МІНІСТУРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет
«Києво-Могилянська Академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

Освітній ступінь – бакалавр

На тему: **Модель агрегованих витрат на рекламу за методом Грейнджера**

Виконала студентка 4-го року навчання
освітньо-наукової програми «Прикладна математика»
спеціальності 113 Прикладна математика
Яковенко Катерина Віталіївна
Керівник: Дрінь Світлана Сергіївна,
кандидат фіз.-мат. Наук, старший викладач
Рецензент _____
Кваліфікаційна робота захищена з оцінкою _____
Секретар ЕК _____
«___» _____ 2023 р.

Київ – 2023

ЗМІСТ

ВСТУП **4**

1. Математичне моделювання та математична модель **5**

1.1. Приклади математичного моделювання **6**

2. Лаг та лагові змінні **8**

2.1. Причини лагів **10**

3. Причинність за Грейнджером **12**

4. Статистичні оцінки регресійних моделей **14**

4.1. Коефіцієнт детермінації **14**

4.2. Поняття F-критерію Фішера **14**

4.3. Залишкове стандартне відхилення **15**

5. Симультативні моделі **17**

6. Розрахунок невідомих парметрів багатофакторної регресії за мнк ... **19**

7. Гетероскедастичність та її виявлення **22**

7.1. Способи вилучення гетероскедастичності **25**

8. Практичне застосування **27**

Список літератури **41**

Тема: Модель агрегованих витрат на рекламу за методом Грейнджера

Календарний план виконання роботи:

№ п/п	Назва етапу дипломного проекту (роботи)	Термін виконання етапу	Примітка
1.	Визначення теми кваліфікаційної роботи.	29.10.2022	
2.	Визначення літератури, з якою буде відбуватись робота.	01.02.2023	
3.	Огляд та аналіз літератури за темою.	27.02.2023	
4.	Виконання практичної частини роботи.	30.03.2023	
5.	Оформлення практичної частини роботи.	15.04.2023	
6.	Оформлення теоретичної частини роботи.	30.04.2023	
7.	Створення презентації кваліфікаційної роботи.	05.05.2023	
8.	Захист кваліфікаційної роботи.	05.06.2023	

Студент _____

Керівник _____

“_____”_____

ВСТУП

Актуальність теми. Тема визначення впливу одного фактору на інший не залишає в спокої науковців та фахівців різних галузей вже не один десяток років. Так само й тема можливості прогнозування значень одного фактору, беручи до уваги значення іншого фактору за попередній період, так звані лагові значення. Саме тому професор Клайв Вільям Джон Грейнджер, який є лауреатом Нобелівської премії з економіки у 2003 році, розробив концепцію причинно-наслідкового зв'язку для покращення ефективності прогнозування. Дана концепція була названа на його честь «Тест на визначення причинності за Грейнджером».

Даний тест останніми роками дуже часто використовується у багатьох галузях, що свідчить про його ефективність. Таким чином тест Грейнджера вже з 2003 року використовують у медицині, для роботи з нейронними системами тощо.

Мета і завдання дослідження. Метою є визначення існування впливу об'єму продажів на кількість рекламних оголошень та наскільки довгий цей вплив. А також дослідити його у зворотному напрямку, чи існує вплив кількості рекламних оголошень на об'єми продажів. Завданням було застосувати тест Грейнджера та визначити наявність впливу та його довжину. На основі отриманих даних побудувати модель з лаговим ефектом та оцінити її відповідним методом.

Об'єкт дослідження. Реальний набір даних з даними за п'ять місяців (липень - листопад) платної пошукової кампанії 2021 року для одного з торгових центрів США.

Методи дослідження. Тест Грейнджера, метод найменших квадратів, кореляційні матриці, діаграми розсіювання.

1. МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА МАТЕМАТИЧНА МОДЕЛЬ

Математичне моделювання - це метод, який представляє та пояснює реальні системи, явища, використовуючи математичні формули, підходи та описи. Багато людей, серед яких науковці, econometristи, математики та інші, використовують математичні моделі для вивчення, проведення аналізу та прогнозування поведінки та подій. Також мат. моделі використовуються для вирішення складних проблем та знаходження відповідей на питання. Математична модель складається з наступних основних кроків:

1. Необхідно визначити проблему.
2. Зробити певні припущення.
3. Визначте змінні, які будуть використовуватись у моделі.
4. Обчислити рішення.
5. Проаналізувати та оцінити модель, її результати, задля перевірки її точності.

Математичні моделі бувають різними. Класифікуються мат. моделі в залежності від їхньої структури, зокрема:

- Лінійна модель: Зв'язок між усіма змінними моделі є лінійним або прямим і послідовним;
- Нелінійна модель: Зв'язок між усіма змінними моделі не є лінійним;
- Статична модель: не враховується час;
- Динамічні: враховуються зміни в часі;
- Детермінована модель: з відомими змінними;
- Стохастична модель: з невідомими змінними;

- Дедуктивна модель: базується на теорії;
- Індуктивна модель: базується на спостереженні або досвіді;
- Плаваюча модель: не є ні дедуктивною, ні індуктивною.

Математичне моделювання демонструє важливість математики для знаходження відповідей на питання і вирішення проблем.

1.1. Приклади математичного моделювання

Нижче наведено 2 приклади математичного моделювання, які вирішують реальні життєві питання та проблеми:

1. Де заправити машину бензином?

У даній моделі проблема полягає в тому, що в автомобілі закінчується паливо, у нашому випадку – бензин, і необхідно визначити, яка АЗС є найкращим варіантом для заправки. Для того, аби створити формула, спершу необхідно визначити ключові змінні. Наприклад, (1) пробіг автомобіля на бензині, (2) відстань до найближчих заправок, (3) ціна бензину на кожній з них. Вивчивши цю інформацію, з'являється можливість розрахувати вартість покупки бензину на кожній АЗС.

Потім необхідно проаналізувати та оцінити отримані варіанти та рішення. Наприклад, якщо до заправки з найдешевшим бензином їхати на 5 км далі, чи варто їхати так далеко заради бензину за нижчою ціною? Можна змінити підхід, щоб врахувати інші фактори, такі як вартість часу, витраченого на дорогу до АЗС, або вплив на навколишнє середовище під час руху до найдешевшої заправки. Отже, необхідно розглянути всі можливі варіанти та вибрати найбільш раціональний.

2. Який розмір коробки кращий?

Уявімо, компанія відправляє свою продукцію клієнтам у коробках розміром 8 х 10 х 12 дюймів і товщиною 0,75 сантиметрів. Операційний відділ нещодавно порекомендував перейти на коробки товщиною 0,5

сантиметрів. Задача стоїть, щоб визначити, скільки місця в коробках вивільниться, якщо відбудеться ця зміна.

У даній моделі можна використати розміри коробки і формулу для розрахунку об'єму, щоб порівняти поточний об'єм коробки із запропонованим новим об'ємом. Потім використати рівняння для обчислення відсотків, щоб визначити відсоток збільшення простору. Але також важливо ставити запитання до реальних проблем. Наприклад, чи варто взагалі збільшувати простір за рахунок зменшення міцності коробки?

2. ЛАГ ТА ЛАГОВІ ЗМІННІ

Багато економічних процесів характеризуються наявністю ефекту від впливу певних факторів на показники, які визначають процес. Але цей ефект виявляється зовсім не одразу, а поступово, протягом або після певного періоду часу. Дане явище називають ***лагом.***

Лаг – показник, що відображає відставання або випереджання в часі одного явища у порівнянні з іншим явищем, яке з ним пов'язано.

В економіці досить рідким є явище миттєвої залежності змінної y, що є залежною, від іншої незалежної змінної x. Частіше трапляється випадки, коли значення змінної y змінюється протягом невеликого проміжку часу після зміни незалежної змінної x. Як вже раніше зазначалось, даний проміжок називається ***часовим лагом***.

Якщо ж давати точне визначення лагу, то лаг – це економічний показник, який показує відставання або ж навпаки випередження в часі одного економічного показника (явища) у порівнянні з іншим показником.

До прикладу, розглянемо витрати та доходи. Якщо різко змінити доходи, то функція споживання теж зміниться, але дана зміна відбудеться не пропорційно доходу. Якщо збільшити дохід, то і витрати значно зростуть, щоб задовольняти поточні потреби. З часом витрати можуть зменшитись, якщо, наприклад, плануються великі витрати. Або ж, якщо вкласти кошти в капітальне будівництво, то й прибуток від нього не миттєвий. Ці все приклади до того, що багато де є саме той лаговий ефект.

