**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**

**імені ІГОРЯ СІКОРСЬКОГО»**

**Інститут прикладного системного аналізу**

**Кафедра математичних методів системного аналізу**

**Звіт**

**про виконання лабораторної роботи**

**на тему «Регресійний аналіз за неоднорідних та корельованих спостережень»**

**з дисципліни «Моделювання складних систем»**

**Виконали:**

студенти IV курсу, групи КІ-02

Панасенко Володимир Володимирович,

Жутник Дар'я Олександрівна,

Наход Олексій Олександрович,

Задорожний Олег Андрійович,

Трач Євгеній Сергійович,

Ткаченко Дарина Володимирівна,

Дубик Наталія Анатоліївна,

Скоропадський Артем Олександрович,

Висоцький Дмітрій,

Васильківський Олексій Олександрович

**Перевірила:**

проф., д.т.н. Чумаченко Олена Іллівна

Київ – 2024

## МЕТА

Знайти і оцінити регресійну модель для даної вибірки, використовуючи різні методи оцінки коефіцієнтів регресії, та провести статистичний аналіз отриманих результатів.

## ПОСТАНОВКА ЗАДАЧІ

1. Перевірити вибірку на однорідність (критерій Бартлета, Кохрена, модифікація М-критерію Бартлета, граничний критерій неоднорідності дисперсій Харисона і Мак-Кейба, параметричний критерій Голдфільда і Квандта).
2. Оцінити коефіцієнти регресії за допомогою методу найменших квадратів.
3. Оцінити коефіцієнти регресії за допомогою узагальненого методу найменших квадратів.
4. Виконати статистичний аналіз результатів оцінювання.
5. Оцінити коефіцієнти регресії першого порядку.
6. Здійснити оцінювання узагальненої регресійної моделі.

## ТЕОРЕТИЧНІ ВІДОМОСТІ

### Критерій Бартлета

Критерій Бартлета – це статистичний тест, який використовується для перевірки гіпотези про однорідність дисперсій у групах даних. Цей критерій чутливий до нормального розподілу даних.

Його головна ідея полягає в порівнянні дисперсій між групами даних. Якщо дисперсії однорідні (тобто різниці між групами нестатистично значущі), то значення критерію буде незначним, що свідчить про прийняття нульової гіпотези про однорідність дисперсій. В іншому випадку, якщо дисперсії статистично відрізняються, значення критерію буде великим, що означатиме відхилення від нульової гіпотези.

Формально, для застосування критерію Бартлета використовують таку послідовність кроків:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії всіх груп даних однорідні, тоді як альтернативна гіпотеза стверджує, що дисперсії відрізняються хоча б у двох групах.
2. Обчислити значення критерію Бартлета.
3. Порівняти обчислене значення критерію з критичним значенням з відповідної таблиці для визначення статистичної значущості результату.
4. Зробити висновок про однорідність або неоднорідність дисперсій в групах даних на основі порівняння обчисленого значення критерію та його критичного значення.

Зокрема, для застосування критерію Бартлета важливо мати рівність кількості спостережень у різних групах та нормальний розподіл даних.

### Критерій Кохрена

Критерій Кохрена – це статистичний тест, який використовується для перевірки однорідності дисперсій у групах даних.

Основна ідея критерію полягає у порівнянні максимальної та мінімальної дисперсій між групами. Якщо ці значення статистично не відрізняються, то припускається, що дисперсії у всіх групах однорідні.

Процедура застосування критерію Кохрена така:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії у всіх групах однорідні, альтернативна гіпотеза – що є хоча б одна відмінність між дисперсіями.
2. Обчислити максимальну та мінімальну дисперсії в групах.
3. Обчислити вибіркову дисперсію по кожній групі.
4. Обчислити статистику Кохрена шляхом ділення максимальної вибіркової дисперсії на суму всіх вибіркових дисперсій.
5. Порівняти обчислене значення статистики Кохрена з критичним значенням з таблиці для визначення статистичної значущості результату.
6. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення статистики Кохрена та його критичного значення.

