

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«ПОБУДОВА ГЕДОНІЧНОГО ІНДЕКСУ ЦІН ДЛЯ
НОУТБУКІВ ЗА ДОПОМОГОЮ МЕТОДІВ МНОЖИННОЇ РЕГРЕСІЇ»**

Виконала: студентка 4-го року
навчання

освітньої програми «Прикладна
математика»,
спеціальності 113 Прикладна
математика

Метелюк Софія Володимирівна

Керівник: Дрінь С. С.,
кандидат фіз.-мат. наук, ст. викладач

Рецензент: Неживенко О.В.

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
« ____ » _____ 20 ____ р.

Київ – 2022

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра математики факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри математики,
проф., д.ф-м.н.

_____ Б. В. Олійник
(підпис)

«_____» _____ 202_ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

На кваліфікаційну роботу

студенту 4-го курсу бакалаврської програми, факультету інформатики
Софія Метелюк

Розробити комп'ютерну програму, яка за вхідними даними виміряє рівень
зміни якості деякого товару.

Вихідні дані:

- похибка вимірювання – 5 %;
- коефіцієнт, який показує зміну в якості товару;

Зміст ТЧ до дипломної роботи:

Зміст

Анотація

Вступ

1. Огляд існуючих теорій з цього питання;
2. Розробка структурної моделі гедонічного індексу цін для ноутбуків за допомогою множинної регресії;
3. Розробка алгоритму програми;
4. Програма для знаходження коефіцієнта якості товару;
5. Результати, які будуть отримані при дослідженні реальних української даних;

Висновки

Список літератури

Додатки

Дата видачі «08» листопада 2021 р. Керівник Дрінь С.С. _____
(підпис)

Завдання отримала Метелюк С.В. _____
(підпис)

Тема: Побудова гедонічного індексу цін для ноутбуків за допомогою методів множинної регресії

Календарний план виконання роботи:

№ п/п	Назва етапу кваліфікаційної роботи	Термін виконання етапу	Примітка
1.	Отримання завдання на кваліфікаційну роботу.	08.11.2021	
2.	Огляд технічної літератури за темою роботи.	16.11.2021	
3.	Виконати аналіз сучасних методів	22.11.2021	
3.	Розробка моделі	14.12.2021	
4.	Розробка алгоритму	18.01.2022	
5.	Програмування розробленого алгоритму	09.02.2022	
6.	Виконання порівняльного аналізу результатів, для різних років	12.04.2022	
7.	Написання пояснювальної роботи.	18.04.2022	
8.	Створення слайдів для доповіді та написання доповіді.	24.04.2022	
8.	Аналіз отриманих результатів з керівником, написання доповіді та попередній захист кваліфікаційної роботи.	30.04.2022	
10.	Корегування роботи за результатами попереднього захисту.	12.05.2022	
11.	Остаточне оформлення пояснювальної роботи та слайдів.	15.06.2022	
12.	Захист кваліфікаційної роботи (проекту)	05.07.2022	

Студент: Метелюк С.В.

Керівник: Дрінь С.С.

“ ”

Зміст

Перелік прийнятих скорочень.....	5
Анотація	6
Вступ.....	7
1. Огляд існуючих теорій з цього питання	8
1.1. Розвиток гедонічної теорії	8
1.2. Гедонічне рівняння	11
1.2.1. Основні означення.....	11
1.2.2. Основні означення для аналізу багатофакторної лінійної регресії 13	
1.3. Проблематика оцінювання гедонічного індексу	20
1.3.1. Проблема гетероскедастичності	20
1.3.2. Вибір функціональної форми.....	21
1.3.3. Вибір пояснюючих змінних, ефекти від їх включення чи неправомірного виключення із моделі регресії	23
2. Розробка структурної моделі гедонічного індексу цін для ноутбуків за допомогою множинної регресії	24
2.1. Дані.....	26
3. Дослідження реальних українських даних.....	31
Висновки	38
Використана література.....	40

Перелік прийнятих скорочень

DV- dummy variables – фіктивні змінні;

R^2 – коефіцієнт детермінації;

BLUE-оцінки -Best Linear Unbiased Estimators, найкращі із усіх оцінок в класі лінійно незсунутих оцінок, що отримані при застосуванні МНК;

МНК – метод найменших квадратів;

УМНК – узагальнений метод найменших квадратів;

GPU – graphics processing units – графічний процесор;

RAM - Random-access memory - оперативна пам'ять;

OS – operating system – операційна система;

Анотація

В цій роботі досліджено побудову гедонічних індексів для ноутбуків для українського ринку, знайдено гедонічна ціна для характеристик за допомогою множинної регресії, запропоновано рішення для розв'язання проблеми гетероскедастичні, вибору функціональної форми. Використано трансформацію Бокса-Кокса, звичайний метод найменших квадратів (МНК), узагальнений метод найменших квадратів (УМНК). Порівняно дві специфікації гедонічної регресії та отримано оцінені параметри для кожної із запропонованих. Результатом роботи є отримання напівлогарифмічної моделі із оціненими гедонічними цінами для ноутбуків на основі реальних українських даних.

Ключові слова: гедонічні індекси, гедонічне рівняння, трансформація Бокса-Кокса, УМНК.

Вступ

В будь-який час цікаво знати ціни змінились через інфляцію чи наявного покращення якості. Для товарів в яких фізичні якості майже не змінюються протягом довгого проміжку часу, наприклад, для сирих матеріалів та напівфабрикатів, зміна якості не є актуальною. Та для продуктів у яких зміна якості відбувається швидко, точне виявлення якості викликає проблему.

Власне саме така проблема і зумовила мету нашої роботи, яка полягає у тому, щоб створити набір інструментів, які дозволять спрогнозувати ціну при зміні в якості товару так, щоб це було відокремлено від інфляції.

У першому розділі описано існуючі роботи з теми, описано розвиток гедонічної теорії, основні означення із теми та необхідні визначення для проведення другого розділу роботи. Також описано головні проблеми, які виникають при дослідженні реальних даних та зроблено припущення на основі оглянутої літератури, якими методами можна вирішити наявну проблему. У другому розділі описано зібрані дані, запропоновано специфікації моделей для аналізу та показано результати роботи.

Слід зазначити, що літератури на дану тематику є чимало, і серед неї у нашій роботі варто було б особливо виділити [1], [2],[3]-[12].

1. Огляд існуючих теорій з цього питання

1.1. Розвиток гедонічної теорії

Матеріал цього й наступного підрозділів має реферативний характер.

Використовуються теоретичні положення з робіт [1], [2] та [3].

Прийнято вважати, що першим, хто використав математичну статистику для опису співвідношення якості до ціни є Фредерік Вог (Frederick V. Waugh) у 1927-1928[5]. Його цікавило як якість овочів впливає на їх ціну. За допомогою множинного регресійного аналізу Вог досліджував вплив фізичних характеристик певних овочів на ціну. Метою його дослідження було знаходження вагомих факторів, що впливають на низьку або високу ціну.

Вог побудував регресію від фізичних характеристик овочів, які були пов'язані на його думку з реальними і видимими відмінностями в якості, такими як розмір, форма, колір, степінь зрілості і т.д.

Його дослідження зосереджувалося на змінах співвідношення ціна-якість в конкретних момент часу, проте все частіше потрібно оцінити зміну ціна-якість в часі, наприклад, через технічний прогрес. Для цього у рівняння можна ввести фіктивні змінні. Такий тип регресійного аналізу відомий як гедонічний аналіз цін.

У 1938 р. Економіст Е. Корт (Andrew T. Court) проводить дослідження метою якого є оцінка впливу зміни цін на автомобілі на об'єми продажів. Через ряд проблем Карту потрібно було знайти альтернативну процедуру для вибору пояснювальної змінної в рівнянні попиту. Так Корт прийшов до «гедонічного мислення», метою якого є досягнення найбільшого рівня задоволеності суспільства в цілому. Тоді Корт ввів визначення терміну «гедонічне порівняння» як такого що визнає «потенційний внесок будь-якого товару в благополуччя та ступінь задоволення покупця та всього суспільства»[1].

Корт застосував гедонічний аналіз і описав спосіб трактувати коефіцієнти при характеристиках та при фіктивних змінних, що визначали час. Також було застосовано логарифмічну специфікацію.

