ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД

«ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ»

МІНІСТЕРСТВА ОСВІТИ І НАУКИ УКРАЇНИ

Кафедра загальної математики

**Індивідуальна робота**

з математичного моделювання

на тему:

«Проста лінійна регресія. Метод найменших квадратів»

Виконали:

студенти 4-го курсу

групи 4214

математичного факультету

спеціальності «Математика»

Сіменович Т.В.

Паназдир А.В.

Леонов А.

Запоріжжя, 2017 р.

ЗМІСТ

1. ПРОСТА ЛІНІЙНА РЕГРЕСІЯ…………………………………………..3
2. МЕТОД НАЙМЕНШИХ КВАДРАТІВ…………………………………..5
3. ПРАКТИЧНА ЧАСТИНА………………………………………………..10
4. ЛАБОРАТОРНА РОБОТА…………………………………………….....19

**1 ПРОСТА ЛІНІЙНА РЕГРЕСІЯ**

***Регресією*** назвемо однобічну стохастичну залежність однієї випадкової змінної від другої або декількох інших випадкових змінних.

Однобічна стохастична залежність виражається за допомогою функції, яка на відміну від сурової математичної залежності, називається ***функцією регресії*.**

Слід зазначити, що функція регресії тільки формально встановлює від-повідність між змінними, хоча вони можуть і не перебувати в причиново-наслі-дкових відношеннях.

***Регресію*** між двома змінними, назвемо ***простою (парною)*** регресією і будемо записувати у вигляді

 (1.1)

Змінну  називають по-різному - ***залежна, результативна, регресанд***, тоб-то, змінна, що підлягає поясненню, а також, завбачувана змінна, або цільова функція. Змінну називають незалежною, пояснюючою або передбачуваною змінною, регресором або факторною ознакою.

Випадкову змінну u назвемо збурюваною (збуренням), якщо 

Зараз ми в змозі сформулювати ***основну задачу регресійного аналізу та******етапи її вирішення.***

**Основна задача регресійного аналізу:**  Встановити вид функції, яка б найкраще описувалаусереднену масову течію економічного процесу.

**I-й етап (попередній):** проведення обґрунтованого якісного економічно-го аналізу досліджуваного процесу; попередня обробка статистичної інформації про цей процес;

**II- й етап**: на основі проведеного аналізу сформулювати гіпотезу про видфункції, яка описує цей процес;

**Ш-й етап:** прийняття гіпотези або відмова від неї шляхом статистичноїперевірки за емпіричними даними;

**ІV-й етап:** визначення кінцевої функції регресії, оцінка невідомихзначень залежної змінної, отриманих теоретично обґрунтованих і статистично надійних різноманітних прогнозів, щодо поведінки досліджуваного процесу.

Перейдемо тепер безпосередньо до класифікації форм регресії, а саме до лінійної регресії.

Лінійною регресією назвемо регресіювиду:

 (1.2)

|  |  |
| --- | --- |
|  |  |
| де – пояснювальні змінні;  – оцінки відповідних |  |
| параметрів регресії  – в лінійній моделі |  |
|  | (1.3) |  |

Тобто

 ****** (1.4)

У випадку парної (простої) регресії вираз (1.2) набуває вигляду:

 (1.5)

де .

Одним з можливих методів отримання оцінок (знаходження числових значень коефіцієнтів  у регресійній моделі) є *метод найменших* *квадратів* (МНК).

**2 МЕТОД НАЙМЕНШИХ КВАДРАТІВ**

Усереднення несумісних розв’язання надлишкової системи рівнянь може бути здійснено різними способами, наприклад, "на око", методом медіанних центрів, методом максимальної правдоподібності і т. ін. Однак одним з найбільш потужних методів є розроблений в 1795-1805 рр. французьким математиком Андрієн М. Лежандром (1752-1833) і німецьким математиком Карлом Фрідрихом Гауссом (1777-1855) метод регресійного аналізу, чи, як часто його називають, метод найменших квадратів (МНК).

Зазначимо, що застосування МНК, якому близько 200 років, гальмувалося через труднощі, пов'язані зі значними обсягами обчислень, і тільки в середині XX століття цей метод набув широкого розповсюдження у зв'язку з появою ЕОМ.

На сьогодні існує досить багато різноманітних програмних продуктів, які дають можливість реалізувати на ЕОМ цей метод. Тому їх застосування настільки різноманітне: статистика, економетрія, оцінка похибок вимірювань тощо.

Для розуміння цього методу розглянемо простий приклад з використанням МНК.

Припустимо, після попереднього аналізу досліджуваного економічного процесу на основі статистичних даних про його поведінку ми вирішили, що математична модель цього процесу має вигляд

 (2.1)

Знайдемо "найкращі" значення параметрів вибраної нами моделі за допомогою МНК. Це нелінійна (квазілінійна) модель – є *.* Зведемо її до лінійної. Для цього покладемо  і *.* У результаті отримаємо лінійну модель (для лінійних моделей існують потужні алгебраїчні засоби для їх дослідження), яка має вигляд

****** (2.2)

Між розрахованими за моделлю значеннями  та експериментальними

відрахуваннями  будуть спостерігатись відхилення. Позначимо їх як

 (2.3)

Далі будемо називати їх *залишками.* Вони включають вплив неврахованих факторів - змінних, випадкових перешкод та помилок спо-стереження тощо. Їхні значення можуть змінюватись від одного спостереження до іншого.