Критерій Кохрена є досить простим у застосуванні та може бути використаний для перевірки однорідності дисперсій у випадку, коли кількість груп не більше деякого певного обмеження, зазвичай до десяти.

### Модифікація М-критерію Бартлета

Модифікація М-критерію Бартлета – це адаптація критерію Бартлета для випадку, коли дані мають нормальний розподіл, але не мають однорідність дисперсій.

Процедура застосування модифікації М-критерію Бартлета аналогічна до звичайного критерію Бартлета, але перед застосуванням необхідно виконати попередню нормалізацію даних. Після цього кроки такі:

1. Обчислити вибіркові дисперсії по кожній групі.
2. Обчислити середню вибіркову дисперсію.
3. Обчислити статистику М-критерію Бартлета, яка базується на порівнянні середньої вибіркової дисперсії з максимальною вибірковою дисперсією.
4. Порівняти обчислене значення статистики М-критерію Бартлета з критичним значенням з таблиці для визначення статистичної значущості результату.
5. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення статистики М-критерію Бартлета та його критичного значення.

Цей підхід дозволяє враховувати нормальний розподіл даних та вирішує проблему нерівномірності дисперсій у групах.

### Граничний критерій неоднорідності дисперсій Харисона і Мак-Кейба

Граничний критерій неоднорідності дисперсій Харісона і Мак-Кейба – це статистичний тест, який використовується для оцінки однорідності дисперсій у групах даних. Він оцінює максимальне відношення дисперсій між групами.

Процедура застосування цього критерію наступна:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що всі групи мають однакову дисперсію, тоді як альтернативна гіпотеза стверджує, що є хоча б одна відмінність між дисперсіями.
2. Обчислити вибіркові дисперсії по кожній групі.
3. Знайти максимальне і мінімальне значення вибіркових дисперсій.
4. Обчислити відношення максимальної до мінімальної вибіркової дисперсії.
5. Знайти критичне значення граничного відношення дисперсій відповідно до обраного рівня значущості та кількості груп.
6. Порівняти обчислене відношення з критичним значенням.
7. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого відношення з критичним значенням.

Цей критерій виявляється корисним в тих випадках, коли потрібно оцінити неоднорідність дисперсій у групах, особливо коли розподіл даних може бути важко інтерпретувати за допомогою інших методів.

### Параметричний критерій Голдфільда і Квандта

Параметричний критерій Голдфільда і Квандта – це статистичний тест, який використовується для перевірки гіпотези про однорідність дисперсій у групах даних. Цей критерій передбачає, що дані мають нормальний розподіл.

Процедура застосування критерію Голдфільда і Квандта наступна:

1. Сформулювати нульову та альтернативну гіпотези. Нульова гіпотеза припускає, що дисперсії в усіх групах однакові, тоді як альтернативна гіпотеза стверджує, що є хоча б одна відмінність між дисперсіями.
2. Обчислити вибіркові дисперсії по кожній групі.
3. Впорядкувати вибіркові дисперсії за зростанням.
4. Визначити кількість спостережень у кожній групі, включаючи всі групи.
5. Знайти критичне значення для потрібного рівня значущості та кількості груп у таблиці критичних значень для критерію Голдфільда і Квандта.
6. Порівняти обчислене критерієм значення тесту з критичним значенням.
7. Зробити висновок щодо однорідності або неоднорідності дисперсій в групах на основі порівняння обчисленого значення критерію з критичним.

Цей критерій є корисним у випадках, коли потрібно перевірити однорідність дисперсій у групах і дані відповідають умовам нормального розподілу. Він може бути особливо корисним, коли кількість груп велика і групи мають різний обсяг.

### Метод найменших квадратів

Метод найменших квадратів (МНК) – це стандартний метод оцінювання параметрів регресійної моделі, який полягає в мінімізації суми квадратів відхилень між спостережуваними значеннями залежної змінної і значеннями, передбаченими моделлю.