Розглянемо дві моделі:

1. Модель, в якої показник у певний момент часу t, який необхідно дослідити, визначається за допомогою поточних та попередніх значень незалежних змінних, будуть називати дистрибутивно-лаговою моделлю:

$$y_t = \alpha + \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + u_t$$

2. Модель, в якої показник у певний момент часу t, який необхідно дослідити, визначається за допомогою власних попередніх значень, будуть називати авторегресійною або динамічною моделлю:

$$y_t = \alpha + \alpha_0 * x_t + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \cdots + u_t$$

Також важливо розділяти скінченні та нескінченні лагові моделі.

Економетрична модель, в якої беруть її незалежні змінні за декілька попередніх періодів, буде називатись скінченною моделлю або моделлю з кінцевим лагом.

Економетрична модель, в якої її незалежні змінні, що не обмежуються певним періодом, буде називатись нескінченною лаговою моделлю. Дане означення є досить загальним, однак на практиці робота із нескінченними лаговими моделями викликає багато труднощів, оскільки необхідно враховувати велику кількість факторів, складність структури та ін.

Коефіцієнти $\alpha_j,\ j = 0, 1, 2, \ldots$ називаються коефіцієнтами лага.

Послідовність $\{\alpha_j,\ j = 0, 1, 2, \ldots\}$ називається структурою лага.

Дистрибутивно-лагові моделі використовуються для опису економічних процесів при незмінних (стабільних) умовах. Якщо ж необхідно враховувати поточні умови функціонування, потрібно звернутись до узагальнених моделей.

Узагальнена модель розподіленого лага – модель, що містить лагові змінні та змінні, що впливають на показник дослідження, тобто враховує ще й поточні умови функціонування.

$$y_t = \sum_{\tau \geq 0} \alpha_\tau x_{t-\tau} + \sum_{i=1}^{m} b_i y_{t,i} + u_t$$

При оцінці даних рівнянь труднощів додають обмеження для коефіцієнтів при лагових змінних.

2.1. Причини лагів

У даному розділі розглянемо 3 основні причини появи лагів.

1. Психологічні причині. Відповідно до інерційних висновків, після зростання доходів або зниження цін люди не змінюють власні споживацькі звички. Можна допустити, що це відбувається внаслідок того, що сам процес зміни може створити певне миттєве незадоволення. Саме тому ті, хто зірвали джек-пот та раптово стали міліонерами, можуть навіть не змінити нічого у своєю способі або стилі життя, оскільки не одразу стає зрозуміло, як впоратись із таким неочікуваним подарунком. Звісно, що з плином часу, можна навчитись жити відповідно до нових можливостей. Реакція на збільшення доходу також може бути залежною від того, наскільки постійна дана зміна, або ж вона тимчасова. Якщо такий подарунок є одноразовим збільшенням, то можна або ж зберегти весь «прибуток», або ж повністю його витратити.
2. Технологічні причини. Розглянемо ситуацію, коли ціна капіталу залежить від скорочення праці. У даному випадку скорочення праці робить заміну капіталу працею можливою з точки зору економіки. Авжеж, щоб збільшити капітал, необхідно певний проміжок часу. Крім того, за умови, що очікується тимчасовий спад цін, а фірми передбачають (сподіваються) серйозне зростання ціни капіталу після тимчасового спаду, то можна й не поспішати із заміною капіталу працею. Звідси можна зробити певний висновок, що неповні знання теж пояснюють часові лаги.
3. Інституціональні причини. Наведемо приклад. Люди, які зберігають власні заощадження на довгострокових рахунках (будемо вважати 1-2 роки) фактично мають «зв’язані руки». Незважаючи на те, ринок може мати й інші умови, які приносять більші прибутки. Ще одним подібним прикладом є ситуація, коли

роботодавці дають на вибір працівникам декілька варіантів оформлення медичного страхуванні. Після вибору одного з представлених працівник не може змінити свій вибір на користь іншого принаймні протягом 1 року.

Отже, з вищезазначених причин лаги відіграють важливу роль в економіці. Це чітко відображено в методології коротко- та довгострокової економіки. Звідси можна зробити висновок, що короткострокові ціни, або еластичність, з доходом, як правило, менші за довгострокові ціни, еластичність.

3. ПРИЧИННІСТЬ ЗА ГРЕЙНДЖЕРОМ

У регресійному аналізі розглядається залежність однієї змінної від інших, але з цього не випливає, що одна змінна є причиною для іншої або ж навпаки. У такому випадку постає питання щодо можливості визначення причинно-наслідкового зв'язку та напряму даного зв'язку за допомогою статистичних перевірок.

Знаходження даного причинно-наслідкового зв'язку пропонується перевірити за тестом Грейнджера. Тест Грейнджера робить припущення щодо наявності потрібної інформації у часових рядах для прогнозування відповідних змінних. Даний тест оцінює дві регресії:

$$\mathrm{X}(t) = \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{j=1}^{n} \beta_j X_{t-j} + \varepsilon_{1t} \quad (1)$$

$$\mathrm{Y}(t) = \sum_{i=1}^{n} \lambda_i Y_{t-i} + \sum_{j=1}^{n} \delta_j X_{t-j} + \varepsilon_{2t} \quad (2)$$

де ε_{1t} та ε_{2t} випадкові величини, до того ж не корельовані між собою.

У рівнянні (1) можна побачити, що поточне значення X залежить від своїх попередніх значень та попередніх значень Y. Аналогічно й для рівняння (2), поточне значення Y залежить від власних попередніх значень та попередніх значень X. Звідси можна розглянути декілька випадків причинності.

1. ***Однобічна причинність від Y до X.*** Дана причинність наявна за умови, що оцінені коефіцієнти при Y у (1) рівнянні не дорівнюють нулю як група ($\sum \alpha_i \neq 0$), і при цьому оцінені коефіцієнти при X у (2) рівнянні дорівнюють нулю ($\sum \delta_i = 0$).
2. ***Однобічна причинність від X до Y.*** Дана причинність наявна за умови, що оцінені коефіцієнти при Y у (1) рівнянні дорівнюють

нулю як група ($\sum \alpha_i = 0$), і при цьому оцінені коефіцієнти при X у (2) рівнянні не дорівнюють нулю ($\sum \delta_i \neq 0$).

3. ***Зворотний зв'язок (двобічна причинність).*** Дана причинність наявна за умови, що множини оцінених коефіцієнтів при X та Y не дорівнюють нулю для обох регресій.
4. ***Відсутність причинного зв'язку між X та Y.*** Дана причинність означає, що не існує залежності між X та Y, це можливо за умови, що множини оцінених коефіцієнтів при X та Y дорівнюють нулю для обох регресій.

Даний тест, як вже було зазначено вище, проводиться на часових рядах для різних значень лагу. Нульовою гіпотезою (H_0) для даного тесту є: часовий ряд X не є причиною за Грейнджером для прогнозуванням часового ряду Y. У свою ж чергу, альтернативною гіпотезою (H_1) для даного тесту є: часовий ряд X є причиною за Грейнджером для прогнозуванням часового ряду Y.

Тест Грейнджера проводиться для різної кількості лагів, що визначити, яка кількість лагів буде значущою для тої чи іншої моделі. Висновок щодо причинності робиться на основи отриманих результатів, а саме, за допомогою значення F-статистики та його p-value. Необхідно переглянути результати тесту для всіх лагів і там, де буде найбільший перепад значення F-статистики (детальніше про F-статистику дивитись розділ 4.2) та не буде перевищення рівню значущості (5% або 10%), то ту кількість лагів і взяти для визначення причинно-наслідкового зв'язку за Грейнджером. Таким чином, якщо p-value < 0,05 (для 5-ти % рівня значущості), то можна відкинути нульову гіпотезу і зробити висновок про наявність причинного зв'язку між факторами. Якщо ж p-value > 0,05 (для 5-ти % рівня значущості), то нульову гіпотезу відкинути не можна, що свідчить при відсутність причинного зв'язку між факторами.

4. СТАТИСТИЧНІ ОЦІНКИ РЕГРЕСІЙНИХ МОДЕЛЕЙ

4.1. Коефіцієнт детермінації

Коефіцієнт детермінації – це статистичний показник, за допомогою якого можна оцінити зв'язок між двома змінними, певною мірою він використовується як один з критерії адекватності моделі.

$$R^2 = \frac{\sigma_{\text{регр.}}{}^2}{\sigma_{\text{заг.}}{}^2} = \frac{SSR}{SST}$$

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

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

де $\hat{y}$ – прогнозоване значення залежної змінної, $\bar{y}$ – середнє значення залежної змінної, y_i – спостережувальна залежна змінна.

Таким чином, чим ближче значення R^2 до одиниці, тим краще описується модель і можна зробити висновок про її адекватність.

Якщо ж значення коефіцієнта детермінації не є однозначним, наприклад, 0,5 або ж 0,4 та інші, то в такому випадку потрібно звернутись до інших критерії, за допомогою яких можна оцінити зв'язок між змінними.