Наступним дослідженням, яке можна виділити було зроблене Грегорі Чоу у 1967 р. (Gregory Chow). Цілю дослідження було пояснення росту попиту на комп'ютерні послуги в США в період 1955-1965 рр., якщо точніше, то відділення один від іншого двох ефектів: зміна попиту, яка відбулася б незалежно від технічного прогресу, і зміна попиту, яка викликана технічним прогресом (покращення якості). Для цього Чоу необхідно було побудувати індекс цін на комп'ютерні послуги за допомогою гедонічного аналізу цін і методів множинної регресії. Чоу використав кілька методів побудови індексів зі поправкою на якість, а також вивів метод для побудову середньорічного темпу росту цін на комп'ютери на основі отриманих результатів [1].

У цей же час до гедонічних індексів починає звертатися все більше вчених та економістів, кількість наукових праць на цю тему різко зростає. Важливо сказати про роботу Цві Грілішес. Поняття гедонічної гіпотези було сформульоване Макото Охта, Цві Грілішес (Makoto Ohta, Zvi Griliches) у 1967 році так: «гедонічна гіпотеза припускає, що товар можна розглядати як зібрання різних характеристик чи атрибутів, для яких імпліцитні (неявні) ціни можуть бути отримані з цін на різні версії одного товару, що містить різні рівні специфічних характеристик. Можливість розбити товар на характеристики (disaggregate) та визначити ціну його компонентів полегшує побудову цінових індексів та вимірювання зміни ціни по різних версія того самого товару.»[3]

Також не можна пропустити роботи Дж. Триплета (Jack E. Triplett). Він ввів визначення, які нині вважаються загальноприйнятими у теорії гедонічних індексів.

«Термін «гедонічний метод» включає будь-яке використання в економічних вимірюваннях (дослідженнях) «гедонічної функції»,

$$P = h(c)$$

Де P – ціна – кожне P_{ijt} де для кожної j -тої «варіації» чи «моделі» i -го сервісу чи товару (наприклад, ціна на різні моделі автомобілів) доступних у час t . Матриця c , яка має по одному рядку, що складається із характеристик для кожної з моделей» [4]

Варто зазначити, що є ще один відомий метод, окрім гедонічного, який може побудувати індекс цін, із врахуваннями зміни якості.

Метод відповідних моделей - “Matched model” method вважається традиційним способом побудови індексу цін, у перекладі метод відповідних(узгоджених) моделей. При конструюванні індексу використовуються дані про ціни тільки тих моделей виробів та їх різновидів, характеристики якості яких не змінилися протягом двох сумісних періодів.

Ідея цього методу полягає в тому, що узгодженість моделей виробів гарантує, що будь-яка різниця в ціні, що з’явиться протягом двох проміжків часу, буде відображати лише зміни самих цін, а не зміни, пов’язані із тим, що в дійсності було куплено. Наприклад, в США індекси цін виробників (Producer Price Indexes, PPI), які використовуються при дефляції багатьох компонентів виробничого устаткування довготривалого використання, будуються на основі методу відповідних моделей.

Проблема методу, які перешкоджають правильному обчисленню зміни якості:

- 1) Може з’явитися, коли зміна цін, за якими спостерігають для відповідних моделей виробів неточно відображають коливання цін, які властиві всім моделям виробів.

2) Може з'явитися, якщо моделі виробів, які не є насправді ідентичними, спостерігаються як відповідні. Вона з'являється, коли інформація про деякі специфікації і характеристики виробів є недоступною або не береться в обчислення. Тобто здійснюється приведення до відповідності товарів насправді по суті різних.

Вирішення однієї проблеми, тобто так званий компроміс, завжди буде приводити до іншої. Проте регресійний аналіз здатний значно пом'якшити вплив такого протиріччя.[1]

Впливовим аргументом для розвитку гедонічного методу замість методу відповідних моделей став той факт, що для обчислення гедонічного індексу нам потрібно мати дані про ціну та характеристики групи товарів і якщо із часом, якась певна модель із цієї групи вийде із ринку і буде заміненена на іншу, ми все одно зможемо коректно застосувати гедонічний метод. У сучасному світі, де виробники техніки змушені постійно змінювати комплектацію і випускати на ринок нові товари, а відповідно прибирати із продажу старі, саме гедонічний метод має більше перспектив.

Однак варто зауважити, що є товари або випадки, у яких результати гедонічного методу та методу відповідних моделей може давати однакові результати[6].

1.2. Гедонічне рівняння

1.2.1. Основні означення

В основі гедонічної гіпотези лежить припущення, що кожен товар характеризується набором всіх його характеристик. Кожен товар x можна подати як набір його характеристик $X = (x_1, \dots, x_K)$. [7]

Висувається припущення, що для кожного товару є таке співвідношення між функції f від його характеристик та ціною товару.

$$p = f(X)$$

Ця функція визначає гедонічну регресію або гедонічне відношення.

Важливо зазначити про концепцію імпліцитної (неявної) ціни або гедонічної ціни. Ці ціни визначення частковими похідними від гедонічної функції

$$\frac{\partial p}{\partial x_k}(x) = \frac{\partial f}{\partial x_k}(x) \quad (k=1, \dots, K)$$

Гедонічна ціна $\frac{\partial f}{\partial x_k}(x)$ визначає як сильно ціна p товару зміниться якщо цьому товару покращити або додати більше характеристики x_k за умови не змінності всіх інших характеристик.[7]

Гедонічну регресію може приймати різні функціональні форми, нижче подано найчастіше використовувані:

Лінійна $p = \beta_0 + \sum \beta_k x_k$, де гедонічна ціна β_k

Експоненційна (exponential) $p = \beta_0 \prod e^{\beta_k x_k} \Rightarrow \ln p = \ln \beta_0 + \sum \beta_k x_k$ де гедонічна ціна $\beta_k p$

Степенева (power function) $p = \beta_0 \prod x_k^{\beta_k} \Rightarrow \ln p = \ln \beta_0 + \sum \beta_k \ln x_k$
де гедонічна ціна $\beta_k (p/x_k)$

Логарифмічна $p = \beta_0 + \sum \beta_k \ln x_k$ де гедонічна ціна β_k/x_k . [7]

У науковій літературі часто можна зустріти іншу нотацію для гедонічної регресії, її основна різниця полягає в тому, що автор одразу подає фіктивні змінні(DV) [1][6].

Типова гедонічна регресія описує ціни кожної варіації або моделі товару, що вироблений або продається у момент часу t ($P_{m,t}, m = 1, \dots, M$) як функцію від характеристик (as a function of the quantities of its characteristic)

($C_{k,m,t}, k = 1, \dots, K$) та фіктивних змінних для часу ($D_{m,t}, t = 1, \dots, T$).

В напівлогарифмічній специфікації це співвідношення представлено як:

$$\ln P_{m,t} = \sum_k \beta_k C_{k,m,t} + \sum_t \alpha_t D_{m,t} + \varepsilon$$

Де $D_{m,t} = \begin{cases} 1, \text{ якщо ціна моделі } m \text{ спостерігається у момент } t \\ 0, \text{ у зворотньому випадку} \end{cases}$

α_t та β_k – оцінювані параметри, ε – незалежна однаково розподілена нормальна похибка. Кожна модель має K характеристик, які впливають на вартість. Для характеристик типово обираються кількісні змінні (наприклад, для комп'ютерів це буде кількість пам'яті, швидкість процесору, вага), проте можна вводити фіктивні змінні для показання впливу або навпаки відсутності впливу певних атрибутів в конкретній моделі (наприклад, для додаткової кеш пам'яті).

$\sum_k \beta_k C_{k,m,t}$ відповідають за різницю в якостях кожної моделі, і відповідно тоді весь інший вплив на ціну здійснюється фіктивними змінними для часу та похибкою.

Гедонічна регресія – це особливе застосування лінійної множинної (багатофакторної) регресії, далі подано основні теоретичні відомості про множинну регресію, на які спиратиметься наступний аналіз.

1.2.2. Основні означення для аналізу багатофакторної лінійної регресії

Матеріал цього й наступного підрозділів має реферативний характер.

