Эконометрическое исследование зависимости стоимости недвижимости в Санкт-Петербурге от характеристик объекта

Исследование подготовили:

Денисенко Дмитрий Сергеевич

Шестяев Егор Алексеевич

Пилюгин Василий Германович

Оглавление

Актуальность темы исследования.....2
Формулировка гипотез..... 3
Данные..... 4
Разведочный анализ данных (EDA)..... 6
Методы исследования..... 10
Основная часть исследования - построение регрессионной модели, ее тестирование и эндогенность.....11
Проверка гипотез..... 19

Актуальность темы исследования

Санкт-Петербург является вторым по величине жилищным рынком страны - объем предложения только первичного жилья в Питере превышает 3 млн. кв.м. Средневзвешенная цена “квадрата” жилой площади к концу 2024 г. достигала примерно 263 тыс. руб. и за последние 5 лет выросла практически вдвое.

Макроэкономические и регуляторные условия, безусловно, находятся в постоянной динамике: завершение льготных ипотечных программ, ужесточение семейной ипотеки и рост рыночных ставок резко снизили долю ипотечных сделок до 44 % - минимального уровня последних лет. А новые градостроительные правила и ограничения на строительство в историческом центре постоянно меняют “ценовую карту” города.

Однако неизменными детерминантами стоимости недвижимости остаются характеристики объекта: престижность района, в котором находится тот или объект, его удаленность от метро, общая площадь квартиры, жилая площадь, площадь кухни, число комнат, этаж, год постройки дома, его тип и тип отделки в квартире.

Также стоит отметить и то, что недвижимость является одним из главных инвестиционных активов российских домохозяйств, поэтому четкое понимание влияния той или иной детерминанты на стоимость объектов позволит населению принимать более обоснованные решения о вложении в тот или иной объект, снижая риск возникновения “пузырей” на рынке недвижимости. В то же время прозрачность факторов ценообразования повышает доверие к рынку и снижает информационное неравенство между профессиональными игроками и обычными гражданами.

Источники:

- *https://www.rbc.ru/spb_sz/07/12/2024/6751b4799a79471d1ad6dba5*
- *https://stroygaz.ru/news/dwelling/za-pyat-let-novostroyki-v-peterburge-podorozhali-bolee-chem-v-dva-raza*
- *https://nikoliers.ru/analytics/itogi-2024-rossiya-sankt-peterburg-zhilaya-nedvizhimost*

Формулировка гипотез

В ходе нашего исследования мы с командой планируем проверить следующие гипотезы:

- **“Премия исторического центра”** - *объекты, находящиеся в пределах исторического и культурного “центра” Санкт-Петербурга будут стоить дороже, чем объекты, находящиеся в иных районах города ввиду ограниченности предложения, высокой культурной и инфраструктурной ценности локаций и постоянно вводимых ограничений на строительство новых объектов в исторических районах города*
- **“Средняя этажность - детерминант высокой стоимости жилья”** - *квартиры, находящиеся на не самых низких и в то же время на не самых высоких этажах будут иметь более высокую стоимость, чем недвижимость на первых или последних этажах дома. Внизу - шум и пыль, а вверху - ветер, долгое ожидание лифта и нагрев крыши в теплое время года*
- **“Историческая и ценовая значимость памятников архитектуры”** *- в центральных районах города наибольшую ценность, которая выражается в более высокой стоимости, имеют дореволюционные дома, которые являются памятниками уникальной архитектуры Санкт-Петербурга. А на окраинах города ситуация обратная: там большую ценность имеют новостройки, так как старый фонд не несет никакого культурного “шарма”.*

Данные

Выборка представляет из себя набор данных об объявлениях о продаже недвижимости в Санкт-Петербурге (**1380 объектов**), размещенных на крупнейшем российском сервисе по подбору недвижимости **“Циан” на момент 13.04.2025** (**тип данных: cross-section**).