4.2. Поняття F-критерію Фішера

F-статистика – це статистичний показник, який отримується в результаті проведення регресійного аналізу для того, щоб визначити чи суттєво відрізняються середні значення між двома вибірками. Даний тест є схожим до t-тесту (t-статистика). t-тест покаже, чи можна вважати одну змінну статистично значущою, а от F-статистика покаже, чи можна вважати групу змінних значущою.

При визначенні значущості результатів значення F-статистики необхідно використовувати лише разом зі значенням p-value. Оскільки,

якщо було отримано значущий результат, то це не означає, що значущими у моделі будуть всі змінні. F-статистика лише порівнює загальний вплив групи змінних разом.

Перевірка моделі на її адекватність методом F-критерію Фішера відбувається наступними кроками:

1. Спершу розраховується величина F-відношення за формулою

$$F_{(1,n-2)} = \frac{MSR}{MSE}$$

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

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

де де $\hat{y}$ – прогнозоване значення залежної змінної, $\bar{y}$ – середнє значення залежної змінної, y_i – спостережувальна залежна змінна, (n-2) – ступені вільності, що пов'язані зі значенням MSR та MSE.

2. Далі визначається рівень значущості. Якщо визначено, що рівень значущості 5%, то це означатиме, що допустима помилка не більше за 5% випадків, а в решті 95% випадків висновки будуть вірними.
3. На даному кроці необхідно розрахувати критичне значення ($F_{\text{кр.}}$). Для цього потрібно використати статистичні таблиці F-розподілу Фішера для (1, n-2) ступенів вільності і відповідним рівнем значущості.
4. Якщо обчислене значення F-критерію буде більше за критичне значення $F_{\text{кр.}}$, тоді можна відкинути нульову гіпотезу, що $\beta_0 = 0$, і визначити, що побудована модель є адекватною.

4.3. Залишкове стандартне відхилення

Залишкове стандартне відхилення – статистичний показник, який використовують, щоб описати різницю між стандартними відхиленнями

спостережувальних значень та стандартними відхиленнями прогнозованих значень.

$$Residual = Y - Y_{est}$$

$$S_{res} = \sqrt{\frac{\sum(Y - Y_{est})^2}{n - 2}}$$

де S_{res} – залишкове стандартне відхилення, Y – спостережувальне значення, Y_{est} – прогнозоване значення, n – кількість даних у вибірці.

5. СИМУЛЬТАТИВНІ МОДЕЛІ

Система симультативних рівнянь або система одночасних рівнянь – система, яка описує взаємну залежність між факторами (змінними).

Класичний регресійний аналіз має певні припущення, і одне з них (4-те) говорить про незалежність факторів та випадкових величин.

$$cov(x, \varepsilon) \neq 0$$

Якщо застосовувати метод найменших квадратів (далі МНК) до одного рівняння, то необхідно вважати, що незалежні змінні є екзогенними та існує виключно односторонній зв'язок між залежною змінною y та незалежними змінними x. Якщо ж дані умови не справджуються, тим самим змінні x виражаються змінною y, тоді 4-те припущення регресійного аналізу про незалежність факторів та випадкових величин порушується. У такому випадку визначення невідомих параметрів моделі призводить до неефективних оцінок, які мають відхилення.

Для випадку, коли y це деяка функція від x, а x це деяка функція від y, тоді використання одного регресійного рівняння, щоб визначити зв'язок між змінними y та x, неприпустимо. Необхідно перейти від моделі з одним регресійним рівнянням до моделі з декількома регресійними рівняннями. У новій моделі з декількома регресійними рівняннями допускаються рівняння, які містять ендогенні та екзогенні змінні. Дана модель рівнянь буде мати назву ***система одночасних або симультативних рівнянь***.

Нижче пропонується розглянути приклади використання симультативних систем, які відображають значення одночасних зав'язків при описі економічних подій. А також визначити, чому з'являється порушення припущення про незалежність факторів та випадкових величин.

Припустимо, що є необхідність в оцінці пропозиції грошової маси. Для того, аби уникнути інфляції, уряд контролює даний процес. Звідси

можна припустити, що реальний рівень доходів буде основним детермінантом у рішенні стосовно грошової пропозиції.

Запишемо функцію грошової пропозиції наступним чином:

$$M = \beta_0 + \beta_1 y + \varepsilon,$$

де M – грошова пропозиція;

y – рівень реального доходу.

Проте рівняння грошової пропозиції не можна розглядати, як повна модель, оскільки рівень реального доходу буде залежати від грошової пропозиції та деяких інших чинників, як, наприклад, соціальна політика уряду або ж інвестиції підприємців. Отже, y не є екзогенною змінною, тобто між M та y існує залежність.

Запишемо рівняння даної залежності:

$$y = \alpha_0 + \alpha_1 M + \alpha_2 I_t + \cdots + \upsilon$$

Тепер підставимо в рівняння для y вираз для M і отримаємо:

$$y = \alpha_0 + \alpha_1(\beta_0 + \beta_1 y + \varepsilon) + \alpha_2 I_t + \cdots + \upsilon$$

Звідси бачимо, що змінна y має залежність від випадкової величини ε, що знову ж таки порушує 4-те припущення класичної регресії.

6. РОЗРАХУНОК НЕВІДОМИХ ПАРМЕТРІВ БАГАТОФАКТОРНОЇ РЕГРЕСІЇ ЗА МНК

Розглянемо ряд спостережень, в якому:

- $y = \{y_1,\ y_2, y_3, \dots, y_n\}$ – залежна змінна;
- $x_1 = \{x_{11}, x_{12}, \dots, x_{1n}\}$; $x_2 = \{x_{21}, x_{22}, \dots, x_{2n}\}$, … , $x_p = \{x_{p1}, x_{p2}, \dots, x_{pn}\}$ – незалежні змінні (фактори).

Побудуємо лінійну багатофакторну модель, спираючись на отримані спостереження:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + e$$

де y – залежна змінна, $b_0, b_1, \dots, b_p$ – невідомі параметри, $x_0, x_1, \dots, x_p$ – незалежні змінні, e – випадкова величина.

Далі необхідно знайти невідомі параметри, використовуючи метод найменших квадратів (МНК). Суть методу найменших квадратів полягає в тому, аби мінімізувати суму квадратів відхилень фактичних даних від даних, які отримуються з регресійної моделі.

За МНК тоді маємо:

$$F(b_0, b_1, \dots, b_n) = min \sum_{i=1}^{n} {e_i}^2 = min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - \cdots - b_p x_p)^2$$

Щоб мінімізувати значення вищезазначеного виразу, необхідно взяти частинні похідні функції $F(b_0, b_1, \dots, b_n)$ та прирівняти їх до 0. Після цього отримуємо систему $(p + 1)$ нормальних рівнянь з $(p - 1)$ невідомими:

$$nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + \cdots + b_p \sum_{i=1}^{n} x_{pi} = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} {x_{1i}}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_p \sum_{i=1}^{n} x_{1i}x_{pi} = \sum_{i=1}^{n} x_{1i}y_{1i}$$

$$b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{2i}x_{1i} + b_2 \sum_{i=1}^{n} {x_{2i}}^2 + \cdots + b_p \sum_{i=1}^{n} x_{2i}x_{pi} = \sum_{i=1}^{n} x_{2i}y_i$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$b_0 \sum_{i=1}^{n} x_{2p} + b_1 \sum_{i=1}^{n} x_{pi}x_{1i} + b_2 \sum_{i=1}^{n} x_{pi}x_{2i} + \cdots + b_p \sum_{i=1}^{n} {x_{pi}}^2 = \sum_{i=1}^{n} x_{pi}y_i$$

Далі розпишемо в матричному вигляді вищезазначені рівняння:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{pi} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} {x_{1i}}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{pi} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} x_{pi} & \sum_{i=1}^{n} x_{pi}x_{1i} & \sum_{i=1}^{n} x_{pi}x_{12} & \cdots & \sum_{i=1}^{n} {x_{pi}}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_p \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{21} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pm} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_p \end{pmatrix}$$

Це ж рівняння можна записати й в скороченому вигляді:

$$X'Xb = X'y$$

Помножимо обидві частини даного рівняння на обернену матрицю $(X'X)^{-1}$ і отримаємо:

$$(X'X)^{-1}(X'X)b = (X'X)^{-1}(X'y)$$

Оскільки $(X'X)^{-1}(X'X) = I$ – одиничній матриці, яка має розмір $(p + 1) \times (p - 1)$, то матимемо:

$$Ib = (X'X)^{-1}(X'y)$$
$$b = (X'X)^{-1}(X'y)$$

Отримане рівняння є основним у визначенні невідомих параметрів у матричному вигляді. Аби отримати вектор невідомих параметрів, необхідно диференціювати рівняння за кожним параметром $(b_0, b_1, \dots, b_p)$, а потім частинні похідні прирівняти до 0.