Використовуються теоретичні положення з робіт [1], [12].

Узагальнена багатофакторна лінійна регресія буде мати вигляд:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon,$$

y – залежна змінна;

$x_1, x_2, \dots, x_p$ – незалежні змінні або фактори;

$\beta_0, \beta_1, \dots, \beta_p$ – параметри моделі, які потрібно оцінити;

ε – неспостережувана випадкова величина;

Для класичної багатофакторної регресійної моделі, яка є узагальненням простої регресійної моделі, всі основні припущення зберігаються, проте із модифікацією. Ці припущення становлять основу класичного регресійного аналізу, вони походять із теореми (умов) Гауса-Маркова.

Спершу подамо Умови Гауса-Маркова:

Щоб регресійний аналіз на основі звичайного методу найменших квадратів давав найкращі із усіх оцінок $b_0, b_1, \dots, b_n$ модель повинна відповідати наступним вимогам:

- 1) Математичне сподівання випадкової помилки в усіх спостереженнях має дорівнювати нулю:

$$E(\varepsilon_i | x_{1i}, x_{2i}, \dots, x_{pi}) = 0 \text{ для кожного } i$$

- 2) Дисперсія помилки постійна для всіх спостережень:

$$\text{var}(\varepsilon_i) = \sigma^2 = \text{const}$$

- 3) Відсутність автокореляції похибок: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0; i \neq j$

- 4) Випадкова величина (помилка) має бути розподілена незалежно від пояснюваних змінних.

Основні припущення у багатофакторному регресійному аналізі за формулюванням [12]:

Щоб регресійний аналіз на основі звичайного методу найменших квадратів давав найкращі із усіх оцінок в класі лінійно незсунутих оцінок (Best Linear Unbiased Estimators, BLUE-оцінки), модель повинна відповідати наступним вимогам:

Припущення 1: Математичне сподівання випадкової величини ε дорівнює 0.

$$E(\varepsilon_i | x_{1i}, x_{2i}, \dots, x_{pi}) = 0 \text{ для кожного } i$$

Припущення 2: Випадкові величини незалежні між собою.

$$cov(\varepsilon_i, \varepsilon_j) = 0; \quad i \neq j$$

Припущення 3: Модель гомоскедастична, тобто має однакову дисперсію для будь-якого спостереження:

$$var(\varepsilon_i) = \sigma^2 = const$$

Припущення 4: Коваріація між випадковою величиною ε_i та кожною незалежною змінною дорівнює 0.

Припущення 5: Модель повинна бути правильно специфікована.

Припущення 6: Випадкова величина ε_i відповідає нормальному закону розподілу з нульовим математичним сподіванням і дисперсією.

Припущення 7: Відсутність мультиколінеарності між фактором x , тобто фактори повинні бути незалежними між собою. Іншими словами, не повинно бути точного лінійного зв'язку між двома або більше факторами.

Математично відсутність колінеарності між двома факторами x_1 та x_2 визначають такими чином, що не існує чисел γ_1 γ_2 , які одночасно не дорівнюють 0, для яких би виконувалася тотожність:

$$\gamma_1 x_{1i} + \gamma_2 x_{2i} = 0, \quad i = \overline{1, n}$$

Якщо тотожність виконується тільки тоді, коли $\gamma_1 = \gamma_2 = 0$, то x_1 та x_2 лінійно незалежні або неколінеарні.

1.2.2.1. Коефіцієнт детермінації

Міра ступеню відповідності даних, отриманих з регресійної моделі, фактичним даним є коефіцієнт множинної кореляції, який визначається як коефіцієнт кореляції між y та $\hat{y}$.

$$R = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}$$

R^2 - Квадрат коефіцієнту множинної кореляції, як і у випадку простої регресії, називають коефіцієнтом детермінації.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Для порівняння значень коефіцієнтів детермінації у різних моделях треба коригувати коефіцієнт кореляції з урахуванням кількості факторів x , які входять у різні моделі, тобто зменшити вплив залежності значення коефіцієнта детермінації від кількості факторів. Для цього вводиться спеціальний оцінений коефіцієнт детермінації (Adjusted R^2), який має вигляд:

$$Adj. R^2 = \bar{R}^2 = 1 - \frac{\frac{\sum_{i=1}^n e_i^2}{n - k}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(n - 1)}}$$

k - кількість параметрів регресійної моделі, включаючи перетин. Оцінений коефіцієнт детермінації коригується з урахуванням ступенів вільності суми квадратів залишків та загальної суми квадратів.

Порівняння коефіцієнтів детермінації у різних моделях має кілька особливостей. Для порівняння коефіцієнта детермінації у двох різних специфікаціях лінійній та напівлогарифмічній потрібно врахувати, що коефіцієнт детермінації у напівлогарифмічній моделі вимірює частку дисперсії $\ln y$, яку можна пояснити факторами $x_{1i}, x_{2i}, \dots, x_{pi}$, тоді як у моделі лінійній він вимірює частку дисперсії y . Це не тотожні поняття. Зміна в $\ln y$ забезпечує відносну зміну y , тоді як зміна в самому y є абсолютною зміною.

Таким чином $\frac{var(\hat{y}_i)}{var(y_i)} \neq \frac{var(\ln \hat{y}_i)}{var(\ln y_i)}$. Для того, щоб порівняти коефіцієнти

детермінації регресійних моделей необхідно, знайти $\ln \hat{y}_i$ у

напівлогарифмічній специфікації, розрахувати коефіцієнт детермінації між антилогарифмом та значенням y_i . Знайдений таким чином коефіцієнт

детермінації можна порівнювати з коефіцієнтом детермінації лінійної регресійної моделі.

Другий спосіб (зворотній): обчислити для напівлогарифмічної моделі $\ln \hat{y}_i$, розрахувати $\hat{y}_i$, обчислити R^2 за формулою

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

а його в свою чергу можна порівняти з R^2 знайденим для лінійної моделі.

Для порівняння моделей у цій робота також буде використано SSE.

Сума квадратів помилок SSE (Sum of Squared Errors)

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Цей критерій особливо поширений при виборі оптимальних моделей прогнозування. З декількох моделей обирається, так, яка дає найменшу суму квадратів помилок.

Середні квадратом називається сума квадратів, поділена на відповідний їй ступінь вільності. Середнім квадратом помилок (MSE) називається сума квадратів помилок, поділена на відповідний ступінь вільності.

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

$$RSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

1.2.2.2. Перевірка адекватності багатофакторної регресійної моделі

Для здійснення перевірки на адекватність багатфакторної регресійної моделі використовується F-критерій Фішера. При цьому нуль-гіпотеза узагальнюється і має вигляд:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

Проти альтернативної гіпотези H_1 : хоча б одне значення β_i відмінне від нуля. Якщо нуль-гіпотеза H_0 не правильна, то тоді правильна гіпотеза H_1 , що не всі параметри незначною мірою відрізняються від нуля, що дає підставу вважати, що відібрані фактори пояснюють зміну залежності величини y . Для перевірки H_0 -гіпотези розраховується F-статистика Фішера з p та $(n-p-1)$ ступенями вільності.

$$F_{p,n-p-1} = \frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - p - 1}}$$

p - кількість факторів, які увійшли у модель, n – загальна кількість спостережень; За F-таблицями Фішера шукаємо критичне значення $F_{кр}$ з p та $(n - p - 1)$ ступенями вільності, задавши попередньо рівень помилки $\alpha \cdot 100\%$, або рівень довіри $(1 - \alpha) \cdot 100\%$.

Якщо $F > F_{кр}$, то нуль гіпотеза відкидається, що свідчить про адекватність побудованої моделі. Навпаки, вони приймається і модель вважається неадекватною.

1.2.2.3. Узагальнений метод найменших квадратів (УМНК)

Узагальнений метод найменших квадратів враховує інформацію про неоднаковість дисперсії і через це здатний створити BLUE-оцінки.

Далі наводяться формули для випадку простої лінійної регресії у вигляді $y = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Перепишемо у вигляді $y = \beta_0 x_{0i} + \beta_1 x_1 + \varepsilon_i$, де $x_{0i} = 1$ для кожного i .

Припустимо, що наявна гетероскедастичність, та всі дисперсії σ_i^2 відомі.

