномалий.

Круговые диаграммы (Рис.1) отображают, что бинарная переменная, характеризующая категорию дома достаточно сбалансирована, в то время как в переменной, характеризующей расположение квартиры, наблюдается преобладание квартир, находящихся не в ЦАО.

В нашем случае, из-за присутствия больших значений некоторых признаков, гистограммы являются недостаточно информативными для визуального определения нормальности распределения, но так как выборка

достаточно большая, то считаем, что распределение факторов асимптотически стремится к нормальному. (Рис.2)

Для того чтобы выявить выбросы в данных, рассмотрим диаграммы типа «ящик с усами» для регрессоров в модели. (Рис.3) По ящикам с усами видим, что выбросы остаются, но их количество небольшое и значения реальные, поэтому оставляем в датасете для дальнейшего анализа.

Анализируя корреляционную матрицу и тепловую карту корреляции (Рис.4), делаем вывод, что среди наших переменных не наблюдается сильной корреляции с ценой (сильная статистическая зависимость, если коэффициент корреляции $r > 0,7$), поэтому все параметры будут в дальнейшем включены в модель регрессии. При этом все регрессоры, кроме категории дома, имеют положительную корреляцию с ценой.

Тесты, эндогенность и модель:

Прежде, чем приступить к дальнейшей оценке модели множественной регрессии, мы провели тест Голдфелда-Квандта на нашей первоначальной модели, чтобы убедиться в отсутствии гетероскедастичности. По результатам теста p-value получилось 0.999, что больше уровня значимости 0.05, значит, нулевая гипотеза не отвергается при любом разумном уровне значимости, следовательно, тест Голдфелда-Квандта не выявил гетероскедастичность. Такой же вывод получаем, посмотрев на график остатков (Рис.5). Гомоскедастичность предполагает, что дисперсия остатков не изменяется для различных наблюдений.

Также на основе данного графика мы можем сделать вывод об отсутствии автокорреляции и нормальности распределения остатков регрессионной модели, поскольку среднее значение остатков равно 0. Кроме того, гистограмма распределения остатков довольно хорошо напоминает нормальное распределение (Рис.6).

Перейдем к проблеме эндогенности. Ранее мы предположили, что в нашей регрессионной модели пропущена переменная «качество дома». С точки зрения модели пропуск признака «качество дома» приводит к следующему: эта переменная попадает в ошибки регрессии, что делает оценки коэффициентов смещенными и несостоятельными. Несомненно, «качество дома» коррелирует с ценой на квартиру, так как потенциальные

покупатели готовы заплатить больше за квартиру в качественном доме с современными удобствами, который находится в хорошем состоянии. Таким образом, получается, что переменная «категория дома», включенная в модель, коррелирует с ошибками регрессии, что влечет за собой искажение ошибок, а, значит, и зависимой переменной, что приводит к смещению оценки коэффициента при переменной «категория дома».

Решить проблему эндогенности можно, подобрав инструментальную переменную к переменной «категория дома». Инструмент должен быть скоррелирован с типом дома и не скоррелирован с ошибкой (то есть с ценой квартиры напрямую), исходя из этого, в качестве инструмента для переменной «качество дома» можно взять переменную «год постройки дома».

Теперь оценим качество подобранного инструмента. Ранее мы установили, что корреляция между ценой квартиры и годом постройки дома равна 0.18, это говорит об очень слабой статистической зависимости, следовательно, подобранный инструмент валидный. Проверим, релевантность инструмента. В начале оценим уравнение регрессии, в котором экзогенные переменные и инструмент предсказывают эндогенную переменную, затем вычислим расчетное значение тестовой F-статистики для проверки следующей гипотезы: H0: $\hat{\beta}_{\text{год постройки дома}} = 0$, если эта гипотеза отвергается, то это значит, что инструмент (год постройки дома) вносит существенный вклад в объяснение изменений эндогенной переменной. Это говорит о релевантности инструмента. Обычно если расчетное значение тестовой F-статистики для проверки нулевой гипотезы больше 10, то инструменты признаются релевантными. F-статистика для теста на незначимость инструмента получилась равна 19.18, что больше 10. Можно заключить, что инструмент является релевантным.