Варто також зазначити, що параметр b_0 розраховується аналогічно до регресії з використанням середніх значень. Таким чином маємо:

$$b_0 = \bar{y} - b_1\overline{x_1} - \cdots - b_p\overline{x_p}$$

7. ГЕТЕРОСКЕДАСТИЧНІСТЬ ТА ЇЇ ВИЯВЛЕННЯ

У моделі класичної лінійної регресії є припущення щодо сталості кожної випадкової величини ε_i. Дане припущення говорить про наявність гомоскедастичності між факторами. Має воно наступний вигляд:

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

Якщо ж дане припущення порушується, то з'являється місце гетероскедастичності між факторами. Можна записати так:

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

У статистиці гетероскедастичність з'являється за умови, що стандартні відхилення прогнозованої змінної, що відстежувались для різних значень незалежної змінної або в різні попередні періоди часу, не є постійними ($\neq const$).

Якщо ж наявна гетероскедастичність, то візуалізація залишкових похибок буде мати тенденцію до розсіювання в часі. Подібну візуалізацію наведено нижче (Рис. 1).

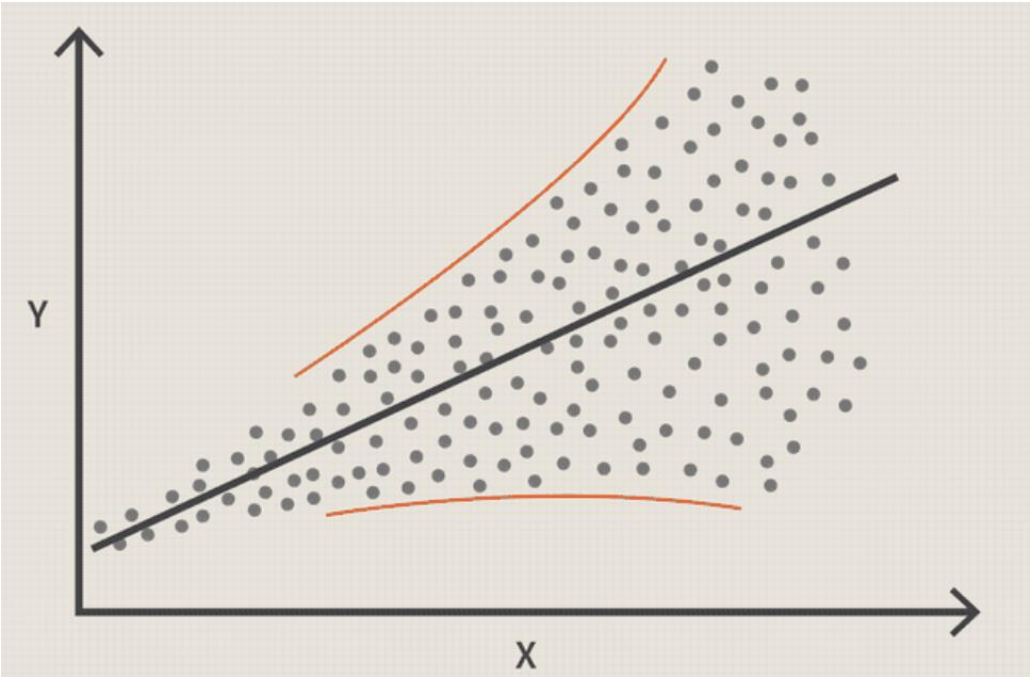

Рис. 1 "Візуалізація гетероскедастичності"

Графічне представлення розкиду спостережень залежить від форми гетероскедастичності, тобто від форми зв'язку між дисперсією та незалежними змінними. Нижче наведено різні форми гетероскедастичності:

- При зростанні дисперсії ε (Рис. 2);
- При спаданні дисперсії ε (Рис. 3);
- Складний випадок – спершу дисперсія ε зменшується при зростанні x, але після певного рівня починає зростати при зростанні x (Рис. 4).

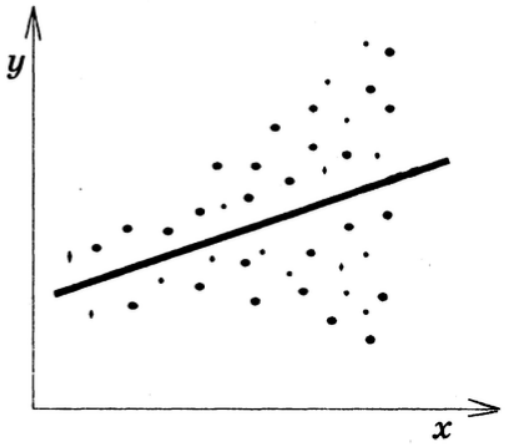

Рис. 2 "Зростання дисперсії ε"

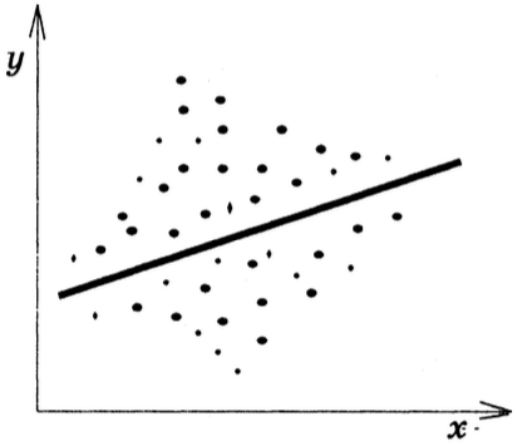

Рис. 3 "Спадання дисперсії ε"

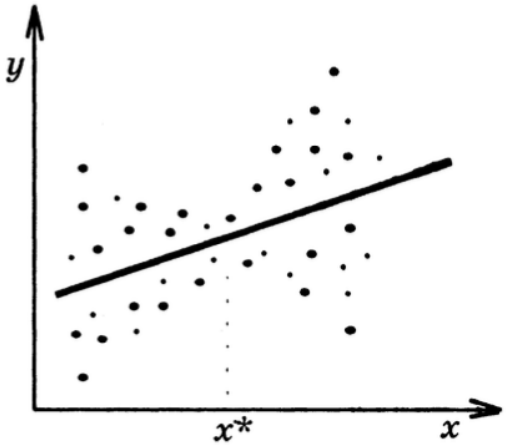

Рис. 4 "Складний випадок"

7.1. Способи вилучення гетероскедастичності

Якщо було виявлено гетероскадестичність за допомогою будь-якого тесту, то, щоб її позбутись, необхідно змінити початкову модель. Змінити модель потрібно саме так, щоб помилки мали постійну дисперсію. А далі вже невідомі параметри можна вирахувати за допомогою МНК.

Спосіб трансформації моделі дуже залежить від того, яка саме наявна форма гетероскедастичності. Тобто потрібно встановити форму залежності дисперсії ${\sigma_{\varepsilon i}}^2$ та незалежних змінних.

Розглянемо один зі способів трансформації моделі задля вилучення гетероскедастичності. Уявімо, що наша початкова модель має вигляд:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

в якій ε_i гетероскедастична, але дана випадкова величина задовольняє решту припущень лінійної регресії.

Нехай гетероскедастичність має наступну форму:

$$E(\varepsilon_i)^2 = {\sigma_\varepsilon}^2 = k^2 x_i$$

Трансформуємо початкову модель шляхом ділення на $\sqrt{x}$. Тоді трансформована модель матиме наступний вигляд:

$$\frac{y}{\sqrt{x}} = \frac{\beta_0}{\sqrt{x}} + \frac{\beta_1 x}{\sqrt{x}} + \frac{\varepsilon}{\sqrt{x}}$$

Отже, для отриманої трансформованої моделі маємо випадкову величину $\frac{\varepsilon}{\sqrt{x}}$ гомоскедастичну, яке має сталу дисперсію k^2:

$$E\left(\frac{\varepsilon_i}{\sqrt{x_i}}\right)^2 = \frac{1}{x_i}E(\varepsilon_i{}^2) = \frac{1}{x_i}\left(\sigma_{\varepsilon_i}{}^2\right) = k^2$$

Таким чином, виконавши вищенаведені кроки, можна позбутись гетероскедастичності.

8. ПРАКТИЧНЕ ЗАСТОСУВАННЯ

Мною був знайдений та використаний датасет, що має назву «Shopping Mall Paid Search Campaign Dataset».

Опис датасету: реальний набір даних з даними за п’ять місяців (липень - листопад) платної пошукової кампанії 2021 року для одного з торгових центрів США. Торговий центр просував свої товари, використовуючи платні пошукові купони та промокоди.