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{x_{0i}}{\sigma_i} + \beta_1 \frac{x_1}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}, \text{ якщо } \frac{y_i}{\sigma_i} \rightarrow y_i^*, \text{ то } y_i^* = \beta_0^* x_{0i}^* + \beta_1^* x_1^* + \varepsilon_i^*,$$

Позначення β_0^*, β_1^* використані, що відрізняти від звичайних β_0, β_1 , отриманих методом найменших квадратів. Завдяки трансформації, дисперсія трансформованої помилки $\varepsilon^* = const$ є сталою величиною, отже зберігається припущення про гомоскедастичність із умов Гауса-Маркова, що дозволяє перейти до класичної регресії і застосувати метод найменших квадратів, який тепер буде називатися узагальненим методом найменших квадратів, бо, дійсно, $var(\varepsilon_i^*) = E(\varepsilon_i^*)^2 = E\left(\frac{\varepsilon_i}{\sigma_i}\right)^2 = \frac{1}{\sigma_i^2} E(\varepsilon_i^2)$.

Припускаємо, що σ_i^2 відомі, тоді

$$\frac{1}{\sigma_i^2} E(\varepsilon_i^2) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1$$

Тепер оцінки параметрів будуть BLUE-оцінками, відомими як УНК-оцінками. Щоб знайти невідомі параметри за методом узагальнених найменших квадратів для $y_i^* = \beta_0^* x_{0i}^* + \beta_1^* x_1^* + \varepsilon_i^*$, треба мінімізувати

$$\sum_{i=1}^n e_i^* = (y_i^* - b_0^* x_{0i}^* + b_1^* x_1^*)^2 \rightarrow \min \text{ або}$$

$$\sum_{i=1}^n \left(\frac{e_i}{\sigma_i}\right)^2 = \left[\left(\frac{y_i}{\sigma_i}\right) - b_0^* \left(\frac{x_{0i}}{\sigma_i}\right) - b_1^* \left(\frac{x_1}{\sigma_i}\right)\right]^2 \rightarrow \min$$

І тоді можна переписати в наступне, де $\gamma_i = \frac{1}{\sigma_i^2}$

$$\sum_{i=1}^n \gamma_i e_i^2 = \sum_{i=1}^n \gamma_i (y_i - b_0^* - b_1^* x_i)$$

Тобто в узагальненому методі найменших квадратів мінімізуємо зважену суму квадратів відхилень з вагами, обернено пропорційними до σ_i . У

літературі останній вираз називають «зважені найменші квадрати (ЗНК)», а оцінки параметрів, отримані за цим методом, - ЗНК-оцінки. У контексті гетероскедастичності можна користуватися обома термінами: і ЗНК, і УНК. Загалом, ЗНК є особливим випадком методу узагальнених найменших квадратів.

УНК-оцінка для β_1^* буде

$$b_1^* = \frac{(\sum_{i=1}^n \gamma_i)(\sum_{i=1}^n \gamma_i x_i y_i) - (\sum_{i=1}^n \gamma_i x_i)(\sum_{i=1}^n \gamma_i y_i)}{(\sum_{i=1}^n \gamma_i)(\sum_{i=1}^n \gamma_i x_i^2) - (\sum_{i=1}^n \gamma_i x_i)^2}$$

І дисперсія параметра буде дорівнювати:

$$var(b_1^*) = \frac{(\sum_{i=1}^n \gamma_i)}{(\sum_{i=1}^n \gamma_i)(\sum_{i=1}^n \gamma_i x_i^2) - (\sum_{i=1}^n \gamma_i x_i)^2}$$

1.3. Проблематика оцінювання гедонічного індексу

1.3.1. Проблема гетероскедастичності

Припущення 3 про сталість дисперсії кожної випадкової величини ε_i , тобто гомоскедастичність, для багатofакторної моделі можна записати так:

$$var(\varepsilon_i) = E\{(\varepsilon_i - E(\varepsilon_i))^2\} = \sigma_\varepsilon^2 = const$$

Якщо це припущення не задовольняється у якомусь окремому випадку, то має місце гетероскедастичність:

$$var(\varepsilon_i) = \sigma_{\varepsilon_i}^2 \neq const$$

Суть припущення гомоскедастичності полягає в тому, що варіація кожної ε_i навколо її математичного сподівання не залежить від значення x . Дисперсія кожної ε_i зберігається сталою незалежно від малих чи великих значень факторів: σ_ε^2 не є функцією x_{ij} , тобто $\sigma_\varepsilon^2 \neq f(x_{1i}, x_{2i}, \dots, x_{pi})$

Гетероскедастичність випадкових помилок моделі – це порушення припущення про сталість дисперсії випадкових помилок моделі, дисперсія залишків моделі не є сталою. Якщо застосувати метод найменших квадратів (МНК\ OLS), то отримані МНК-оцінки параметрів моделі будуть неефективними, не матимуть найменшу дисперсію серед незсунутих оцінок, як це має бути за умови гомоскедастичності.

Для того, щоб перевірити наявність гетероскедастичності, можна скористуватися наступними тестами. Вибір тесту залежить від припущень про появу гетероскедастичності. Відомими тестами є тест Бреуша-Пагана (Breusch-Pagan Test), тест Глейзера (Gleizer test), тест або критерій Вайта (White test), коефіцієнт кореляції рангу Спірмена (Spearman's rank correlation coefficient).

Під час застосування будь-якого з цих тестів перевірці підлягає основна гіпотеза $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_n^2$ проти альтернативної $H_1: \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \dots \neq \sigma_n^2$, де σ^2 - дисперсія [10]

Для того щоб оцінки параметрів були ефективними, а оцінки стандартних помилок були вірними необхідно замість звичайного МНК використовувати узагальнений МНК (УМНК\GLS – generalized least squares), якщо є автокореляція залишків, або зважений МНК, якщо помилки незалежні, але не однаково розподілені.[1]

1.3.2. Вибір функціональної форми

Вище було подано кілька функціональних форм для гедонічної регресії. Вибір функціональної форми може бути складністю при побудові гедонічної моделі. Вог брав до уваги тільки лінійну форму, Корт порівняв напів логарифмічну з лінійною та визначив, що із автомобільними даними напівлогарифмічна форма є більш підходяща.

Коул та колеги порівняли результати основані на лінійній, напівлінійній та log-log специфікаціях і виявили, що в загальному випадку log-log специфікація є підходящою.

Є певна статистична основа, яка може бути використана, щоб порівняти альтернативні функціональні форми і не лише log-log, напівлогарифмічну та лінійну специфікацію. Цей метод відомий як трансформація Бокс-Кокс або Бокса-Тідвела (Box-Cox or Box-Tidwell transformation).

Трансформація Бокса-Кокса включає трансформацію змінної від X_i до X_i^* , Y_i до Y_i^* наступним чином

$$X_i^* = \frac{X_i^\lambda - 1}{\lambda}, \quad Y_i^* = \frac{Y_i^\lambda - 1}{\lambda}$$

де λ це оцінюваний параметр. Якщо $\lambda = 1$ для всіх змінних, то функціональна форма є лінійною. Якщо $\lambda \rightarrow 0$ для кожної змінної, то $X_i^* \rightarrow \ln X_i$, $Y_i^* \rightarrow \ln Y_i$, що призводить до специфікації log-log.

В процедурі Бокса-Тідвела для окремих X можна застосовувати різні λ . [1]

При використанні трансформації Бокса-Кокса необхідно, щоб всі значення вхідної послідовності були додатніми та відмінними від нуля, якщо це не виконується, то послідовність можна зсунути в додатню область на величину, яка гарантує «додатність» всіх її значень.

Трансформація Бокса-Кокса – це нелінійна трансформація змінної, яка дозволяє вибрати між двома функціональними формами моделі lin-lin та log-lin. Формула для трансформації має вигляд:

$$x^{(\lambda)} = g(x, \lambda) = \frac{x^\lambda - 1}{\lambda}$$

Найчастіше значення параметру варіюється наступним чином $-5 \leq \lambda \leq 5$.

Для того, щоб в результаті трансформації Бокса-Кокса закон розподілу результуючої послідовності був максимально приближеним до нормального закону розподілу, необхідно вибрати оптимальне значення параметру λ .