Для борьбы с эндогенностью использовался двухшаговый метод наименьших квадратов (2SLS). На первом шаге матрица X, содержащая эндогенную информацию, проецировалась на Z. Z — это матрица без эндогенной информации, которая включает подобранный нами инструмент. На втором шаге мы оценили параметры той модели, которая нас интересовала изначально. Только теперь в правой части вместо эндогенных регрессоров ставятся их предсказанные значения из регрессий первого

шага, то есть $Y = \hat{X}\beta + \epsilon$. Корреляция между потенциально эндогенными переменными и остатками финальной модели приблизительно равна 0.

Correlation with Residuals	
floor	0
floors_count	0
rooms_count	0
total_meters	0
kitchen_meters	0
is_CAO	0
house_category	-0.07

Кроме того, на основе графиков остатков (Рис.7) мы можем сделать вывод об отсутствии автокорреляции и гетероскедастичности, а также о нормальности распределения остатков регрессионной модели (Рис.8), поскольку среднее значение остатков равно 0.

Чтобы понять, действительно ли нужно применять оценки, полученные методом 2SLS, был проведен тест Хаусмана. Нулевая гипотеза заключается в том, что факторы модели экзогенны (матрица возможных эндогенных переменных и ошибки не скоррелированы), альтернативная — что эндогенны (матрица возможных эндогенных переменных и ошибки скоррелированы). В начале мы получили прогнозируемые остатки регрессии с инструментальной переменной ($\hat{r}$), затем оценили регрессию: $Y = X\beta_1 + \hat{r}\beta_2 + \epsilon$ и, наконец, проверили, значительно ли коэффициент при $\hat{r}$ отличается от 0, используя F-тест с 1 степенью свободы. По результатам теста Хаусмана получили: $F_{\text{расчетное}} = 0.097$, p-value = 0.755. P-value > 0.05, значит, нулевая гипотеза не отвергается при любом разумном уровне значимости. Следовательно, матрица возможных эндогенных переменных и ошибки не скоррелированы. Таким образом, $\widehat{\beta_{OLS}}$ - эффективная и состоятельная, $\widehat{\beta_{2SLS}}$ – состоятельная, то есть разница между оценками мала. На основе теста Хаусмана нет доказательства того, что переменная "категория дома" эндогенная, поэтому делаем выбор в пользу модели, оцененной при помощи метода наименьших квадратов (Табл.2).

Для начала проверим гипотезу о незначимости регрессии в целом. Тест Фишера позволяет проверить незначимость регрессии в целом, то есть установить, равны ли коэффициенты одновременно при всех регрессорах нулю. Если коэффициенты признаются равными нулю (H0), регрессия считается незначимой, если коэффициент хотя бы при одном регрессоре

отличен от нуля, регрессия значима(H1). Расчетное значение F-статистики для этой модели регрессии равно 94.17, p-value равно 0, что меньше уровня значимости 0.05, H0 отвергается, значит, регрессия значима и адекватна.

По результатам регрессии следующие переменные оказались значимы – **floors_count, total_meters, is_CAO, house_category**. (Табл.2)

Переменная **floors_count** положительно влияет на зависимую переменную, то есть чем больше этажей в здании, тем она дороже. Можно предположить, что это связанно с тем, что современные дома строят более высокими.

Выводы:

В результате исследования было выявлено, что коэффициент перед переменной is_CAO является значимым и положительным. То есть квартиры, расположенные на кольцевой линии метро и внутри неё стоят дороже, чем остальные. Значит, гипотеза “Чем ближе квартира к центру города, тем она дороже” подтвердилась. Можно предположить наличие данной зависимости из-за того, что квартиры, которые находятся вблизи центра города, могут быть востребованы из-за легкого доступа к инфраструктуре, транспорту, культурным и развлекательным объектам.

Гипотеза “Квартиры, расположенные на более высоких/последних этажах домов, имеют более высокую цену по сравнению с квартирами на первых этажах” не подтвердилась, поскольку коэффициент перед переменной “floor” оказался незначим.