Датасет складається з 12-ти колонок (атрибутів), 190-та значень.

Атрибути датасету:

1. Ad Group – категорія рекламної кампанії по пристроям, на які було запущено кампанію;
2. Month – місяць, коли було запущено рекламну кампанію;
3. Impressions – кількість показів рекламного оголошення;
4. Clicks – кількість кліків на рекламне оголошення;
5. CTR – відсоток користувачів, що перейшли (клікнули) на рекламне оголошення серед всіх, хто побачив рекламне оголошення;
6. Conversions – кількість покупок, здійснених по рекламному оголошенню;
7. Conv Rate - відсоток користувачів, які здійснили покупку, з усіх, хто перейшов за оголошенням;
8. Costs – витрати на рекламу;
9. CPC – вартість одного кліку;
10. Revenue – дохід особи, яка запускала рекламну кампанію;
11. Sale Amount – прибуток торгового центру з рекламної кампанії;
12. Profit&Loss – прибуток особи, яка запускала рекламну кампанію.

Підготовка даних

Спершу було проведено дослідження на існування пропущених даних, таких даних не було виявлено. Потім було створено новий датафрейм, який складався з наступних атрибутів:

1. Ad Group – категорія рекламної кампанії по пристроям, на які було запущено кампанію;
2. Month – місяць, коли було запущено рекламну кампанію;
3. Impressions – кількість показів рекламного оголошення;
4. Sale Amount – прибуток торгового центру з рекламної кампанії.

Далі на основі створеного датафрейму було розроблено 2 нових датасети. Перший був для рекламних кампаній на мобільних пристроях, другий – на комп'ютерних пристроях.

Далі проводилась робота безпосередньо з двома новими датасетами. Було виявлено необхідність у зміні типів атрибутів, таким чином було змінено на факторинй тип наступні атрибути:

1. Ad Group – 2 рівня;
2. Month – 5 рівнів.

Тест Грейнджера

Далі для обох датасетів було застосовано спершу тест Грейнджера, в якому:

H_0 (нульова гіпотеза): об'єм продажів не є причиною за Грейнджером для прогнозування кількості рекламних оголошень;

H_1 (альтернативна гіпотеза): об'єм продажів є причиною за Грейнджером для прогнозування кількості рекламних оголошень.

Результати тесту по датасету для комп'ютерних пристроїв показали, що лагового ефекту немає і відхилити нульову гіпотезу неможливо навіть для значення в 1 лаг.

```
> grangertest(Impressions ~ Sale_Amount, order = 1, data = Desk)
Granger causality test

Model 1: Impressions ~ Lags(Impressions, 1:1) + Lags(Sale_Amount, 1:1)
Model 2: Impressions ~ Lags(Impressions, 1:1)
  Res.Df Df      F Pr(>F)
1      1
2      2 -1 0.0214 0.9076
```

Рис. 5 "Застосування тесту Грейнджера для датасету Деск"

Результати тесту по датасету для мобільних пристроїв показали, що існує лаговий ефект в одне значення, тобто можна відхилити нульову гіпотезу для значення в 1 лаг.

```
> grangertest(Impressions ~ Sale_Amount, order = 1, data = Mob)
Granger causality test

Model 1: Impressions ~ Lags(Impressions, 1:1) + Lags(Sale_Amount, 1:1)
Model 2: Impressions ~ Lags(Impressions, 1:1)
  Res.Df Df      F  Pr(>F)
1      1
2      2 -1 170.61 0.04864 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рис. 6 "Застсоування тесту Грейнджера для датасету Моб"

Таким чином було побудовано наступну модель, базуючись на даних тесту Грейнджера:

$$A_t = \alpha_0 + \alpha_1 S_t + \alpha_2 S_{t-1} + \varepsilon_t$$

де A_t – кількість рекламних оголошень за поточний місяць, S_t – об’єм продажів за поточний місяць, S_{t-1} – об’єм продажів за попередній місяць, ε_t – шум.

Далі було застосовано тест Грейнджера для тих самих датасетів, але змінено гіпотези:

H_0 (нульова гіпотеза): кількость рекламних оголошень не є причиною за Грейнджером для прогнозування об’єму продажів;

H_1 (альтернативна гіпотеза): кількость рекламних оголошень є причиною за Грейнджером для прогнозування об’єму продажів.

Результати тесту по датасету для комп'ютерних пристроїв показали, що лагового ефекту немає і відхилити нульову гіпотезу неможливо навіть для значення в 1 лаг.

```
> grangertest(Sale_Amount ~ Impressions, order = 1, data = Desk)
Granger causality test

Model 1: Sale_Amount ~ Lags(Sale_Amount, 1:1) + Lags(Impressions, 1:1)
Model 2: Sale_Amount ~ Lags(Sale_Amount, 1:1)
  Res.Df Df      F Pr(>F)
1      1
2      2 -1 0.0317 0.8877
```

Рис. 7 "Застосування тесту Грейнджера для датасету Деск"

Результати тесту по датасету для мобільних пристроїв показали, що лагового ефекту немає і відхилити нульову гіпотезу неможливо навіть для значення в 1 лаг.

```
> grangertest(Sale_Amount ~ Impressions, order = 1, data = Mob)
Granger causality test

Model 1: Sale_Amount ~ Lags(Sale_Amount, 1:1) + Lags(Impressions, 1:1)
Model 2: Sale_Amount ~ Lags(Sale_Amount, 1:1)
  Res.Df Df      F  Pr(>F)
1      1
2      2 -1 43.182 0.09614 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рис. 8 "Застсоування тесту Грейнджера для датасету Моб"

Тож звідси побудувати модель із лаговими значеннями не вдалось.

Далі відбувається робота із побудованою раніше моделлю:

$$A_t = \alpha_0 + \alpha_1 S_t + \alpha_2 S_{t-1} + \varepsilon_t$$

Було побудовано діаграму розсіювання для визначення зв'язку між факторами.

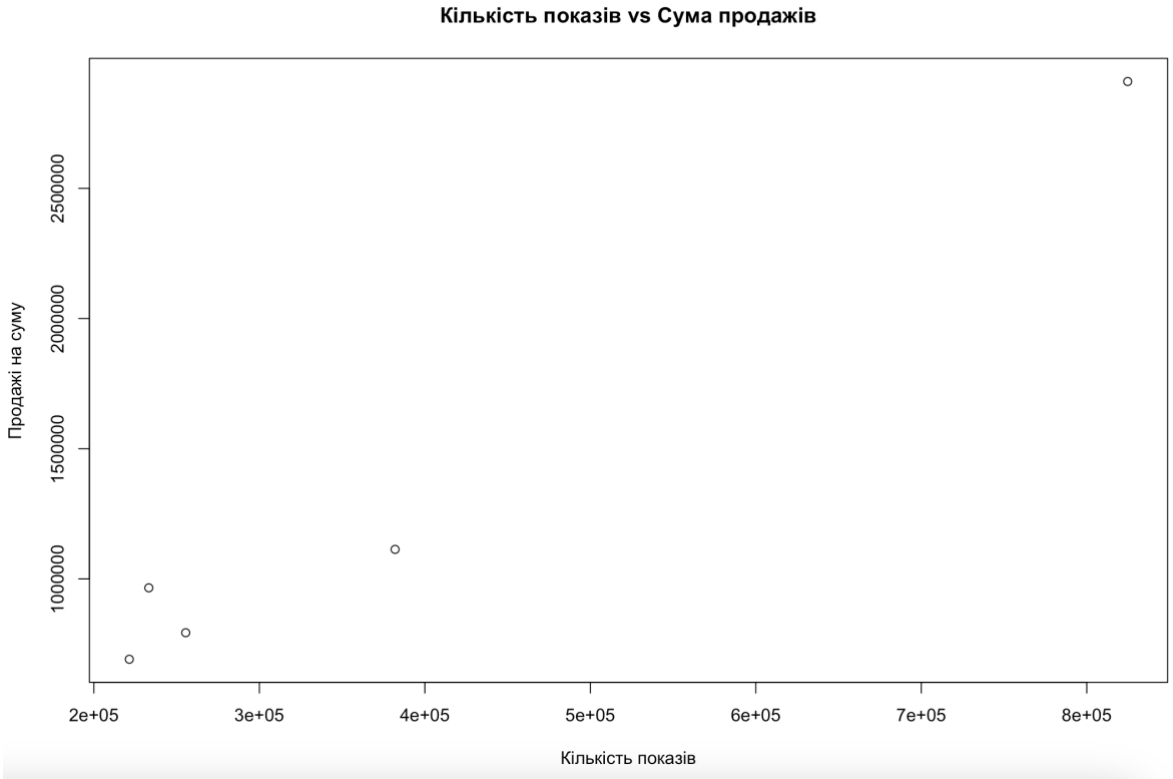

Рис. 9 "Діаграма розсіювання для кількості рекламних оголошень та об'єму продажів"