Одним із способів визначення оптимального значення λ є максимізація логарифму функції правдоподібності:

$$f(x, \lambda) = -\frac{N}{2} \ln \left[\sum_{i=0}^{N-1} \frac{(x_i(\lambda) - \tilde{x}_i(\lambda))^2}{N} \right] + (\lambda - 1) \sum_{i=0}^{N-1} \ln(x_i), \quad \text{де } \tilde{x}_i(\lambda) = \frac{1}{N} \sum_{i=0}^{N-1} x_i(\lambda)$$

Тобто необхідно вибрати таке значення параметру λ лямбда, при якому дана функція набуває максимального значення.

Якщо $\lambda = 1$, то $g(x, 1) = \frac{x^1 - 1}{1} = x - 1$

Якщо $\lambda = -1$, то $g(x, 1) = \frac{x^{-1} - 1}{-1} = -\frac{1}{x} - 1$

Особливої уваги заслуговує випадок $\lambda = 0$, тоді ліміт функції

$$\lim_{\lambda \rightarrow 0} g(x, \lambda) = \frac{\lim_{\lambda \rightarrow 0} (x^\lambda - 1)}{\lim_{\lambda \rightarrow 0} \lambda} = \frac{0}{0}$$

Тоді

$$\lim_{\lambda \rightarrow 0} g(x, \lambda) = \frac{\lim_{\lambda \rightarrow 0} (x^\lambda - 1)}{\lim_{\lambda \rightarrow 0} \lambda} = \lim_{\lambda \rightarrow 0} (x^\lambda \log(x)) = \log(x). [11]$$

1.3.3. Вибір пояснюючих змінних, ефекти від їх включення чи неправомірного виключення із моделі регресії

Неможливо специфікувати всі змінні, які відносяться до якості товару, які мають входити в гедонічне рівняння ціни. Неправомірне невключення в гедонічне рівняння суттєвих змінних (з точки зору опису якості продукту) може призвести до зміщення оцінок параметрів цього рівняння. [1]

Взагалі, часто бувають ситуації, коли пояснюючі зміни, що використовуються в гедонічному регресійному рівнянні, самі по собі не є мірою якості, але вважаються сильно корельований з тим, що переважає в якості для споживача. Наприклад, в дослідженні по автомобілях Курта, вага автомобіля використовувалася як міра якості. Споживачі не судять про якість по вазі, але тим не менше в дійсності вага часто сильно корелює з такими показниками якості, як безпека і потужність.[1]

Тому при гедонічному аналізі необхідно приділяти окрему увагу всім обраним пояснюючим змінним і шукати всі характеристики, які впливають на якість.

2. Розробка структурної моделі гедонічного індексу цін для ноутбуків за допомогою множинної регресії

Для дослідження було вибрано ноутбуки, адже вони добре представлені на українському ринку, вони мають багато різних характеристик та досить часто випускаються нові моделі, що унеможлиблює використання методу відповідних моделей. В різних країнах у побудові Індексу Продуктових Цін (Producer Price Indexes (PPI)) використовується гедонічна регресія в сполученні із традиційним методом відповідних моделей при побудові індексу для персональних комп'ютерів та ноутбуків. Персональні комп'ютери розвиваються ще із 60-70 рр. 20 століття і про них вже є багато досліджень, як наприклад, робота Грегорі Чоу. А от ноутбуки ввійшли на ринок у повній силі за останні два десятиліття. Використання гедонічної регресії для побудови цінових індексів виправдано тоді, коли товар має дуже високий рівень зміни якості, тобто там, де метод відповідних моделей неможливо застосувати коректно, адже на двох проміжках часу не завжди можна знайти нову відповідну модель, якраз таки через швидкий технічний прогрес та постійну зміну характеристик і якості цих характеристик. Тому використання гедонічної регресії для ноутбуків є виправданим.

Часто пропонована у літературі гедонічна модель для персональних комп'ютерів та ноутбуків:

$$P_i = \beta_0 + \sum_j \beta_j x_{ij} + \sum_k \alpha_k D_{ik} + \varepsilon_i$$

P_i ціна і-го ноутбуку;

β_0 - константа;

β_j - регресійний коефіцієнт для кількісної характеристики j ;

α_k - регресійний коефіцієнт для якісної характеристики k або для часу;

D_{ik} - фіктивна змінна – Dummy variable – набуває значення 1 або 0;

ε_i - похибка;

Проте як зазначає Берндт «порівняння напряму якості моделей з різними функціональними формами немає можливості, так як зміна логарифму ціни не те саме, що зміна самої ціни», тому спробуємо дослідити і наступну модель:

$$\ln P_i = \beta_0 + \sum_j \beta_j x_{ij} + \sum_k \alpha_k D_{ik} + \varepsilon_i$$

На ринку ноутбуків є певне розділення на ноутбуки для ігор «геймінгові», ноутбуки середньої категорії, ультра тонкі «іміджеві» ноутбуки.

І хоча легко виділити ще інші категорії, зупинимось на цих, адже між ними спостерігається найбільша різниця в характеристиках і наявності певних характеристик взагалі.

Графічні карти. Особливою проблемою при побудові індексів є графічні карти. Їх є багато варіацій і поколінь, дешевші і дорожчі, на ринку постійно з'являються нові, тож при побудові моделі потрібно обов'язково врахувати їх вплив. Для передбачування ціни це не ефективно, адже нові моделі постійно замінюються старі і побудова моделі на основі даних на сьогодні, не буде

ефективна для передбачення навіть на найближчий час. Рішенням для цього може бути розділення ринку графічних карт на кілька великих категорій за схожістю і при появі нової моделі відносити до однієї із груп і відповідно мати фіктивну зміну для категорії, а не для моделі, що має приблизити результати до реального життя. [8]

2.1. Дані

Дані були зібрані для аналізу у два моменти часу: перший збір на початку 2021 року, другий збір у червні 2022 року. Датасет був зібраний за допомогою технології веб-скрапінгу із українських сайтів електронної комерції: rozetka.com.ua, moyo.ua, itbox.ua. Для збору, очищення та підготовки даних використовувалися дві мови програмування R та Python.

Веб-скрапінг – це збірна назва для методів автоматизації збирання даних за допомогою мов програмування із веб-сторінок.

Зібрано дані про 4000 ноутбуки усіх цінових сегментів.

Дані, які були зібрані, їх класифікація (кількісна чи якісна (категоріальна) та рівні характеристик:

- 1) lap_name – повна назва ноутбуку, якою вона була на сайті, із якого було зроблено веб-скрапінг;
- 2) Link- посилання на товар;
- 3) screen_size - кількісна характеристика, діагональ екрану в дюймах;
- 4) screen_type - якісна характеристика, тип екрану: TN+film, IPS, WVA,OLED;
- 5) refresh_rate - якісна характеристика, частота оновлення екрану: 60,120,144, 240 Гц;
- 6) OS – якісна характеристика, операційна система: Без ОС, macOS Big Sur, Windows 10 Home, Windows 10 Pro, DOS, Linux ;

- 7) intel_generation – якісна характеристика, покоління процесору інтел:
10-оє Ice Lake, 10-оє Comet Lake, 11-оє Tiger Lake, 8-оє Whiskey Lake,
NA(отже, процесор бренду AMD);
- 8) RAM_size – кількісна характеристика, розмір пам'яті – 8, 16 Gb;
- 9) RAM_type – якісна характеристика, тип пам'яті - DDR4-2400 МГц,
DDR4-3200 МГц, DDR4, DDR4-2666 МГц, LPDDR4X, DDR4-2933
МГц;
- 10) storage_volume – якісна характеристика, 256 Гб SSD, 512 Гб SSD,
1 Тб SSD, 2 Тб SSD, 1 Тб, 512 Гб SSD, 512 Гб SSD, 1 Тб SSD, 1 Тб +
SSD 512 Гб;
- 11) GPU_size - кількісна характристика, 2,4,6,8, 16, NA;
- 12) GPU – моделі: якісна характеристика, NVIDIA GeForce RTX
2070 Super Max-q, NVIDIA GeForce MX450, NVIDIA GeForce RTX
2070 Max-q, NVIDIA GeForce MX330, NVIDIA GeForce MX450,
NVIDIA GeForce GTX 1650 Ti Max-q, Інтегрована відеокарта, NVIDIA
GeForce GTX 1650 Ti, NVIDIA Quadro T1000 Max-q, NVIDIA Quadro
T1000, NVIDIA Quadro T1000 Max-q;
- 13) color - якісна характеристика, колір;
- 14) weight – кількісна характеристика, вага;
- 15) battery – якісна характеристика, тип батареї : знімна та незнімна;
- 16) battery_capacity- кількісна характеристика, ємність батареї;
- 17) **price** – кількісна характеристика, в гривнях;
- 18) brand – бренд;
- 19) GPU_integrated – якісна характеристика, наявність GPU – так\ні;
- 20) Cores - якісна характеристика, кількість ядер у процесорі – 2, 4, 6,
8
- 21) storage_type – якісна характеристика, 128 GB SSD, 256 GB SSD,
512 GB SSD, 1 TB SSD (основна частина із storage_volume)
- 22) OS_paid – якісна характеристика, наявність операційної системи
в комплекті – так\ні