Описание переменных представлено в таблице ниже:

Группа переменных	Переменная	Описание
Идентификация	***url***	*Уникальная ссылка на объявление*
	author, author_type	*Имя продавца и его тип (developer, real_estate_agent, realtor, representative_developer, homeowner)*
Локация	***district***	*Городской район СПб*
	street, house_number	*Адрес объекта*
	metro	*Ближайшая станция метро*
	residental_complex	*Название ЖК, если объект находится в новостройке*
“Физические” характеристики	***total_meters, living_meters, kitchen_meters***	*Общая, жилая площадь и площадь кухни*
	rooms_count	*Число комнат (студия: -1, 1, 2, 3+ комнаты)*
	floor, house_floors_total	*Этаж квартиры и этажность здания*
	year_of_construction	*Год постройки дома*
	house_material_type	*Тип дома (монолитный,*

		панельный, кирпичный, смешанные варианты)
	finish_type	*Отделка квартиры (чистовая, без отделки и т.д.)*
Целевая переменная	***price***	*Запрашиваемая цена за объект недвижимости*

Таблица 1: “Описание переменных”

Описание выборки:

- **Тип продавца:** 53% - девелоперы / официальные представители, 21% - профессиональные агенты / риэлторы, 3% - собственники, 1% - объявления с нераспознанным типом агента
- **Материал здания по декларируемому описанию:** 746 объектов - “монолит” / “монолит-кирпич”, 65 объектов - “панель”, 569 объектов - остальные / не указано
- **Отделка:** 288 объектов - “чистовая”, 133 объекта - “предчистовая”, 205 объектов - “без отделки” / “черновая”, 377 объектов - комбинированные варианты (не содержат явного описания).

Разведочный анализ данных (EDA)

Предварительная обработка данных:

- **Приведение типов** - для потенциально числовых столбцов нами была предпринята попытка перевести каждое значение в число (если не получается, то в NaN)
- **Категоризация и явное указание типов** - “категоризовали” признаки-категории (district, metro и др.) с помощью astype()
- Пропущенные значения - очистили датасет от пропусков, заполнили медианными значениями
- **Добавили новые признаки** - арифметически нашли стоимость недвижимости за кв. м. (price_per_meter) и “возраст дома” (construction_age), добавили бинарные переменные: central (центральный район СПб), middle_floor ("средний этаж") - выше первого, но ниже последнего, old_house - старый жилой фонд (дома с возрастом > 100 лет), и new_build - новострой (отрицательный год постройки), также добавили логарифм цены log_price

Основные дескриптивные статистики:

Переменная	Число наблюдений	Ср. знач.	Станд. отклон.	Минимум	Максимум
total_meters	1380	49	24	14	218
living_meters	1146	22	15	8	120
kitchen_meters	1022	15	8	4	99
rooms_count	1380	1	1	-	3
floor	1380	5	5	-	26
house_floors_total	1380	12	7	3	29
year_of_construction	1342	2016	26	1823	2032
construction_age	1342	9	26	-	202
price	1379	14049255	21053037	3550000	271200000

price_per_meter	1379	247406	149707	45165	1500000

Таблица 2: "Основные описательные статистики"

Графический EDA:

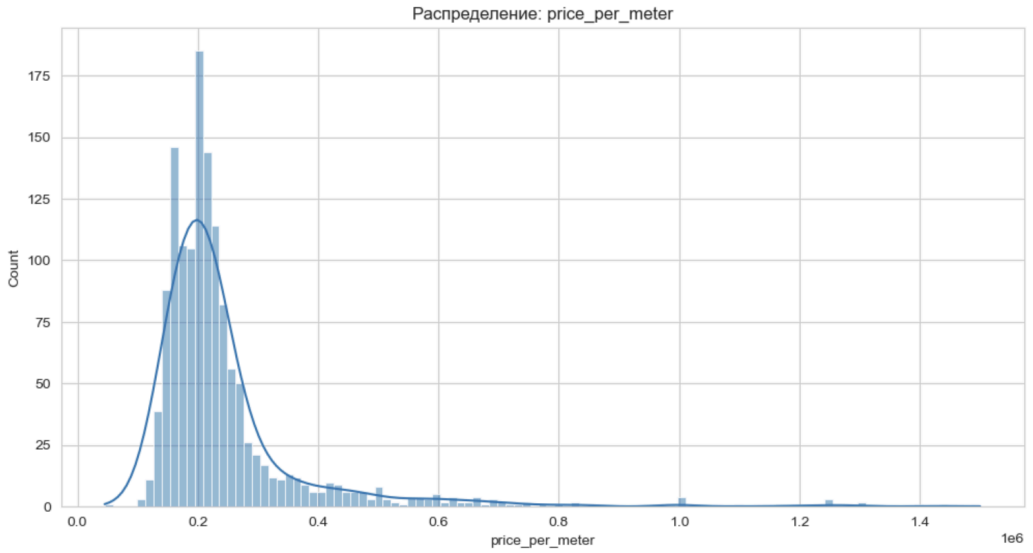

График 1: "Распределение цены недвижимости в СПб за кв. м."