За діаграмою визначено лінійний зв’язок між атрибутами. Для оцінки невідомих параметрів найбільше підходить лінійна модель. Було оцінено побудовану модель методом найменших квадратів. І отримано наступні статистичні результати.

```
> model_2 <- lm(data=Mob, Impressions ~ Sale_Amount+Sale_Amount_t_1_2)
> model_2

Call:
lm(formula = Impressions ~ Sale_Amount + Sale_Amount_t_1_2, data = Mob)

Coefficients:
      (Intercept)        Sale_Amount  Sale_Amount_t_1_2
       -1.273e+05          2.692e-01          1.754e-01

> summary(model_2)

Call:
lm(formula = Impressions ~ Sale_Amount + Sale_Amount_t_1_2, data = Mob)

Residuals:
       1        2        3        4        5
 24109.6 -26047.4  23468.2 -20701.2   -829.1

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.273e+05  8.874e+04  -1.434  0.28793
Sale_Amount        2.692e-01  1.847e-02  14.574  0.00467 **
Sale_Amount_t_1_2  1.754e-01  9.525e-02   1.842  0.20683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33460 on 2 degrees of freedom
Multiple R-squared:  0.9914,    Adjusted R-squared:  0.9828
F-statistic: 115.1 on 2 and 2 DF,  p-value: 0.008616
```

Рис. 10 "Оцінка параметрів моделі з лаговими значеннями за МНК"

За результатами статистичних перевірок можна побачити, що значення в один лаг не є значущим для нашої моделі та погіршує оцінку моделі.

Таким чином, теоретично ми побачили, що коефіцієнт з лаговими значенням в один крок повинен описувати нашу модель (перевірено за допомогою тесту Грейнджера), але практично ми довели, що коефіцієнт з лаговими значенням в один крок є незначущим для нашої моделі. Тому необхідно розглянути іншу модель. А оскільки модель не залежить від лагових змінних, перейдемо до факторного аналізу.

Розглянемо модель для загального датасету:

$$Sale.Amount = \beta_0 + \beta_1 Impressions + \beta_2 Costs + \beta_3 CPC + \beta_4 Clicks + \beta_5 CTR + \beta_6 Conversions + \beta_7 P.L + \beta_8 Revenue + \beta_9 Ad\_Device$$

Далі цю модель було оцінено факторним аналізом.

```
lm(formula = Sale.Amount ~ Ad_Device + Impressions + Costs +
    CPC + Clicks + CTR + Conversions + P.L + Revenue, data = Campaign_res)

Coefficients:
(Intercept)    Ad_Device  Impressions        Costs          CPC       Clicks          CTR  Conversions
 -6.197e+02   -7.561e+02   -7.993e-02   -3.936e+00   -2.179e+02    8.529e-02    1.266e+04    1.528e+00
       P.L      Revenue
 -5.440e+00    2.464e+01

> summary(gen_model)

Call:
lm(formula = Sale.Amount ~ Ad_Device + Impressions + Costs +
    CPC + Clicks + CTR + Conversions + P.L + Revenue, data = Campaign_res)

Residuals:
    Min      1Q  Median      3Q     Max
-7361.4 -1489.5  -211.1   873.5  9773.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.197e+02  1.262e+03  -0.491    0.624
Ad_Device   -7.561e+02  6.070e+02  -1.246    0.215
Impressions -7.993e-02  5.716e-02  -1.398    0.164
Costs       -3.936e+00  5.215e+00  -0.755    0.451
CPC         -2.179e+02  7.852e+02  -0.278    0.782
Clicks       8.529e-02  1.655e-01   0.515    0.607
CTR          1.266e+04  2.650e+03   4.778 3.66e-06 ***
Conversions  1.528e+00  4.413e+00   0.346    0.730
P.L         -5.440e+00  5.225e+00  -1.041    0.299
Revenue      2.464e+01  5.448e+00   4.524 1.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2550 on 180 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9996
F-statistic: 5.079e+04 on 9 and 180 DF,  p-value: < 2.2e-16
```

Рис. 11 "Оцінка параметрів моделі за МНК"

Після застосування статистичних перевірок, можна побачити, що лише два фактори є значущими в моделі:

1. CTR;
2. Revenue.

Також було виявлено, що в моделі такого виду навіть атрибут «Кількість рекламних оголошень» не є значущим, що суперечить усім попереднім перевіркам.

Далі було побудовано кореляційну матрицю, щоб визначити, які саме параметри можна викинути з моделі.

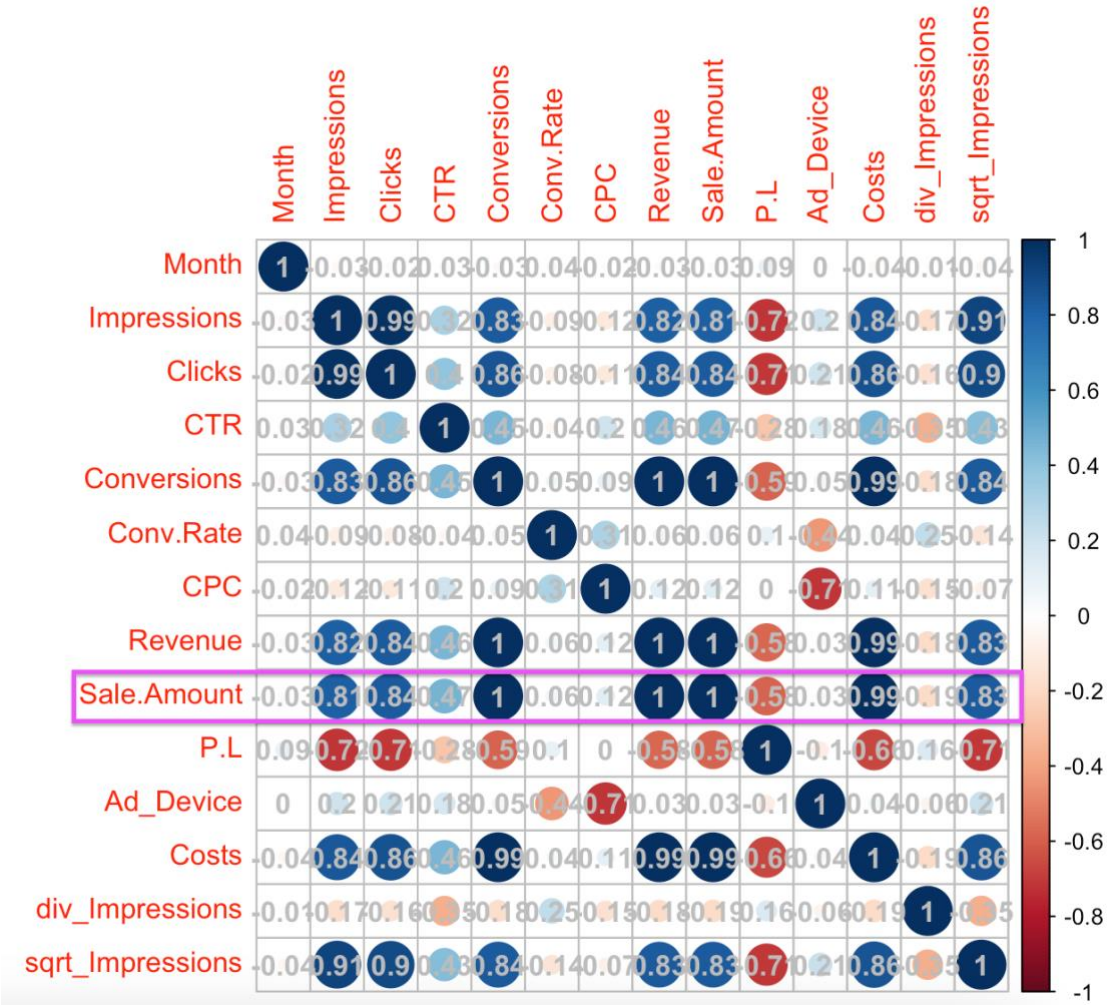

Рис. 12 "Кореляційна матриця"

Пояснення: з кореяційної матриці бачимо, що у Sale.Amount наявна кореляція з:

1. Impressions – будемо враховувати в модель;
2. Clicks – під питанням (велика кореляція з Impressions), але поки враховуємо;
3. CTR – будемо враховувати в модель;
4. Conversions – відкидаємо, оскільки велика кореляція з Costs, для нашої моделі Costs є важливішим;

5. Revenue – відкидаємо, оскільки велика кореляція з Costs, для нашої моделі Costs є важливішим;
6. P.L – будемо враховувати в модель;
7. Costs – будемо враховувати в модель;
8. CPC – відкидаємо, оскільки кореляція мала.