23) RAM_type_n - якісна характеристика, скорочений тип пам'яті – DDR4, LPDDR, NA (скорочено від RAM_type)

Серед них є такі, що мають значну кореляцію. Наприклад, при збільшенні діаметру екрану збільшується вага ноутбуку, такі характеристики було прибрано із остаточного рівняння.

Після дослідження на кореляцію між ними, було відібрано наступні характеристики. Кореляція між відібраними показана на графіку:

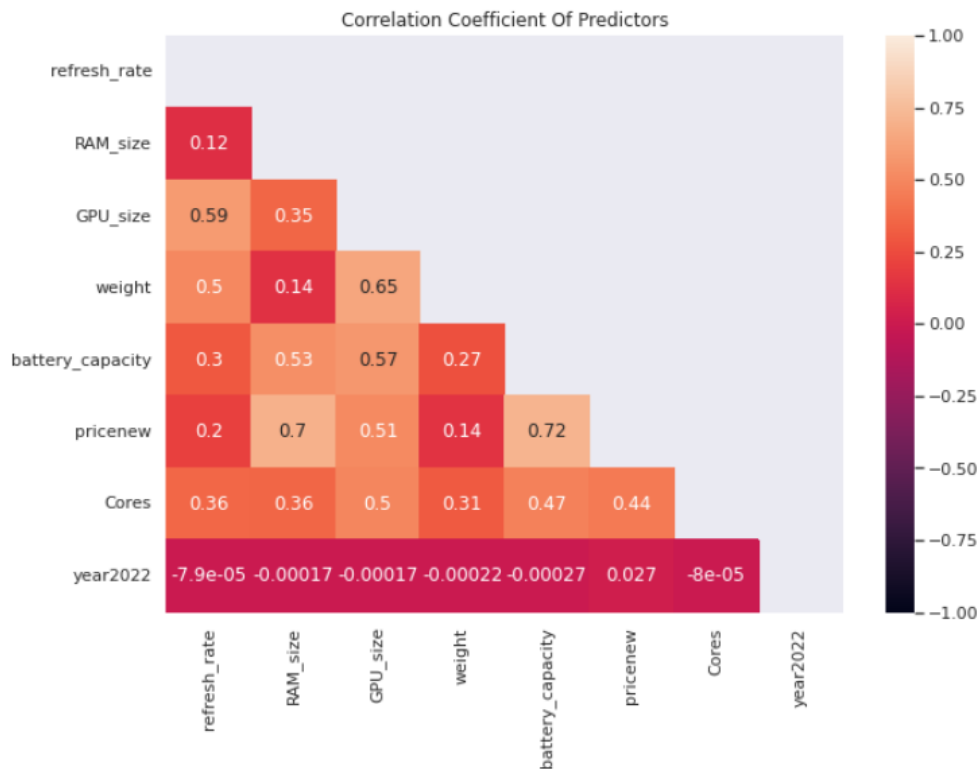


Рис.1: Графік кореляційних коефіцієнтів

В остаточну модель увійшли наступні дані:

- 1) refresh_rate,
- 2) OS;
- 3) RAM_size;
- 4) GPU_size;
- 5) weight;
- 6) battery_capacity;

- 7) **pricenew**;
- 8) **brand** ;
- 9) **storage_type**;
- 10) **year2022** – додана фіктивна змінна (DV), яка відповідає за рік, у якому було зібрано дані про *i*- тий ноутбук.

Серед них є якісні змінні (categorical або factor змінні), тобто ті, які мають обмежену кількість різних значень, які ще прийнято називати рівнями. Для того щоб ввести їх в гедонічну регресію, необхідно подати їх через фіктивні змінні.

Якісна змінна, що містить *k* значень (або рівнів) вводиться у регресійний аналіз як *k*-1 змінна, тобто набору *k*-1 фіктивних змінних (DV). Потім, регресійні коефіцієнти цих *k*-1 змінних відповідають набору лінійних гіпотез про значення змінної. Є багато схем для подання фіктивних змін і введення їх у лінійну регресію[9].

Для візуалізації рівнів зручно використовувати контрастну матрицю (contrast matrix), яка будується за принципом: якщо змінна має *k* рівні, то контрастна матриця буде мати *k*-1 колонок та *k* рядків, що заповнена 0 або 1.

Модель для *i*-го ноутбуку буде мати наступний вигляд:

$$\begin{aligned}
 P_i = & \beta_0 + \beta_1 * RAM_{size} + \beta_2 * weight + \beta_3 * batteryCapacity + \beta_4 * refreshRate \\
 & + \beta_5 * GPUsize + \sum_{i=1}^{m=8} \alpha_i DOS_i \\
 & + \sum_{i=1}^{m=12} \gamma_i DstorageType_i + \sum_{i=1}^{m=11} \varphi_i Dbrand + \beta_6 Dyear2022_i
 \end{aligned}$$

Літера D перед змінною позначає те, що це є dummy-змінні (DV), де *m*+1 кількість рівнів.

Log-lin специфікація буде мати вигляд

$$\begin{aligned}
\ln P_i = & \beta_0 + \beta_1 * RAM_{size} + \beta_2 * weight + \beta_3 * batteryCapacity + \beta_4 \\
& * refreshRate + \beta_5 * GPUsize + \sum_{i=1}^{m=8} \alpha_i DOS_i \\
& + \sum_{i=1}^{m=12} \gamma_i DstorageType_i + \sum_{i=1}^{m=11} \varphi_i Dbrand + \beta_6 Dyear2022_i
\end{aligned}$$

3. Дослідження реальних українських даних

Результати аналізу за методом найменших квадратів (табл.1):

Таблиця 1

	Лінійна специфікація	Напівлогарифмічна
R^2	0.742	0.751
Оцінений R^2 (Adj. R^2)	0.740	0.749
F-статистика	311.6	327.0
SSE	488623218287.7113	757741937.725

Трансформація Бокса-Кокса

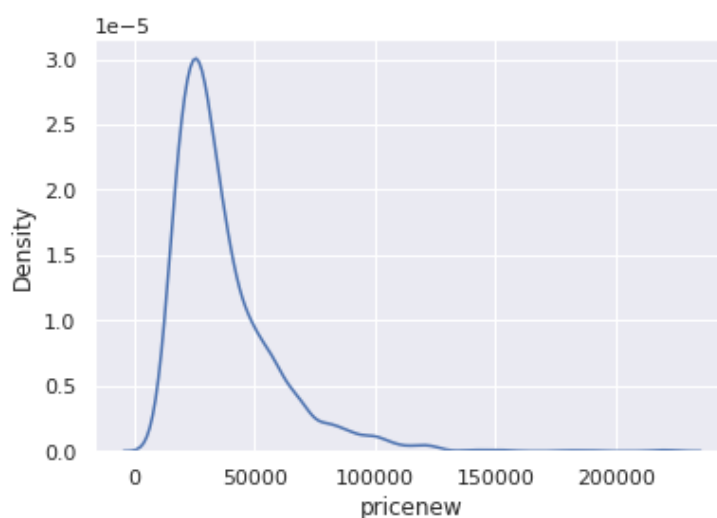


Рис.2 Графік розподілу характеристики «ціна»

Розподіл ціни не схожий на нормальний. Тому застосовується трансформація Бокса-Кокса. Для застосування однопараметричної трансформації Бокса-Кокса треба впевнитися, що послідовність, що буде підлягати зміні, є додатною та не дорівнює нулю: $\min P = 8299$, що повністю задовільняє цю умову.