Литература и источники:

1. Pardoe, I. (2008). Modeling Home Prices Using Realtor Data. Journal of Statistics Education, 16(2). https://doi.org/10.1080/10691898.2008.11889569

2. Ceccato, V. and Wilhelmsson, M. (2011) ‘The impact of crime on apartment prices: evidence from Stockholm, Sweden’, Geografiska Annaler: Series B, Human Geography, 93(1), pp. 81–103

3. Ndegwa, James. (2018). Determinants of Apartment Prices within Housing Estates of Nairobi Metropolitan Area. International Journal of Economics and Finance. 10. 104.

4. Bui, Toan. (2020). A study of factors influencing the price of apartments: Evidence from Vietnam. Management Science Letters. 10. 2287-2292.

5. Хлюпина М.А., Исавнин А.Г. МОДЕЛИРОВАНИЕ ЗАВИСИМОСТИ И АНАЛИЗ ЦЕН НА КВАРТИРЫ ОТ РЯДА ФАКТОРОВ НА ПРИМЕРЕ ГОРОДА ЕЛАБУГА // Фундаментальные исследования. – 2016. – № 5-1. – С. 213-217; URL: https://fundamental-research.ru/ru/article/view?id=40278

6. https://www.cian.ru/

7. https://github.com/lenarsaitov/cianparser

Приложение:

	floor	floors_count	rooms_count	total_meters	price	year_of_constuction	kitchen_meters	is_CAO	house_category
count	554	554	554	554	554	554	554	554	554
mean	7,84	15,28	1,49	45,33	15801105	1993,84	9,96	0,03	0,36
std	5,96	7,94	0,5	12,07	9289276,8	24,35	3,81	0,18	0,48
min	1	3	1	15	5000000	1895	3	0	0
25%	3	9	1	37,2	10800000	1971	7,5	0	0
50%	6	14	1	42	13300000	1997,5	9,45	0	0
75%	11	18	2	53	17500000	2018	11	0	1
max	35	85	2	125	100000000	2024	30	1	1

Табл.1 Основные описательные статистики

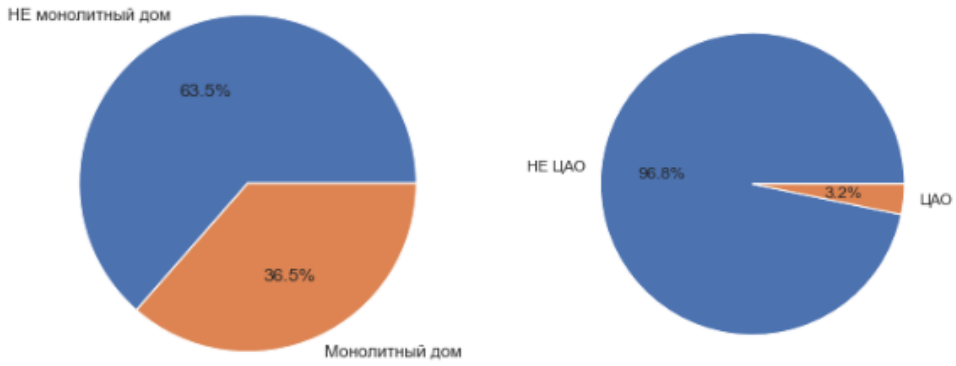

Рис.1 Круговые диаграммы распределения наблюдений по признакам расположение дома и тип дома