У результаті отримаємо наступну модель:

$$Sale.Amount = \beta_0 + \beta_1 Impressions + \beta_1 Costs + \beta_2 Clicks + \beta_3 CTR + \beta_4 P.L + \beta_5 Ad\_Device$$

Потім було оцінено нову отриману модель.

```
> gen_model_2 <- lm(data=Campaign_res, Sale.Amount ~ Ad_Device + Impressions
+                 + Clicks+ CTR + Costs + P.L)
> gen_model_2

Call:
lm(formula = Sale.Amount ~ Ad_Device + Impressions + Clicks +
    CTR + Costs + P.L, data = Campaign_res)

Coefficients:
(Intercept)    Ad_Device  Impressions       Clicks          CTR        Costs          P.L
 -9.812e+02   -7.248e+02   -3.445e-02    3.191e-02    1.242e+04    2.093e+01    1.938e+01

> summary(gen_model_2)

Call:
lm(formula = Sale.Amount ~ Ad_Device + Impressions + Clicks +
    CTR + Costs + P.L, data = Campaign_res)

Residuals:
    Min      1Q  Median      3Q     Max
-7248.6 -1407.8  -286.3   803.6 13353.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.812e+02  8.375e+02  -1.171   0.2429
Ad_Device   -7.248e+02  4.236e+02  -1.711   0.0888 .
Impressions -3.445e-02  4.991e-02  -0.690   0.4910
Clicks       3.191e-02  1.351e-01   0.236   0.8135
CTR          1.242e+04  2.366e+03   5.248 4.23e-07 ***
Costs        2.093e+01  6.509e-02 321.522  < 2e-16 ***
P.L          1.938e+01  3.182e-01  60.915  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2700 on 183 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic: 6.795e+04 on 6 and 183 DF,  p-value: < 2.2e-16
```

Рис. 13 "Оцінка параметрів моделі за МНК"

Після проведення статистичної перевірки, бачимо, що ще деякі фактори залишаються незначущими. Тож було побудовано графік, щоб визначити поведінку параметрів.

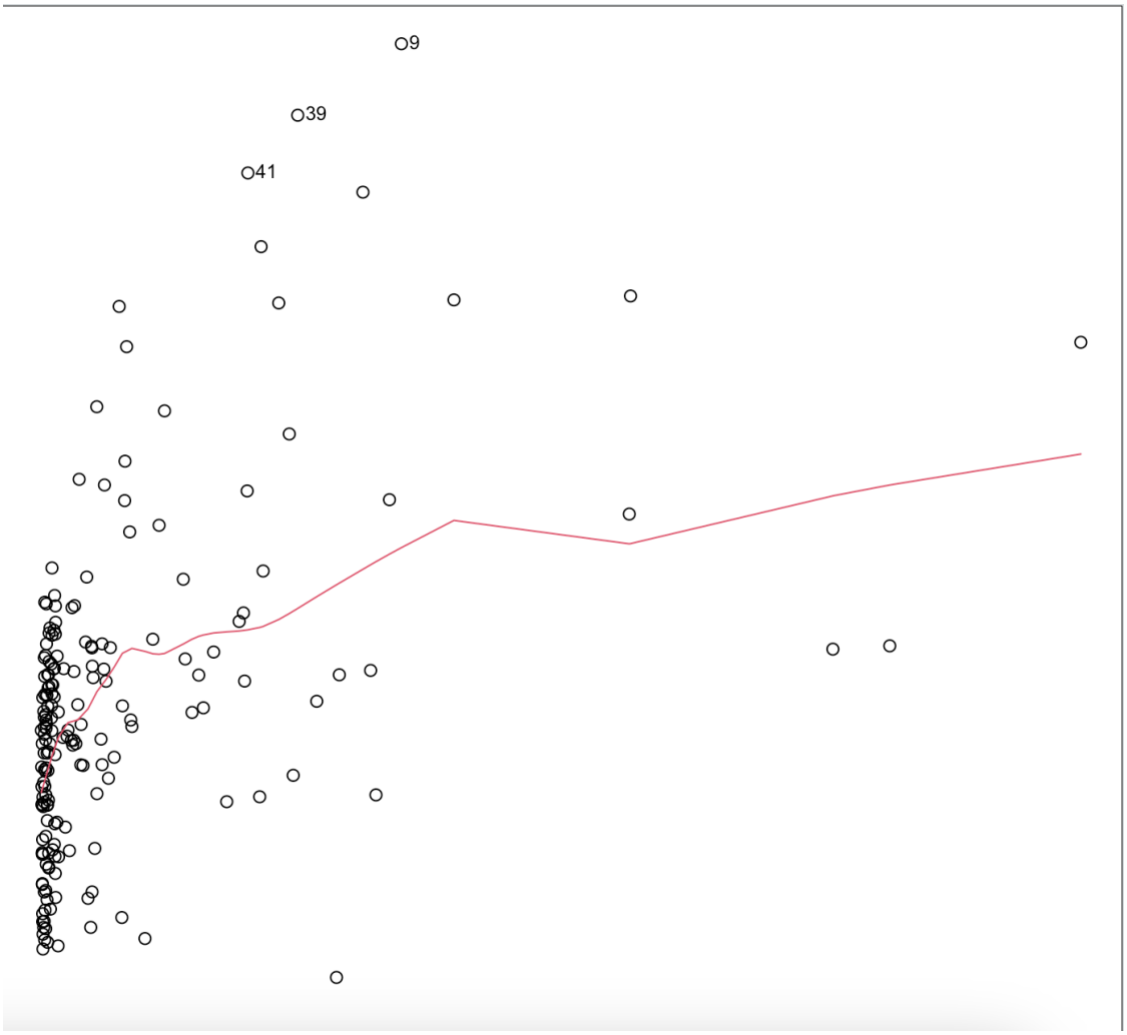

Рис. 14 "Графік для виявлення гетероскедастичності"

Бачимо, що червона лінія не пряма, хоча повинна такою бути, отже, було зроблено припущення, що існує гетероскедастичність. Перевіримо це.

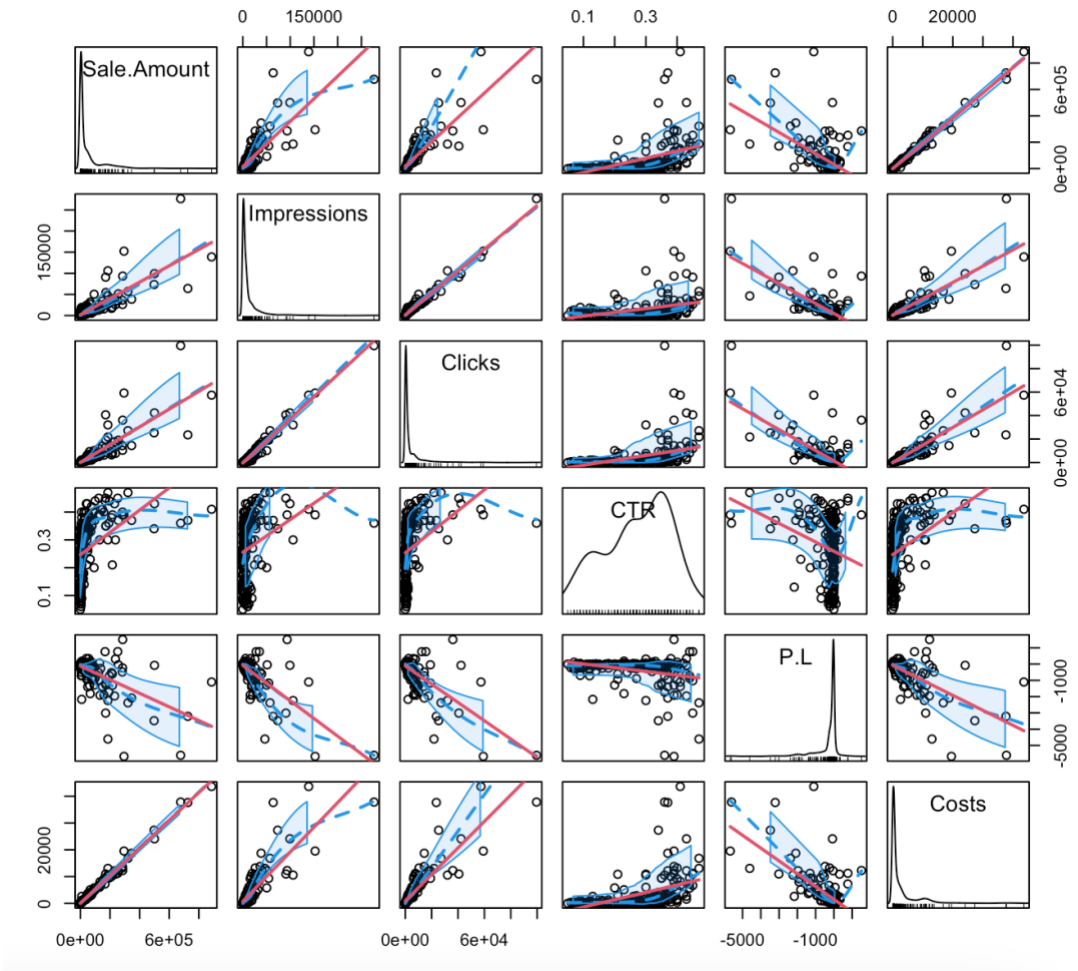

Рис. 15 "Графіки зв'язку між фаткорами"

Можна помітити, що для параметру Impressions та Clicks наявна гетероскедастичність, з параметром CTR потрібно попрацювати, візьмемо, як експоненту, а от з параметром Cost ідеальний лінійний зв’язок. Попрацюємо з параметрами. З параметру impressions візьмемо квадратний корінь. І перевіримо по графікам ще раз.