Оцінки, отримані за допомогою МНК, надають найкращі значення параметрів для регресійної моделі в тому сенсі, що вони мінімізують суму квадратів відхилень. Цей метод є широко використовуваним і дозволяє отримати найкращі оцінки параметрів, якщо припущення про нормальний розподіл похибок виконується.

Для виконання статистичного аналізу результатів оцінювання параметрів регресії за допомогою методу найменших квадратів, можна виконати наступні кроки:

1. Перевірка припущень моделі

Перш ніж переходити до аналізу результатів, слід перевірити припущення, які лежать в основі регресійної моделі, такі як нормальність розподілу похибок, лінійність зв'язку між змінними тощо. Для цього можна використовувати графічні методи (наприклад, Q-Q графіки, графіки залишків), а також статистичні тести (наприклад, тест Шапіро-Уілка для нормальності розподілу похибок).

1. Оцінка значущості параметрів моделі

Для оцінки статистичної значущості кожного з параметрів регресії можна використовувати t-тест. Порівнюючи оцінки коефіцієнтів регресії з їхніми стандартними помилками, можна обчислити t-статистики та визначити статистичну значущість параметрів на певному рівні значущості.

1. Оцінка адекватності моделі

Для оцінки того, наскільки добре побудована модель пояснює варіацію в залежній змінній, можна використовувати різноманітні статистичні метрики, такі як коефіцієнт детермінації (R2), кореляційний коефіцієнт Пірсона, а також F-тест на значущість моделі в цілому.

1. Аналіз залишків моделі

Перевірка залишків моделі є важливим етапом аналізу. Графіки залишків (наприклад, розподіл залишків, графік "залишків-прогнозованих значень") можуть допомогти виявити систематичні відхилення в моделі. Додатково можна застосувати тести на автокореляцію залишків, якщо це відповідає контексту дослідження.

1. Перевірка мультіколінеарності

Якщо в моделі присутня мультіколінеарність (високий рівень кореляції між незалежними змінними), це може спричинити проблеми при оцінці параметрів. Для виявлення мультіколінеарності можна використовувати статистичні тести, такі як визначення факторного числа, або обчислення кореляційної матриці між змінними.

1. Інтерпретація результатів

Нарешті, після виконання усіх статистичних аналізів, слід зробити висновки щодо статистичної значущості та адекватності моделі, а також провести інтерпретацію кожного з коефіцієнтів регресії в контексті дослідження.

### Узагальнений метод найменших квадратів

Узагальнений метод найменших квадратів (УМНК) – це метод оцінювання параметрів регресійної моделі, який використовується у випадку, коли відомо, що похибки моделі не мають гауссівського розподілу. Цей метод є розширенням класичного методу найменших квадратів (МНК), який передбачає нормальний розподіл похибок.

Припустимо, що ми маємо регресійну модель у вигляді:

Основна ідея узагальненого методу найменших квадратів полягає в мінімізації суми квадратів вагованих відхилень між спостережуваними значеннями і значеннями, передбаченими моделлю. Ваги при цьому встановлюються відповідно до властивостей похибок.

Процедура оцінювання параметрів за допомогою УМНК включає наступні кроки:

1. Сформулювати регресійну модель і припущення про похибки моделі.
2. Обрати ваги для мінімізації функції втрат (зазвичай ваги вибираються залежно від природи похибок).
3. Мінімізувати суму квадратів вагованих відхилень між спостережуваними значеннями і значеннями, передбаченими моделлю.
4. Оцінити параметри моделі, що мінімізують цю суму.

Оцінки параметрів, отримані за допомогою УМНК, зазвичай є менш ефективними порівняно з оцінками, отриманими за допомогою МНК у випадку, коли похибки мають нормальний розподіл. Однак УМНК залишається корисним інструментом у випадках, коли це припущення про розподіл похибок не виконується.

Для оцінки узагальненої регресійної моделі можна використовувати методи, такі як метод найменших квадратів з вагами, робустні методи оцінювання (наприклад, регресія Хубера), або методи, які враховують структуру похибок (наприклад, гребнева регресія або ласо).