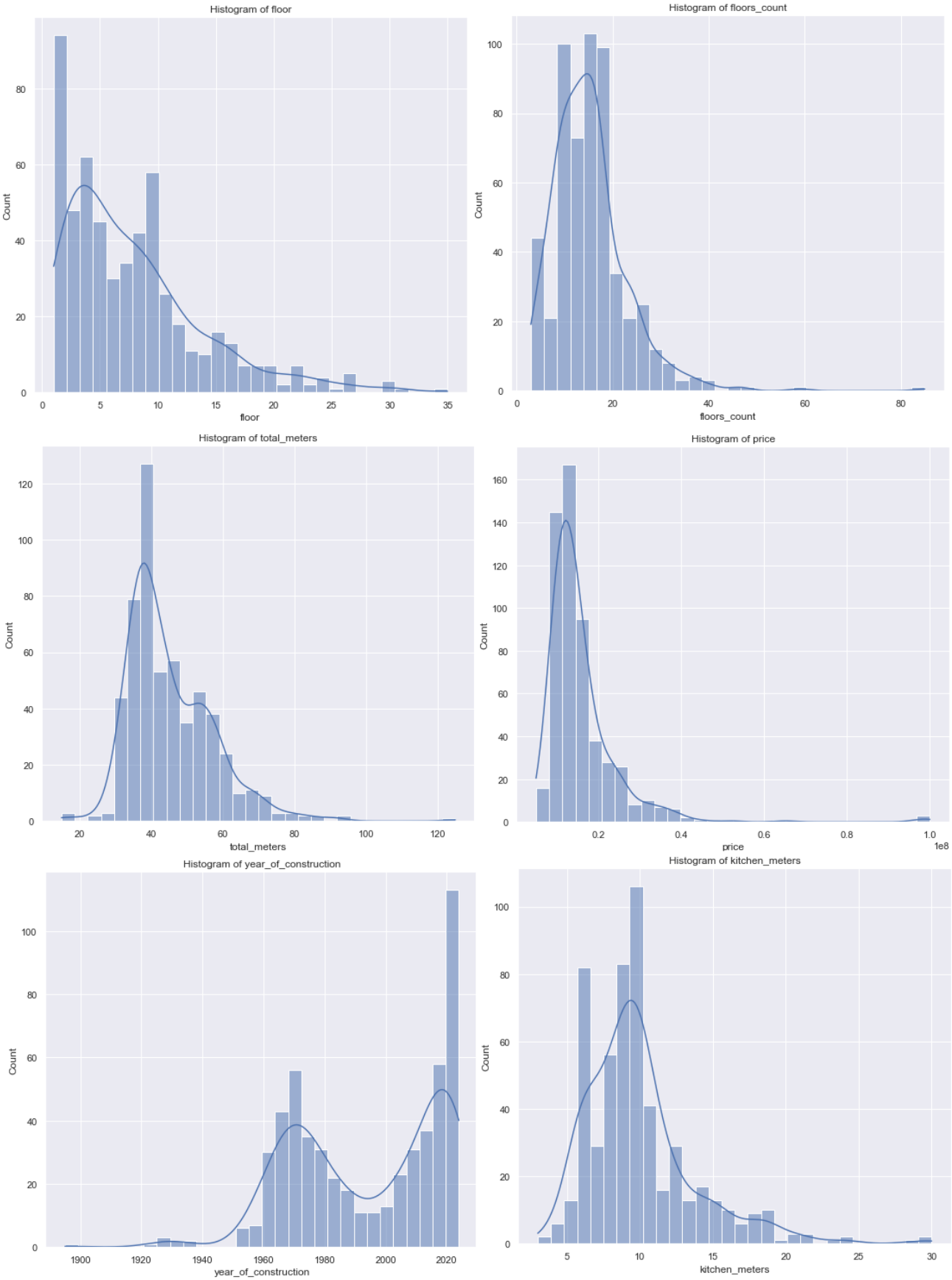

Рис.2 Гистограммы распределения квартир по различным переменным

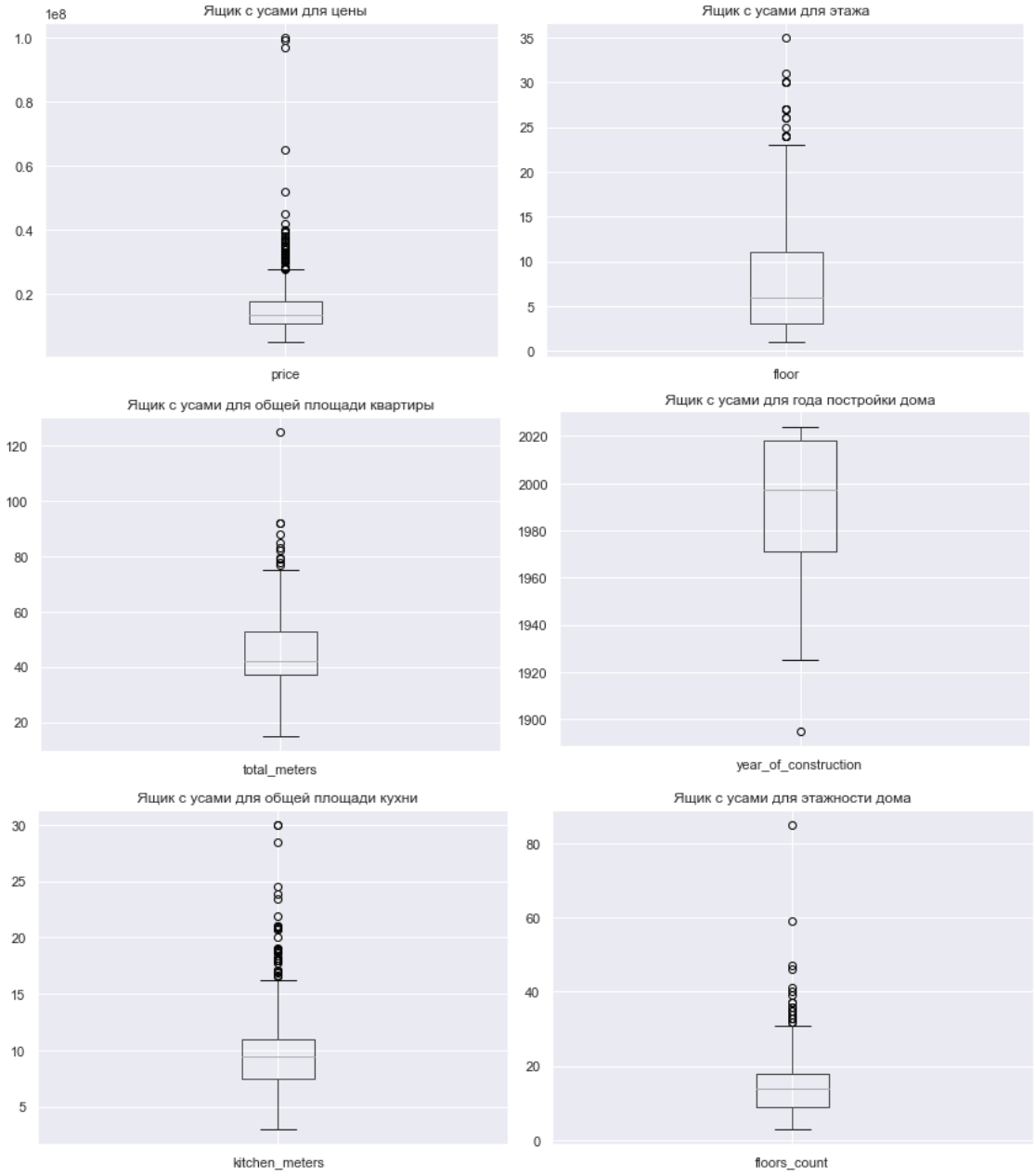

Рис.3 Ящичковые диаграммы

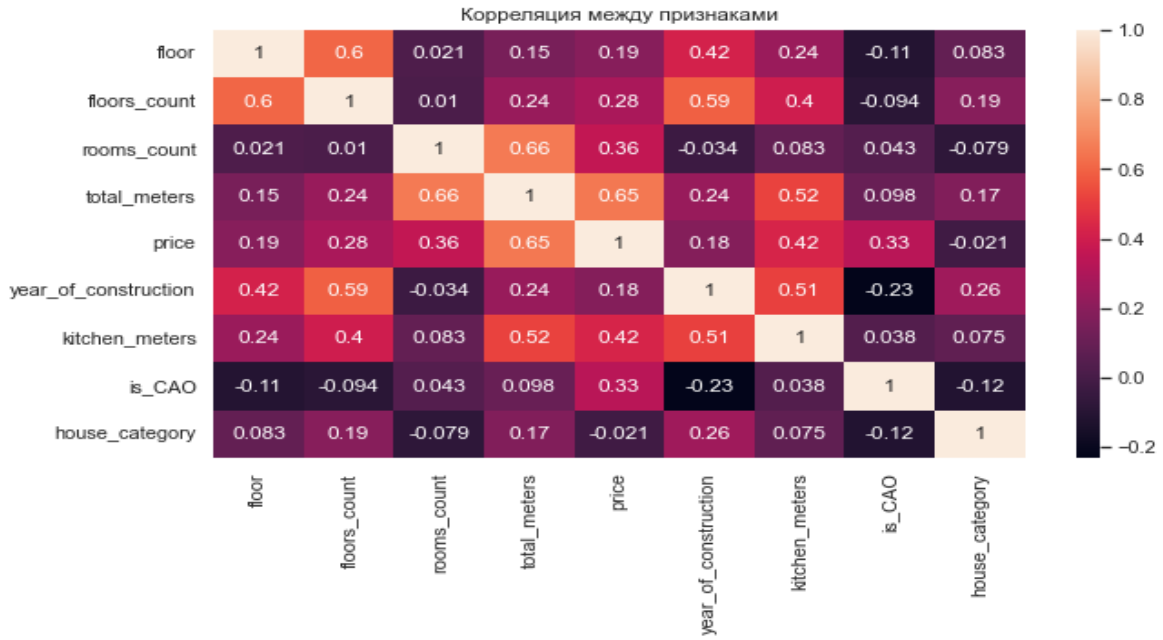

Рис.4 Корреляционная матрица

Рис.5 График остатков для МНК регрессии