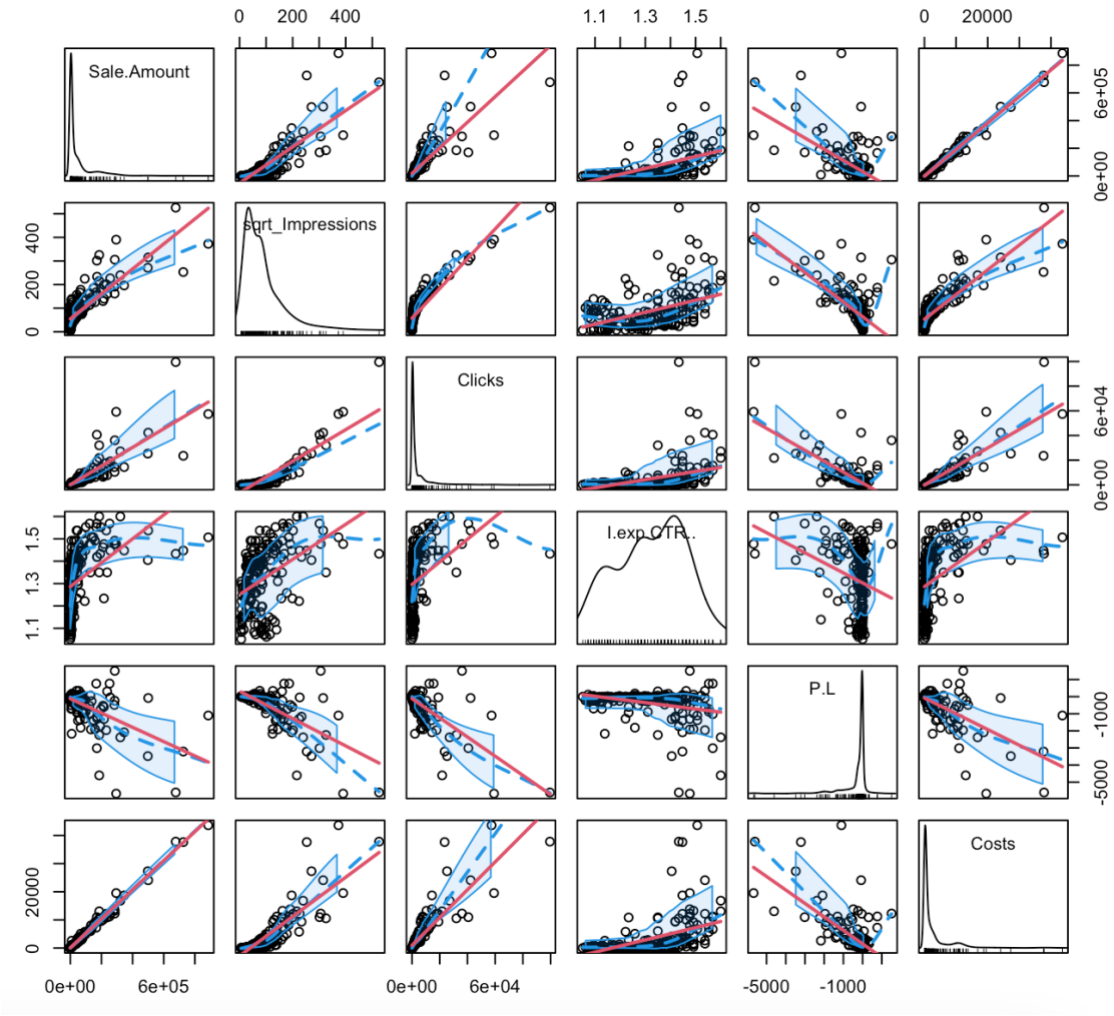

Рис. 16 "Графіки зв'язку між факторами після обробки"

Маємо кращі результати. Тоді перевіримо значущість усіх врахованих параметрів.

```
Call:
lm(formula = Sale.Amount ~ Ad_Device + sqrt_Impressions + Clicks +
    I(exp(CTR)) + Costs + P.L, data = Campaign_res)

Coefficients:
     (Intercept)         Ad_Device  sqrt_Impressions            Clicks       I(exp(CTR))
      -1.160e+04        -9.651e+02         1.762e+01        -1.209e-01         1.015e+04
           Costs               P.L
       2.085e+01         1.961e+01

> summary(final_model)

Call:
lm(formula = Sale.Amount ~ Ad_Device + sqrt_Impressions + Clicks +
    I(exp(CTR)) + Costs + P.L, data = Campaign_res)

Residuals:
    Min      1Q  Median      3Q     Max
-7203.7 -1211.1  -225.3   803.7 13794.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.160e+04  2.014e+03  -5.759 3.52e-08 ***
Ad_Device        -9.651e+02  4.176e+02  -2.311  0.02195 *
sqrt_Impressions  1.762e+01  6.092e+00   2.892  0.00430 **
Clicks           -1.209e-01  4.332e-02  -2.792  0.00579 **
I(exp(CTR))       1.015e+04  1.579e+03   6.431 1.07e-09 ***
Costs             2.085e+01  6.763e-02 308.349  < 2e-16 ***
P.L               1.961e+01  3.087e-01  63.540  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2624 on 183 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9996
F-statistic: 7.191e+04 on 6 and 183 DF,  p-value: < 2.2e-16
```

Рис. 17 "Оцінка параметрів моделі за МНК"

Бачимо, що всі параметри є значущими для нашої моделі. Тож кінцевий вигляд моделі буде таким:

$$Y = -1160 - 9651.1X_1 + 17.62\sqrt{X_2} - 12.09X_3 + 10150e^{X_4} + 20.85X_5 + 1961X_6$$

Y – Sale Amount

X_1 – Ad Group

X_2 – Impressions

X_3 – Clicks

X_4 – CTR

X_5 – Costs

X_6 – Profit&Loss

Як висновок: рекламні оголошення на комп'ютерах дають більший вплив на об'єм продажів, ніж рекламні оголошення на мобільних пристроях.

СПИСОК ЛІТЕРАТУРИ

1. *Shopping Mall Paid Search Campaign Dataset*. (2022, September 29). Kaggle. https://www.kaggle.com/datasets/marceaxl82/shopping-mall-paid-search-campaign-dataset?resource=download
2. Indeed Editorial Team. (2022). Everything You Need To Know About Mathematical Modeling. *Indeed Career Guide*. https://www.indeed.com/career-advice/career-development/mathematical-modeling
3. *Поняття лагу і лагових змінних.* (n.d.). StudFiles. https://studfile.net/preview/9444367/page:21/
4. Вікімедіа, У. П. (2019). Лаг. *uk.wikipedia.org*. https://uk.wikipedia.org/wiki/%D0%9B%D0%B0%D0%B3
5. Вікімедіа, У. П. (2022). Лаг часовий. *uk.wikipedia.org*. https://uk.wikipedia.org/wiki/%D0%9B%D0%B0%D0%B3_%D1%87%D0%B0%D1%81%D0%BE%D0%B2%D0%B8%D0%B9
6. Hayes, A. (2022). Heteroscedasticity Definition: Simple Meaning and Types Explained. *Investopedia*. https://www.investopedia.com/terms/h/heteroskedasticity.asp
7. *F Statistic / F Value: Definition and How to Run an F-Test*. (2023, March 3). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/
8. Kenton, W. (2020). Residual Standard Deviation: Definition, Formula, and Examples. *Investopedia*. https://www.investopedia.com/terms/r/residual-standard-deviation.asp
9. Краснікова Л. І. Економетрика / Л. І. Краснікова, І. Г. Лук'яненко. – Київ: Товариство "Звання", 1998. – 494 с. – (966-7293-08-4).