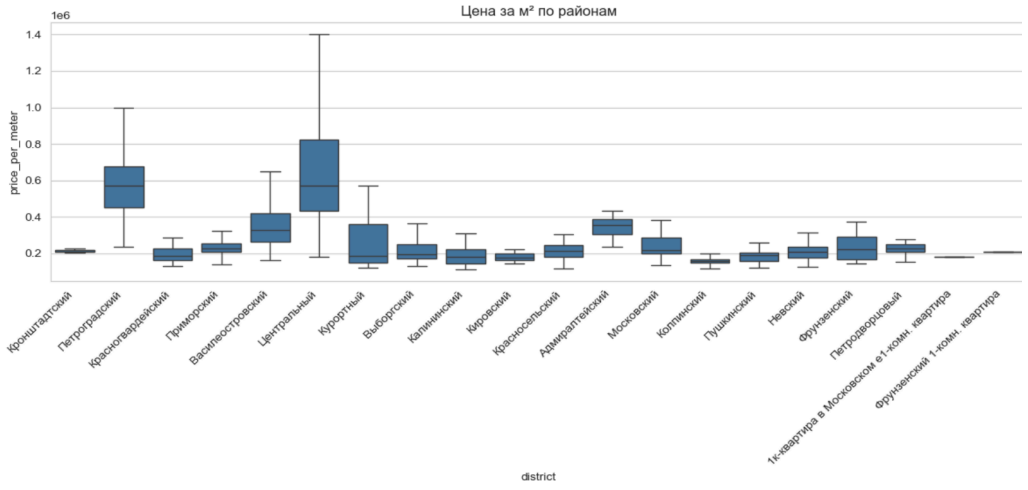

График 2: "Цена за кв. м. недвижимости в СПб по районам"

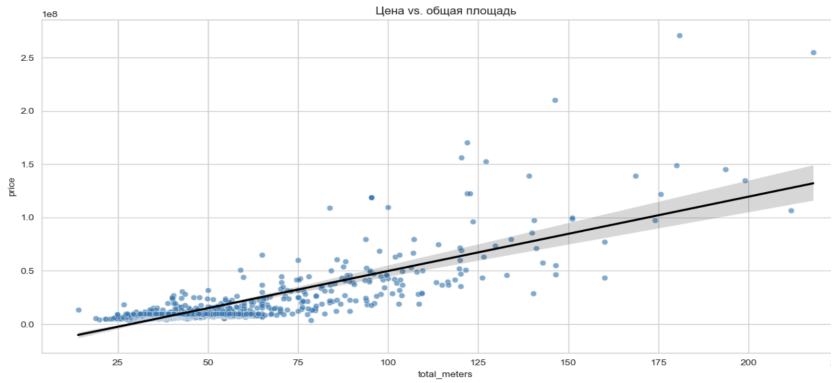

График 3: "Взаимосвязь цены и общей площади недвижимости в СПб"