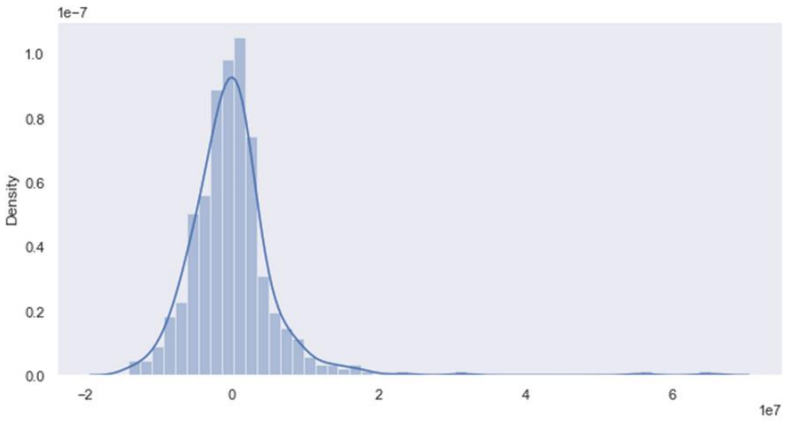

Рис.6 Гистограмма распределения остатков для МНК регрессии

Рис.7 График остатков для 2МНК регрессии

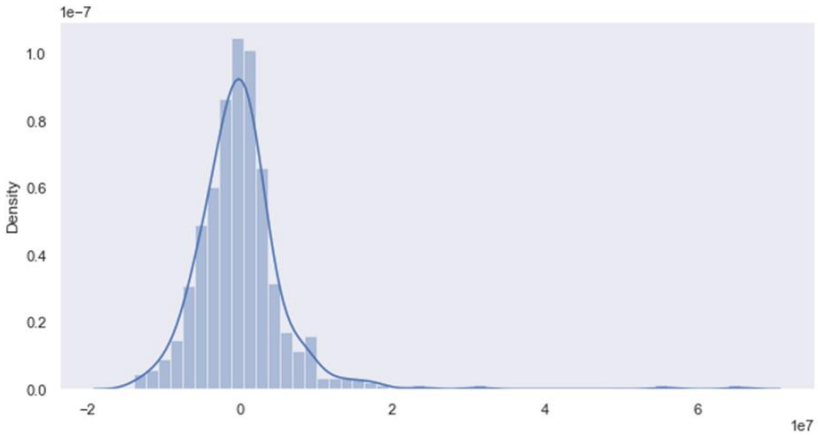

Рис.8 Гистограмма распределения остатков для 2МНК регрессии

Variable	**Coefficient**
Intercept	-7 554 000 (0.000)
Floor	67 130 (0.235)
Floors_count	166 800 (0.000)

Rooms_count	-2 054 000 (0.012)
Total_meters	516 900 (0.000)
Kitchen_meters	48 190 (0.611)
Is_cao	14 120 000 (0.000)
House_category	-2 801 000 (0.000)

R-squared: 0.547

F-statistic: 94.17

(В скобках указано значение p-value)

Табл.2 Результаты МНК регрессии