Основні кроки оцінювання узагальненої регресійної моделі:

1. Вибір моделі

Спочатку необхідно визначити структуру моделі, включаючи вибір залежних і незалежних змінних, а також вибір форми функціонального відношення між ними.

1. Визначення функції ваг

У випадках, коли похибки мають гетероскедастичну або асиметричну структуру, можна використовувати функції ваг, які дають більший вагу менш варіативним спостереженням.

1. Оцінка параметрів моделі

Після визначення моделі та функції ваг можна застосувати методи оцінювання, такі як МНК з вагами або робустні методи, для отримання оцінок параметрів моделі.

1. Оцінка статистичної значимості параметрів

Після отримання оцінок параметрів можна провести тести на їх статистичну значимість, такі як t-тест або тести Вальда.

1. Оцінка адекватності моделі

Для оцінки адекватності моделі можна використовувати різноманітні статистичні тести, такі як тест на гомоскедастичність похибок, тест на адекватність підгонки моделі тощо.

1. Перевірка припущень

Нарешті, необхідно перевірити, чи виконуються припущення, на яких ґрунтується модель, та вирішити, чи вони задовольняються.

## ОПИС ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

### Вхідні дані

У якості вхідних даних були взяті дані про ринок нерухомості з [Kaggle](https://www.kaggle.com/datasets/rukenmissonnier/final-house/data).

Незалежні змінні:

– к-ть спалень

– площа

– відстань до центру міста

– відстань до метро

– поверх

– вік

Залежна змінна:

– вартість

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| № | к-ть спалень | площа (м2) | відстань до центру міста | відстань до метро | поверх | вік (років) | вартість |
| 1 | 1 | 26 | 1287 | 204 | 22 | 67 | 96005 |
| 2 | 1 | 35 | 1855 | 187 | 8 | 30 | 92474 |
| … | … | … | … | … | … | … | … |
| 4307 | 2 | 54 | 255 | 26 | 18 | 0 | 93743 |

### Хід роботи

Імпортуємо необхідні бібліотеки:

import numpy as np

import pandas as pd

import statsmodels.api as sm

from scipy import stats

import matplotlib.pyplot as plt

import seaborn as sns  
  
np.random.seed(0)

Імпортуємо вхідні дані:

df = pd.read\_csv('house.csv')  
df.head()

bedroom\_count net\_sqm center\_distance metro\_distance floor age \  
0 1 26.184098 1286.68 204.003817 22 67   
1 1 34.866901 1855.25 186.980360 8 30   
2 1 36.980709 692.09 111.224999 24 24   
3 1 17.445723 1399.49 237.998760 1 66   
4 1 52.587646 84.65 100.996400 20 3   
  
 price   
0 96004.804557   
1 92473.722568   
2 98112.519942   
3 92118.326874   
4 98976.653176

Проаналізуємо головні статистики:

df.describe()

bedroom\_count net\_sqm center\_distance metro\_distance \  
count 4308.000000 4308.000000 4308.000000 4308.000000   
mean 3.732823 118.941327 1090.544301 105.665358   
std 2.476989 95.469802 555.325830 60.750825   
min 1.000000 10.244518 11.800000 1.245338   
25% 2.000000 52.240659 635.897500 56.156897   
50% 3.000000 91.828201 1143.135000 101.388012   
75% 5.000000 150.987591 1554.280000 148.481068   
max 17.000000 750.971604 1999.840000 330.275317   
  
 floor age price   
count 4308.000000 4308.000000 4308.000000   
mean 9.397168 48.576834 95701.196185   
std 7.517018 27.621465 3919.823988   
min 1.000000 0.000000 86113.592974   
25% 2.000000 25.000000 92934.251436   
50% 8.000000 48.000000 95337.792918   
75% 16.000000 72.000000 97980.675094   
max 24.000000 97.000000 118134.771190

Для перевірки вибірки на однорідність розділимо її на підвибірки на основі bedroom\_count. Групи з мають невелику кількість спостережень, тому об’єднаємо їх у групу “10+”.