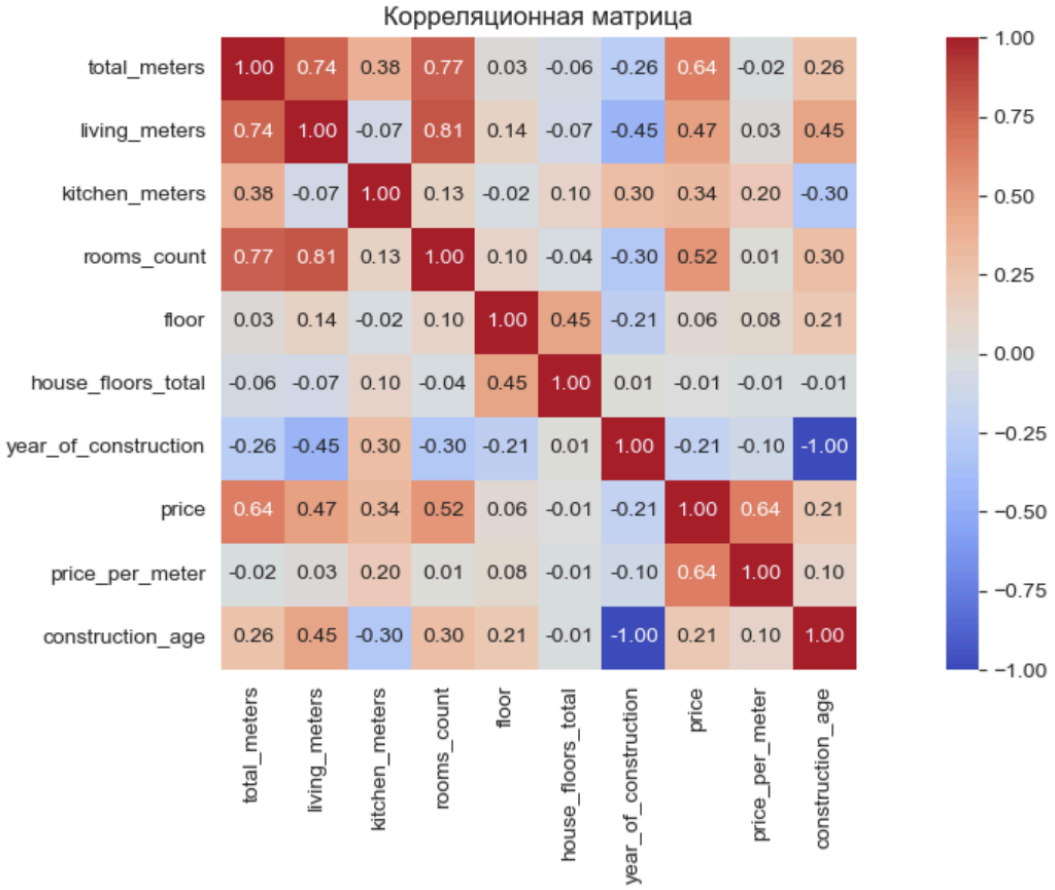

График 4: "Корреляционная матрица переменных"

Предварительные заключения из EDA:

- Вполне логично и ожидаемо, что **распределение цен имеет длинные хвосты** и без трансформаций они будут “тянуть” оценку OLS. Самый простой способ побороть это - взять логарифм от цены либо же отрезать верхний 1% наблюдений
- **Район (district) - самый значимый “неметровый” фактор**
- Вероятно, при добавлении в модель чего-либо “метрового”, кроме total_meters **приведет к мультиколлинеарности.** Можно попробовать побороть это при помощи регуляризации
- Гетероскедастичность и выбросы - на скаттере (График 3) видно, что **разброс цен растет с площадью**. Можно попробовать использовать робастные ошибки, WLS или тесты BP/White
- Наблюдается слабый, но **нестандартный эффект возраста дома**. Такое лучше улавливается нелинейными эффектами.

Методы исследования

- **Потенциальные методы исследования** - оценка “чистого” эффекта факторов на цену (классическая OLS/GLS-модель), учет пространственной зависимости (SAR/SEM, GWR), применение методов многоуровневой иерархии (Mixed-effects, random intercept/slope и др.), изучение неоднородных эффектов вдоль распределения цен (квантильная регрессия QR), использование поправок на селекционное смещение (например, Heckman / Propensity score weighting)
- **Борьба с эндогенностью** - планируются к проведению DWH-тесты, а также тесты на инструментальную значимость переменных (F-тесты, правило Staiger-Stock >= 10 и др.). Также, возможно, будет смысл сравнить робастные и IV-оценки друг с другом
- ***Источники*** *с “технической” информацией, которая может оказаться полезной при применении того или иного метода:*
    - ***Гедоническая регрессия -*** *https://www.jstor.org/stable/1830899, https://www.sciencedirect.com/science/article/pii/S2212609016300383*
    - ***Инструментальные переменные -*** *https://www.nzae.org.nz/wp-content/uploads/2019/07/JamesGraham_HousePricesConsumptionBartikInstrument.pdf*
    - ***SAR / SEM -*** *https://www.sciencedirect.com/science/article/pii/S2211381911000348*
    - ***Geographically Weighted Regression (GWR) -*** *https://www.sciencedirect.com/science/article/abs/pii/S0264837722002101*
    - ***Многоуровневая модель -*** *https://www.researchgate.net/publication/330572553_Analysis_of_Prices_in_the_Housing_Market_Using_Mixed_Models*
    - ***Селективность выборки -*** *https://www.jstor.org/stable/44095499*

Основная часть исследования - построение регрессионной модели, ее тестирование и эндогенность

Модель:

В процессе работы изначально планировалось проверить несколько различных спецификаций регрессионных моделей (LAD / Median Regression, Log-Log, Gls / WLS).