Розподіл цін після трансформації Бокса-Кокса має вигляд (рис.2)

$\lambda = -0.2726$, що приводить до специфікації log-lin, як найкращої.

Проте порівняння далі буде для двох специфікацій, щоб показати всі важливі проблеми і вирішення при гедонічному аналізі.

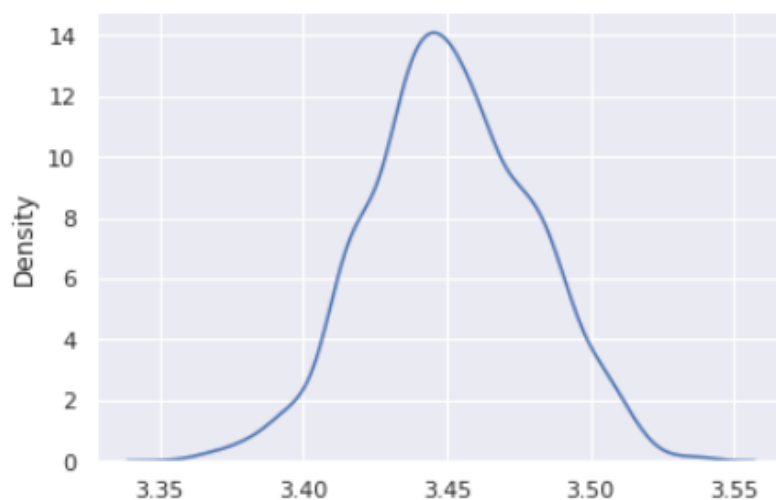


Рис.3 Графік розподілу характеристики «ціна» після трансформації Бокса-Кокса
Результати аналізу після трансформації Бокса-Кокса (табл.2):

Таблица 2

	Лінійна специфікація	Напівлогарифмічна
R^2	0.738	0.751
Оцінений R^2 (Adj. R^2)	0.735	0.748
F-статистика	303.9	302.3
SSE	75795847265.1	637732077352.6272

Із результатів аналізу видно, що напівлогарифмічна підходить краще.

Подамо оцінені параметри для всіх 4 проаналізованих моделей(табл.3):

Таблица 3

	ols(Lin- lin)	ols(log-lin)	Ols(lin(box -cox)-lin)	ols(log(box -cox)-lin)
Intercept	3745.4507	11532.218 4	3.3875	3.3878
OS[T.Endless OS]	8485.4201	1.1496	0.0068	1.0019
OS[T.Linux]	-129.3476	0.9774	-0.0015	0.9996
OS[T.Windows 10 Home]	5748.3481	1.1533	0.0083	1.0024
OS[T.Windows 10 Pro]	6071.9495	1.1673	0.0092	1.0027
OS[T.macOS Big Sur]	-1030.837	1.0095	0.001	1.0003

OS[T.macOS Catali0]	4813.2825	1.024	0.0005	1.0001
OS[T.macOS Mojave]	-10380.0	1.0549	0.0088	1.0027
OS[T.Без ОС]	-481.9075	0.9759	-0.0016	0.9995
storage_type[T.1 TB HDD 128 GB SSD]	-10730.0	0.9656	0.0022	1.0007
storage_type[T.1 TB HDD 256 GB SSD]	-5197.7954	1.2227	0.0172	1.0051
storage_type[T.1 TB HDD 512 GB SSD]	-4958.1313	1.2418	0.0181	1.0054
storage_type[T.1 TB SSD]	3192.1256	1.3760	0.0223	1.0065
storage_type[T.128 GB SSD]	-1975.7565	0.8040	-0.0164	0.9952
storage_type[T.2 TB HDD]	39700.0	2.3363	0.0479	1.0139
storage_type[T.2 TB HDD 512 GB SSD]	-8664.4017	1.0832	0.0089	1.0026
storage_type[T.2 TB HDD N TB SSD]	101.0767	0.8809	-0.0066	0.9981
storage_type[T.2 TB SSD]	12880.0	1.4132	0.0214	1.0062
storage_type[T.256 GB SSD]	-3539.4669	1.0637	0.0066	1.0020
storage_type[T.512 GB HDD]	-5748.962	0.7419	-0.0208	0.9939
storage_type[T.512 GB SSD]	-3057.1835	1.202737	0.0152	1.0045
brand[T.Apple]	21680.0	1.7388	0.0319	1.0092
brand[T.Asus]	-1689.301	0.9374	-0.0041	0.9988
brand[T.Chuwi]	-5993.3908	0.6844	-0.0265	0.9922
brand[T.Dell]	4895.9368	1.1866	0.0106	1.0031
brand[T.Dream Machines]	-1975.6091	1.0764	0.0059	1.0017
brand[T.HP]	3154.8049	1.1111	0.0064	1.0018
brand[T.Huawei]	-5390.7025	0.8640	-0.0084	0.9976
brand[T.Lenovo]	4625.8004	1.0925	0.0047	1.0014
brand[T.MSI]	-4448.7559	0.9826	0.0002	1.0068
brand[T.Razer]	-9449.4695	0.9443	-0.0016	0.9996
brand[T.Xiaomi]	-12870.0	0.7478	-0.016	0.9954
weight	-7246.5775	0.8325	-0.0105	0.997

battery_capacity	544.0149	1.014	0.0008	1.0200
RAM_size	831.3683	1.016535	0.0009	1.0003
refresh_rate	-13.696	1.0006	4.089e-05	1.0000
GPU_size	2467.6161	1.0375	0.0019	1.0005

В моделі наявна гетероскедастичність, що видно із тесту Вайта та Бреуш-Пагана, результати занесені у порівняльну таблицю (табл.4):

Таблиця 4

	Тест Вайта	Тест Бреуш-Пагана
LM-Statistic	640.2943	2353.4128
Test p-value	4.5091986462737405e-111	1.98369470074287e-13
F-Statistic	20.3719	16.6738
F-Test p-value	4.93870050129264e-122	0.00003

Тому використовуємо Узагальнений МНК :

Результати аналізу за методом УМНК (табл.5):

Таблиця 5

	Лінійна специфікація	Напівлогарифмічна
R^2	0.763	0.772
Оцінений R^2 (Adj. R^2)	0.761	0.770
F-статистика	348.5	367.0
SSE	49896263108.8215	75774199.598

Знову результати аналізу для напівлогарифмічної підходять значно краще.

Результати аналізу за узагальненим методом найменших квадратів та трансформованою залежною змінною за методом Бокса-Кокса (табл.6):

Таблиця 6

	Лінійна специфікація	Напівлогарифмічна
R^2	0.758	0.759
Оцінений R^2 (Adj. R^2)	0.753	0.757

F-статистика	334.7	336.8
SSE	7579584700.964	75802577.75

Із усіх порівняних моделей SSE саме для напівлогаримічної після трансформації Бокса-Кокса та використання УМНК замість МНК дало найменше значення, що свідчить про найоптимальнішу модель із всіх.