df['bedroom\_count\_grouped'] = df['bedroom\_count'].apply(lambda x: x if x < 10 else '10+')  
df\_grouped = df.groupby('bedroom\_count\_grouped')['price']  
price\_data\_by\_bedrooms = [group for \_, group in df\_grouped]  
group\_analysis\_merged = pd.DataFrame({  
 'Sample Count': df\_grouped.count(),  
 'Variance': df\_grouped.var()  
})  
group\_analysis\_merged

Sample Count Variance  
bedroom\_count\_grouped   
1 612 7.290650e+06  
2 1130 8.501346e+06  
3 781 8.923768e+06  
4 588 1.071874e+07  
5 301 1.133771e+07  
6 258 1.442870e+07  
7 186 1.287068e+07  
8 156 1.771164e+07  
9 140 1.731087e+07  
10+ 156 2.562785e+07

### Критерій Бартлета

bartlett\_result = stats.bartlett(\*price\_data\_by\_bedrooms)  
bartlett\_result

BartlettResult(statistic=219.14294857226892, pvalue=3.169397674040137e-42)

Значення p\_value вказує на необхідність відхилити нульову гіпотезу, тобто вибірка має неоднорідні дисперсії.

### Критерій Кохрена

def cochran\_c\_test(\*args):  
 variances = [np.var(a, ddof=1) for a in args]  
 max\_variance = max(variances)  
 sum\_variances = sum(variances)  
 C = max\_variance / sum\_variances

return C  
  
cochran\_c\_result = cochran\_c\_test(\*price\_data\_by\_bedrooms)  
cochran\_c\_result

0.1902277226687444

Значення G перевищує відповідне табульоване значення у таблиці Кохрена (), тому відкидаємо нульову гіпотезу, тобто вибірка має неоднорідні дисперсії.

### undefinedМодифікація М-критерію Бартлета

log\_price\_data\_by\_bedrooms = [np.log(group + 1) for group in price\_data\_by\_bedrooms]  
bartlett\_stat\_transformed, bartlett\_p\_transformed = stats.bartlett(\*log\_price\_data\_by\_bedrooms)  
bartlett\_stat\_transformed, bartlett\_p\_transformed

(158.87803607530577, 1.2701629783538752e-29)

Значення p\_value вказує на необхідність відхилити нульову гіпотезу, тобто вибірка має неоднорідні дисперсії.

### Граничний критерій неоднорідності дисперсій Харисона і Мак-Кейба

HMC = 0.36042, p-value < 2.2e-16

Значення p\_value вказує на необхідність відхилити нульову гіпотезу, тобто вибірка має неоднорідні дисперсії.

### Параметричний критерій Голдфільда і Квандта

X = df[['bedroom\_count', 'net\_sqm', 'center\_distance', 'metro\_distance', 'floor', 'age']]  
y = df['price']

sm.stats.diagnostic.het\_goldfeldquandt(y, X, idx=0, drop=0.2)

(0.8429282740148406, 0.9997238150809711, 'increasing')

Значення p\_value не дозволяє відхилити нульову гіпотезу.

### Метод найменших квадратів

results\_ols = sm.OLS(y, sm.add\_constant(X)).fit()  
results\_ols.summary()

<class 'statsmodels.iolib.summary.Summary'>  
"""  
 OLS Regression Results   
==============================================================================  
Dep. Variable: price R-squared: 0.719  
Model: OLS Adj. R-squared: 0.719  
Method: Least Squares F-statistic: 1835.  
Date: Sat, 16 Mar 2024 Prob (F-statistic): 0.00  
Time: 21:34:10 Log-Likelihood: -39021.  
No. Observations: 4308 AIC: 7.806e+04  
Df Residuals: 4301 BIC: 7.810e+04  
Df Model: 6   
Covariance Type: nonrobust   
===================================================================================  
 coef std err t P>|t| [0.025 0.975]  
-----------------------------------------------------------------------------------  
const 9.46e+04 129.991 727.773 0.000 9.43e+04 9.49e+04  
bedroom\_count 308.5671 21.016 14.682 0.000 267.364 349.770  
net\_sqm 25.1994 0.505 49.909 0.000 24.209 26.189  
center\_distance -3.3608 0.070 -47.908 0.000 -3.498 -3.223  
metro\_distance 6.9984 0.642 10.909 0.000 5.741 8.256  
floor 121.6607 4.989 24.385 0.000 111.880 131.442  
age -26.1286 1.149 -22.739 0.000 -28.381 -23.876  
==============================================================================  
Omnibus: 641.697 Durbin-Watson: 1.956  
Prob(Omnibus): 0.000 Jarque-Bera (JB): 7305.816  
Skew: -0.321 Prob(JB): 0.00  
Kurtosis: 9.347 Cond. No. 5.08e+03  
==============================================================================  
  
Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
[2] The condition number is large, 5.08e+03. This might indicate that there are  
strong multicollinearity or other numerical problems.  
"""

Аналіз результатів:

* 1. Загальна інформація
* , тобто модель пояснює приблизно 71,9% дисперсії ціни. Показник має досить високе значення, що вказує на вдале підлаштування моделі.
* , тобто кількість регресорів є вдалою для кількості спостережень.
* , , що вказує на високу важливість моделі.
  1. Коефіцієнти
* const (перетин): 94600. Це базова ціна, коли всі інші змінні предиктора дорівнюють нулю.
* bedroom\_count: кожна додаткова спальня пов’язана зі збільшенням ціни приблизно на 308,5671.
* net\_sqm: за кожен квадратний метр збільшення площі ціна зростає приблизно на 25,1994.
* center\_distance: ціна зменшується приблизно на 3,3608 за кожну одиницю збільшення відстані від центру.
* metro\_distance: ближче розташування до метро збільшує ціну приблизно на 6,9984 одиниці за одиницю відстані.
* floor: перебування на більш високому поверсі пов’язане зі збільшенням ціни, зокрема, кожен поверх вище пов’язаний зі збільшенням приблизно на 121,6607.
* age: ціна зменшується приблизно на 26,1286 одиниць за кожен рік збільшення віку.
* Усі коефіцієнти є статистично значущими (дуже малий для всіх коефіцієнтів).
* Вплив усіх коефіцієнтів збігається з їхнім інтуїтивним сприйняттям.
  1. Діагностика моделі
* P-value у тестах Omnibus та Jarque-Bera вказують на те, що залишки скоріш за все не є нормально розподіленими.
* Коефіцієнт асиметрії вказує на незначну асиметрію залишків.
* Коефіцієнт ексцесу вказує на те, що залишки мають розподіл з важкими хвостами.
* Критерій Дарбіна-Уотсона вказує на відсутність автокореляції в залишках.

### Узагальнений метод найменших квадратів

weights = 1 / (results\_ols.fittedvalues \*\* 2)  
  
results\_wls = sm.WLS(y, sm.add\_constant(X), weights=weights).fit()  
results\_wls.summary()

<class 'statsmodels.iolib.summary.Summary'>  
"""  
 WLS Regression Results   
==============================================================================  
Dep. Variable: price R-squared: 0.716  
Model: WLS Adj. R-squared: 0.715  
Method: Least Squares F-statistic: 1806.  
Date: Sat, 16 Mar 2024 Prob (F-statistic): 0.00  
Time: 21:36:35 Log-Likelihood: -38960.  
No. Observations: 4308 AIC: 7.793e+04  
Df Residuals: 4301 BIC: 7.798e+04  
Df Model: 6   
Covariance Type: nonrobust   
===================================================================================  
 coef std err t P>|t| [0.025 0.975]  
-----------------------------------------------------------------------------------  
const 9.458e+04 128.359 736.877 0.000 9.43e+04 9.48e+04  
bedroom\_count 296.5598 20.911 14.182 0.000 255.564 337.556  
net\_sqm 25.6590 0.514 49.884 0.000 24.651 26.667  
center\_distance -3.3877 0.069 -48.842 0.000 -3.524 -3.252  
metro\_distance 7.3604 0.633 11.636 0.000 6.120 8.601  
floor 122.1195 4.890 24.975 0.000 112.533 131.706  
age -26.1996 1.132 -23.137 0.000 -28.420 -23.979  
==============================================================================  
Omnibus: 547.852 Durbin-Watson: 1.961  
Prob(Omnibus): 0.000 Jarque-Bera (JB): 6076.666  
Skew: -0.126 Prob(JB): 0.00  
Kurtosis: 8.813 Cond. No. 5.16e+03  
==============================================================================  
  
Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
[2] The condition number is large, 5.16e+03. This might indicate that there are  
strong multicollinearity or other numerical problems.  
"""

Аналіз результатів:

* Значення , , , схожі на результати методу найменших квадратів.
* Коефіцієнти регресії приблизно рівні коефіцієнтам, що були отримані за допомогою методу найменших квадратів.
* Загалом схожість результатів МНК і УМНК вказує на стійкість моделі.
* Так як у даних була помічена гетероскедастичність, модель на основі УМНК є більш придатною для проведення аналізу.

Оцінювання регресійної моделі:

residuals = results\_ols.resid

plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)

stats.probplot(residuals, dist="norm", plot=plt)

plt.title('Q-Q Plot')

plt.subplot(1, 2, 2)

sns.scatterplot(x=results\_ols.fittedvalues, y=residuals)

plt.axhline(y=0, color='r', linestyle='--')

plt.title('Residuals vs Fitted')

plt.xlabel('Fitted values')

plt.ylabel('Residuals')

plt.tight\_layout()

plt.show()

* A graph and diagram of a graph

  Description automatically generated with medium confidenceQ-Q графік порівнює розподіл залишків з нормальним розподілом. В ідеальному випадку точки повинні знаходитися на червоній лінії.
* У нашому випадку маємо відхилення нижчих та вищих квартилів на Q-Q графіку від червоної лінії, тобто залишки розподілені за розподілом з важчими хвостами, ніж у нормальному розподілі.
* Графік порівняння залишків з підлаштованими значеннями ціни допомагає перевірити нелінійність та гетероскедастичність. В ідеальному випадку залишки повинні бути розкидані навколо червоної лінії випадковим чином без помітних закономірностей
* У нашому випадку маємо у більшості випадковий розкид, але можемо помітити невелике збільшення дисперсії залишків зі збільшенням значень ціни, що вказує на гетероскедастичність. Це може означати, що модель передбачає певні інтервали цін краще за інші, або те, що в даних відсутня змінна, що пояснює частину дисперсії.

## ВИСНОВКИ

У ході виконання лабораторної роботи була знайдена і оцінена регресійна модель для даної вибірки за допомогою різних методів оцінки коефіцієнтів регресії та проведений статистичний аналіз отриманих результатів.

Отримані результати:

* + 1. Перевірка вибірки на однорідність

За критеріями Бартлета, Кохрена, модифікованим М-критерієм Бартлета та граничним критерієм неоднорідності дисперсій Харисона і Мак-Кейба відхилили нульову гіпотезу про однорідність вибірки, тобто вибірка є неоднорідною. За параметричним критерієм Голдфільда і Квандта не відхилили нульову гіпотезу про однорідність вибірки.

* + 1. Оцінка коефіцієнтів регресії за допомогою методу найменших квадратів

|  |  |
| --- | --- |
| const | 9.46e+04 |
| bedroom\_count | 308.5671 |
| net\_sqm | 25.1994 |
| center\_distance | -3.3608 |
| metro\_distance | 6.9984 |
| floor | 121.6607 |
| age | -26.1286 |

Вплив усіх коефіцієнтів збігається з їхнім інтуїтивним сприйняттям.

* + 1. Статистичний аналіз результатів
* Значення , , , вказують на високу ефективність моделі.
* Аналіз залишків показав, що вони не є нормально розподіленими, мають незначну асиметрію, мають розподіл з важкими хвостами, не мають автокореляції.