Однако результаты классической OLS модели показали, что для наших данных о недвижимости **модель является** одновременно и консистентной, и наиболее **эффективной**:

OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	331.7
Date:	Sat, 17 May 2025	Prob (F-statistic):	5.06e-314
Time:	19:50:15	Log-Likelihood:	-140.46
No. Observations:	1379	AIC:	298.9
Df Residuals:	1370	BIC:	346.0
Df Model:	8		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	15.1735	0.029	525.083	0.000	15.117	15.230
total_meters	0.0175	0.001	31.463	0.000	0.016	0.019
floor	-0.0005	0.001	-0.306	0.760	-0.003	0.002
house_floors_total	-0.0005	0.001	-0.498	0.619	-0.003	0.002
central	0.7277	0.043	16.959	0.000	0.644	0.812
middle_floor	0.0494	0.017	2.989	0.003	0.017	0.082
old_house	-0.1642	0.161	-1.018	0.309	-0.480	0.152
central_old	-0.1809	0.194	-0.933	0.351	-0.561	0.199
outsk_new	-0.0118	0.014	-0.849	0.396	-0.039	0.015

Omnibus:	140.680	Durbin-Watson:	1.541
Prob(Omnibus):	0.000	Jarque-Bera (JB):	352.059
Skew:	0.572	Prob(JB):	3.56e-77
Kurtosis:	5.195	Cond. No.	1.50e+03

Таблица 3: "Результаты оценки OLS-модели"

- **R^2 = 0.823** свидетельствует о том, что модель объясняет 82.3% дисперсии логарифма цены - **это очень хороший показатель**, наша модель весьма неплохо справляется с описанием данных.

Скорректированный Adj. R^2 = 0.822 говорит о том же, учитывая то, что наша OLS-модель включает в себя несколько предикторов

- **F-Stat = 331.7 и p-value, стремящееся к нулю говорят о том, что** в целом предложенная **модель является статистически значимой**, а нулевая гипотеза о том, что все коэффициенты, кроме константы, равны нулю отвергается с достаточно высокой уверенностью
- **Гетероскедастичность присутствует**, частично побороть ее помогло использование робастных (White/HC1) ошибок SE
- **Оценка коэффициентов модели привела к следующим результатам:**

Переменная	p-value	Интерпретация
const	0.000	Константа - значима
total_meters	0.000	Увеличение общей площади на 1 кв.м. повышает логарифм цены на 0.0175, значимо
floor	0.760	Этаж не влияет на стоимость недвижимости, незначимо
house_floors_total	0.619	Общее число этажей в доме не имеет статистически значимого влияния на стоимость жилья
central	0.000	Факт нахождения жилья в центре города положительно влияет на стоимость, значимо
middle_floor	0.003	Факт нахождения жилья на “среднем” этаже хоть и незначительно, но увеличивает стоимость жилья, значимо
old_house	0.309	“Старость” дома -

		незначима
central_old	0.351	“Центр+Старый дом” - незначимо
outsk_new	0.396	“Окраина+Новый дом” - незначимо

**central_old и outsk_new являются “переменными взаимодействия” для проверки гипотез*

Таблица 4: “Оценки коэффициентов модели и их интерпретация”

Тестирование - мультиколлинеарность:

Диагностика мультиколлинеарности при помощи VIF показала, что инфляция ошибок из-за коррелирующих регрессоров отсутствует - все VIF < 10:

Признак	Значение VIF
total_meters	1.49
floor	1.39
house_floors_total	1.51
central	1.68
middle_floor	1.15
old_house	7.2
central_old	7.36
outsk_new	1.16

Таблица 5: “Результаты VIF-тестирования”

Тестирование - эндогенность переменных:

Изначально мы предположили, что **при проверке на эндогенность “под подозрением” может оказаться общая площадь объекта** (total_meters) в силу ряда причин:

- **Теория гедонических цен** гласит, что в классической спецификация цена есть функция от качества, размера, локации и т. п. Размер и цена формально “совместно определяются” покупателем и продавцом. У покупателя есть бюджет и предпочтения - он одновременно решает, сколько метров может / хочет купить и сколько готов заплатить. Такого рода “совместное решение” покупателя и продавца в теории может коррелировать с компонентой ошибки - ненаблюдаемым “уровнем платежеспособности” конкретного покупателя, общим состоянием рынка и др.
- **Некоторая автоселекция по сегменту** - более просторные квартиры, как правило, относятся к дорогому сегменту, находятся в лучших локациях и имеют более качественную отделку, а эти характеристики частично скрыты в “эпсилон”-ошибке. Если эти “скрытые” качества не полностью контролируются в X-матрице, то площадь улавливает их и, как следствие, возникает эндогенность
- У нас есть **living_meters и kitchen_meters** - они **сильно коррелируют с total_meters** (значение коэффициента корреляции Пирсона: 0.74 и 0.38 соответственно), что вполне логично. Считать living_meters и kitchen_meters драйверами стоимости жилья не принято, если в уравнении регрессии уже присутствует общая площадь объекта. Поэтому естественным кажется проводить тесты на эндогенность именно относительно общей площади (total_meters).

Результаты двухэтапного DWH-теста показали, что признаков эндогенности общей площади объекта (total_meters) в нашей выборке не наблюдается:

- **1-ый этап DWH-теста** - регрессировали X = total_meters на инструменты и получили остатки Ui = X1i - X1i^, после чего проверили силу инструментов при помощи F-теста, по результатам которого “сила” инструментов была подтверждена (F-Stat = 778 >> 10) по правилу Staiger-Stock:

OLS Regression Results

Dep. Variable:	total_meters	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	778.2
Date:	Sat, 17 May 2025	Prob (F-statistic):	0.00
Time:	19:56:28	Log-Likelihood:	-5102.5
No. Observations:	1379	AIC:	1.022e+04
Df Residuals:	1369	BIC:	1.028e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.5571	0.974	5.704	0.000	3.646	7.468
living_meters	1.2294	0.022	55.373	0.000	1.186	1.273
kitchen_meters	1.2399	0.044	28.047	0.000	1.153	1.327
floor	-0.0914	0.063	-1.442	0.150	-0.216	0.033
house_floors_total	-0.0782	0.049	-1.587	0.113	-0.175	0.018
central	10.3196	1.041	9.916	0.000	8.278	12.361
middle_floor	0.5736	0.634	0.905	0.366	-0.670	1.817
old_house	-2.1790	4.962	-0.439	0.661	-11.912	7.554
central_old	-3.0803	5.379	-0.573	0.567	-13.633	7.472
outsk_new	-1.3074	0.605	-2.160	0.031	-2.495	-0.120

Omnibus:	945.766	Durbin-Watson:	1.844
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22249.704
Skew:	2.850	Prob(JB):	0.00
Kurtosis:	21.835	Cond. No.	833.

Таблица 6: “Результаты 1-го этапа DWH-теста”

- **2-ой этап DWH-теста** - расширенное структурное уравнение, в котором к исходному уравнению лог-цены добавляется Ui^. Также в параметрах тестов мы использовали ковариационную матрицу HC1, чтобы не полагаться на гомоскедастичность. Далее при помощи t-теста мы проверили нулевую гипотезу gamma = 0 (т. е. total_meters является переменной экзогенной) - смотрим на переменную fs_resid. P > |z| = 0.44 > 0.05, а значит, что H0 - не отвергается, следовательно, связи между площадью и ошибкой структурной регрессии нет и OLS-коэффициент по площади несмещен. Вторая, «классическая» форма DWH-теста (Hausman-F по разнице OLS vs IV-параметров) дает то же самое: Hausman-p > 0.3:

DWH — значимость fs_resid:

	coef	std err	z	P>|z|	[0.025	0.975]
const	15.1860	0.032	481.525	0.000	15.124	15.248
total_meters	0.0172	0.001	27.303	0.000	0.016	0.018
floor	−0.0004	0.001	−0.281	0.779	−0.003	0.002
house_floors_total	−0.0006	0.001	−0.573	0.567	−0.003	0.001
central	0.7384	0.046	16.056	0.000	0.648	0.829
middle_floor	0.0505	0.017	3.017	0.003	0.018	0.083
old_house	−0.1584	0.159	−0.997	0.319	−0.470	0.153
central_old	−0.1863	0.192	−0.970	0.332	−0.563	0.190
outsk_new	−0.0130	0.014	−0.935	0.350	−0.040	0.014
fs_resid	0.0011	0.001	0.772	0.440	−0.002	0.004

Таблица 7: Результаты 2-го этапа DWH-теста”