Оцінені коефіцієнти за УМНК до та після трансформація Бокса-Кокса (табл.7):

Таблиця 7

	gls(lin-lin)	gls(log-lin)	gls(lin(box-cox)-lin)	gls(log(box-cox)-lin)
Intercept	4230,4725	11467,8184	3,3868	3,3872
OS[T.Endless OS]	3210,6876	1,0368	0,0014	1,0004
OS[T.Linux]	-546,1112	0,9686	-0,0020	0,9994
OS[T.Windows 10 Home]	8224,0772	1,2264	0,0119	1,0034
OS[T.Windows 10 Pro]	10860,0000	1,3139	0,0163	1,0047
OS[T.macOS Big Sur]	488,9497	1,0173	0,0010	1,0003
OS[T.macOS Catali0]	8797,6326	0,8766	-0,0117	0,9965
OS[T.macOS Mojave]	-4133,2485	1,3519	0,0259	1,0077
OS[T.Без ОС]	-863,2457	0,9698	-0,0020	0,9994
storage_type[T.1 TB HDD 128 GB SSD]	-8899,2935	1,0027	0,0043	1,0013
storage_type[T.1 TB HDD 256 GB SSD]	-3202,3047	1,2872	0,0202	1,0059
storage_type[T.1 TB HDD 512 GB SSD]	-3637,1420	1,2831	0,0200	1,0059
storage_type[T.1 TB SSD]	3700,7564	1,3971	0,0232	1,0068
storage_type[T.128 GB SSD]	-2147,8652	0,7852	-0,0183	0,9946
storage_type[T.2 TB HDD]	38950,0000	2,3439	0,0483	1,0141
storage_type[T.2 TB HDD 512 GB SSD]	-6002,5990	1,1552	0,0125	1,0037

storage_type[T.2 TB HDD N TB SSD]	1277,0289	0,8809	-0,0070	0,9980
storage_type[T.2 TB SSD]	15280,000 0	1,4269	0,0215	1,0062
storage_type[T.256 GB SSD]	- 2365,9038	1,0838	0,0075	1,0022
storage_type[T.512 GB HDD]	- 2513,6270	0,8123	-0,0153	0,9955
storage_type[T.512 GB SSD]	- 2241,6172	1,2214	0,0160	1,0047
brand[T.Apple]	20750,000 0	1,7561	0,0330	1,0095
brand[T.Asus]	- 1799,2914	0,9406	-0,0038	0,9989
brand[T.Chuwi]	- 9612,6632	0,6392	-0,0301	0,9912
brand[T.Dell]	3856,2321	1,1620	0,0095	1,0028
brand[T.Dream Machines]	- 1212,0452	1,1080	0,0078	1,0023
brand[T.HP]	2379,1364	1,0966	0,0057	1,0017
brand[T.Huawei]	- 8466,0330	0,8114	-0,0118	0,9966
brand[T.Lenovo]	3729,0738	1,0630	0,0031	1,0009
brand[T.MSI]	- 4968,0247	0,9738	-0,0003	0,9999
brand[T.Razer]	- 13290,000 0	0,8778	-0,0057	0,9984
brand[T.Xiaomi]	- 14870,000 0	0,7113	-0,0189	0,9945
weight	- 6827,8120	0,8419	-0,0099	0,9971
battery_capacity	475,6478	1,0129	0,0007	1,0002
RAM_size	910,9040	1,0177	0,0010	1,0003
refresh_rate	-16,8547	1,0006	0,0000	1,0000
GPU_size	2696,2771	1,0401	0,0020	1,0006
year2022	1384,9161	1,0317	0,0019	1,0006

Гедонічне рівняння буде мати вигляд:

$$\begin{aligned}
\ln P_i = & 3,3872 + 1.003 * RAM_{size} - 0.9971 * weight + 1.0002 * batteryCapacity \\
& + 1 * refreshRate + 1.0006 * GPUsize + 1.0004 * DOS_1 - 0.9994 \\
& * DOS_2 + 1.0034 * DOS_3 + 1.047 * DOS_4 + 1.003 * DOS_5 - 0.9965 \\
& * DOS_6 + 1.077 * DOS_7 + 0.9994 * DOS_8 + 1.0013 DstorageType_1 \\
& + 1.0059 DstorageType_2 + 1.0059 DstorageType_3 \\
& + 1.0068 DstorageType_4 + 0.9946 DstorageType_5 \\
& + 1.0141 DstorageType_6 + 1.0037 DstorageType_7 \\
& + 0.9980 DstorageType_8 + 1.0062 DstorageType_9 \\
& + 1.0022 DstorageType_{10} + 0.9955 DstorageType_{11} \\
& + 1.0047 DstorageType_{12} + 1.0095 brand_1 + 0.9989 brand_2 \\
& + 0.9912 brand_3 + 1.0028 brand_4 + 1.0023 brand_5 + 1.0017 brand_6 \\
& + 0.9966 brand_7 + 1.0009 brand_8 + 0.9999 * brand_9 \\
& + 0.9984 brand_{10} + 0.9945 brand_{11} + 1.0006 * year2022
\end{aligned}$$

Для отримання гедонічного індексу бажано мати три і більше часових проміжки, коли були зібрані результати. Гедонічний індекс отриманий із моделі log-lin із використанням УМНК після трансформації Бокса-Кокса дорівнює 1.0006.

Висновки

В цій роботі було досліджено проблему впливу зміни якості на ціну. Була здійснена розробка структурної моделі гедонічного індексу цін для ноутбуків за допомогою множинної регресії, досліджено дві специфікації моделі: лінійна та напівлогарифмічна, здійснено трансформація Бокса-Кокса до залежної змінної та застосовано узагальнений метод найменших квадратів для вирішення проблеми гетероскедастичності залишків. Знайдено гедонічні ціни для моделей. Найкращою моделлю є напівлогарифмічна.

Дані було зібрано із українських сайтів із продажу електроніки: перший збір у першому кварталі 2021 року, другий збір червень 2022 року. Для збору даних та для аналізу використано технологію веб-скрапінгу, мову програмування R та Python.

Використана література

1. Berndt Ernst R. The Practice of Econometrics: Classic and Contemporary / Ernst R. Berndt., 1991. – 702 с.
2. Gregory C. Chow G. Technological Change and the Demand for Computers / Gregory C. Gregory C. Chow. // The American Economic Review (Vol. 57, No. 5, Dec., 1967). – 1967. – С. 1117–1130. <https://www.jstor.org/stable/1814397>
3. Griliches Z. Automobile Prices Revisited: Extensions of the Hedonic Hypothesis / Z. Griliches, M. Ohta // Household Production and Consumption / Z. Griliches, M. Ohta., 1976. – С. 325 – 398. <https://www.nber.org/books-and-chapters/household-production-and-consumption/automobile-prices-revisited-extensions-hedonic-hypothesis>
4. Fifty Years of Economic Measurement: The Jubilee of the Conference on Research in Income and Wealth Volume Author/Editor: Ernst R. Berndt and Jack E. Triplett, editors // University of Chicago Press., January 1991// Hedonic Methods in Statistical Agency Environments: An Intellectual Biopsy Chapter Author: Jack E. Triplett Chapter C. 207 – 238 <http://www.nber.org/chapters/c5977>
5. Frederick V. Waugh Quality Factors Influencing Vegetable Prices. // Journal of Farm Economics/ Oxford University Press. – С. 185–196. <https://www.jstor.org/stable/1230278>
6. Doms, Mark & Aizcorbe, Ana & Corrado, Carol. When Do Matched-Model and Hedonic Techniques Yield Similar Measures? // Federal Reserve Bank of San Francisco, Working Paper Series., 2003. - С.-14.
7. Brachinger, H. W. Statistical theory of hedonic price indexes [DQE Working Paper N° 1]. University of Friburg, Friburg, Switzerland., 2002.- С.14
8. Himpens, Stéphanie, and Jean-Denis Zafar. “Web scraping Laptop Prices To Estimate Hedonic Models And Extensions To Other Predictive Methods.”, 2019. – с. 18

9. R LIBRARY CONTRAST CODING SYSTEMS FOR CATEGORICAL VARIABLES [Електронний ресурс] // UCLA: Statistical Consulting Group. – 2011. – Режим доступу до ресурсу: <https://stats.oarc.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/>.
10. Економетрика [Електронний ресурс] : методичні рекомендації і завдання до самостійної роботи за темою "Проблеми в побудові лінійних множинних регресійних моделей: гетероскедастичність" для студентів усіх спеціальностей першого (бакалаврського) рівня / уклад. І. Л. Лебедева, А. В. Жуков, С. С. Лебедєв. – Харків : ХНЕУ ім. С. Кузнеця, 2019. – 33 с
11. G. E. P. Box. An Analysis of Transformations / G. E. P. Box, D. R. Cox. // Journal of the Royal Statistical Society. Series B (Methodological). – 1964. – С. 211–252.
12. Лук'яненко І. Г., Краснікова Л. І. Л 84 Економетрика: Підручник. — К.: Товариство "Знання", КОО, 1998. — 494 с. (з табл., граф.).