Тестирование - возможное смещение OLS:

Мы решили провести дополнительное тестирование смещения OLS-модели при помощи IV-методов 2SLS, так как IV-оценка остается консистентной, даже когда OLS смещена (в теории смещена). Для этого мы прошли по следующим шагам:

- **Шаг 1** - спрогнозировали площадь инструментами, предварительно определив подозрительный RHS-фактор - total_meters и нашли правдоподобные инструменты - living_meters и kitchen_meters и все прочие экзогенные регрессоры
- **Шаг 2** - заменили фактическую площадь на её прогноз в уравнении цены и сравнили результаты 2SLS с OLS, получив практически идентичные коэффициенты - это значит, что наша изначальная OLS-модель несмещена и тестировать какие-либо иные спецификации особого смысла нет (Hausman-статистика также не отвергает OLS):

IV-/2SLS-оценка (робастные SE):

OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.744
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	226.0
Date:	Sat, 17 May 2025	Prob (F-statistic):	5.10e-244
Time:	20:02:23	Log-Likelihood:	-394.75
No. Observations:	1379	AIC:	807.5
Df Residuals:	1370	BIC:	854.6
Df Model:	8		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	15.1860	0.041	373.255	0.000	15.106	15.266
total_meters_hat	0.0172	0.001	20.737	0.000	0.016	0.019
floor	-0.0004	0.002	-0.230	0.818	-0.004	0.003
house_floors_total	-0.0006	0.001	-0.466	0.641	-0.003	0.002
central	0.7384	0.068	10.816	0.000	0.605	0.872
middle_floor	0.0505	0.020	2.573	0.010	0.012	0.089
old_house	-0.1584	0.177	-0.896	0.370	-0.505	0.188
central_old	-0.1863	0.213	-0.874	0.382	-0.604	0.231
outsk_new	-0.0130	0.016	-0.822	0.411	-0.044	0.018

Omnibus:	328.560	Durbin-Watson:	1.534
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1354.615
Skew:	1.084	Prob(JB):	7.07e-295
Kurtosis:	7.345	Cond. No.	1.48e+03

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.48e+03. This might indicate that there are strong multicollinearity or other numerical problems.

F-статистика 1-й стадии (Staiger-Stock): 778.16

Таблица 8: "Результаты тестирования через 2SLS-методы"

Проверка гипотез

- **“Премия исторического центра” - гипотеза подтверждена.** Квартиры в четырех «центральных» районах Санкт-Петербурга (Адмиралтейский, Центральный, Петроградский, Василеостровский) действительно стоят дороже. Центр СПб сочетает в себе как ограниченное предложение (разного рода “охранные зоны”, запреты на высотное строительство и др.), так и высокую концентрацию офисов, сервисов, разного рода культурных объектов, что делает перечисленные районы крайне привлекательными для жизни и обеспечивает высокий спрос на жилье в этой части города. Оцененный коэффициент в лог-модели оказался равен 0.7277, что в “классической” шкале означает +107% к цене относительно аналогичного объекта за пределами центра. Это, кстати, достаточно близко к надбавке к оценкам других российских и европейских городов (80-120%)
- **"Средняя этажность - детерминант высокой стоимости жилья" - гипотеза также подтверждена.** Низшие этажи “страдают” от шума и пыли, самые верхние - от ветра, жары, долгого ожидания лифта и др. Оцененный коэффициент при регрессии составил 0.0494, что свидетельствует о примерно 5%-ой премии к стоимости жилья на “средних” этажах
- **"Историческая и ценовая значимость памятников архитектуры" - не подтверждена.** Данную гипотезу мы проверяли “перекрестным” образом при помощи переменных взаимодействия - “старый_фонд+центр” и “новострой+окраина”. Коэффициенты при обеих переменных статистически неотличимы от нуля. Такой результат может быть связан, как вариант, с сильной гетерогенностью нашей “шарм-премии” - дореволюционный дом может быть как отреставрированным «до бриллианта», так и требовать капитальных вложений. Также стоит отметить и то, что наша классификация old_house по критерию возраста дома (>= 100 лет) могла оказаться слишком грубой - и туда попали не только архитектурные шедевры Санкт-Петербурга, но и аварийные постройки. А еще - эффект уже мог быть учтен переменной central: модель могла «забрать» всю премию локации в основной коэффициент, а дополнительная разница «старое vs не старое»

оказалась сравнительно слабой.