МНК дає змогу знайти такі значення (оцінки)  вихідних параметрів  моделі, для якої (це пов'язано з тим, що необхідний крите-рій для підбору коефіцієнтів моделі повинен враховувати ту обставину, за якої одержана функція регресії (якщо її представити на графіку) буде якомога ближ-че проходити між експериментально отриманими змінними)

** (2.4)

У подальшому таку модель будемо записувати у вигляді

 (2.5)

Беручи частинні похідні по  і прирівнюючи їх до нуля, одержимо систему з трьох рівнянь з трьома невідомими *,*вирішенням якої і є шу-кані значення оцінок. У нашому випадку, оскільки

 (2.6)

 (2.7)

З наведеної вище системи отримуємо наступну СЛАР

 (2.8)

Застосовуючи для розв’язання цієї системи один з відомих методів щодо

розв’язання системи лінійних рівнянь (наприклад, метод Гаусса, або правило

Крамера - метод визначників, або матричний метод) знаходимо невідомі коефі-цієнти 

 (2.9)

 (2.10)

.

Підставляючи їх значення в загальний вигляд регресії отримуємо лінію, яку називають лінією регресії. Коефіцієнти і  називаються **коефіцієнтами регресії**  по  і  по  відповідно.

Знайдена точка  є точкою, яка задовольняє умові (2.5). Але, у математичному аналізі є теорема, за допомогою якої можна ви значити достатні умови екстремуму функції, у нашому випадку – мінімуму функції .

Така сама методика і міркування застосовуються у випадку знаходження коефіцієнтів регресії 

Щоб оцінити щільність зв’язку між x та y , використовують коефіцієнт кореляції та коефіцієнт детермінації, які показують, наскільки варіація змінної x пояснює варіацію y . Щоб оцінити наскільки добре лінія регресії пояснює зв’язок між x та y , використовують стандартну помилку залишків, яка показує відхилення емпіричних значень від лінії регресії.

При побудові регресійної моделі перевіряється гіпотеза про її адекватність. Для цього можна використовувати F -критерій Фішера. Під час оцінки параметрів регресії перевіряються гіпотези, чи статистично значимо вони відрізняються від нуля. Для цього можна використовувати t -тест Стьюдента. Побудовану регресійну модель можна використовувати для прогнозування величини результативної ознаки y при заданому значенні факторної ознаки x , при цьому бажано будувати інтервал довіри для прогнозу

F-тест Фішера для перевірки моделі на адекватність:



Спостережуване значення t- критерію Стьюдента для перевірки значущості коефіцієнта кореляції:



Спостережуване значення t- критерію Стьюдента для перевірки гіпотези 



Спостережуване значення t- критерію Стьюдента для перевірки гіпотези



**3 ПРАКТИЧНА ЧАСТИНА**

**Приклад 1.** На базі статистичних даних про річний продаж фірмою продукції y (тис. од.) та витратами на наукові дослідження x (тис. гр. од.) вибрати форму однофакторної моделі, оцінити всі її параметри, визначити довірчі інтервали при рівні значущості . Оцінити прогноз  y для наступного значення , побудувати довірчий інтервал для прогнозованого значення.



Розв’язання. Побудуємо точковий графік залежності емпіричних значень y від .



Рис 1.

Форма графіка дозволяє зробити припущення про лінійну форму залежності:



Обчислимо середні значення, дисперсії та середні квадратичні відхилення продаж та витрат на наукові дослідження (допоміжні обчислення наведені в таблиці 1):

1. середні значення:

 

1. дисперсії:





1. середні квадратичні відхилення:

.

Таблиця 1.



1. Обчислимо коефіцієнт кореляції:

.

Оскільки значення коефіцієнта кореляції близькі до 1, то між факторним і результативним показниками існує суттєвий прямий зв’язок.

Зміна факторного показника впливає на  на зміну результативного покзника.

1. *Оцінимо параметри  та*  лінійного рівняння регресії  за допомогою методу найменших квадратів.

,

.

 Рівняння регресії має вид:

.

Аналізуючи параметри рівняння регресії, можна зробити висновок: при збільшенні факторного показника (витрат на наукові дослідження) на 1 (1 тис. гр. од.) результативний показник (продаж продукції) збільшиться на 1,8 (1,8 тис. од.); якщо значення факторного показника будуть дорівнювати 0, то значення результативного дорівнюватимуть 554,42.

1. Перевіримо *адекватність побудованої моделі*

Обчислимо дисперсію похибки:

.

Дисперсія похибки не прямує до нуля, але її значення значно менше дисперсії . Це дозволяє висунути гіпотезу про адекватність побудованої моделі.

Застосуємо критерій Фішера, тобто перевіримо нульову гіпотезу  при альтернативній , де  –– нахил узагальненої регресійної моделі.

Таблиця 2.



Обчислимо спостережуване значення  -критерію:

.

За таблицями  -критерію за рівнем значущості  (надійністю 0,95) та числами ступенів свободи 1 і знаходимо .

Оскільки , то побудована нами регресійна модель може вважатись адекватною з ймовірністю 0,95.

1. Перевіримо значущість параметрів та .

Обчислимо середню квадратичну похибку рівняння регресії:



та оцінки дисперсій параметрів  і :

 ; .

Перевіримо нульову гіпотезу  при альтернативній  за критерієм Стьюдента.

Знайдемо спостережувані значення Критерію Стьюдента для параметрів регресії:

; .

Критичне значення .

Оскільки спостережувані значення більші , то нульова гіпотеза  для кожного параметра відкидається, а отже, обчислені значення  та  є статистично значущими.

Побудуємо довірчі інтервали для параметрів рівняння регресії  та :

,

,

що означає:

 .

1. Обчислимо прогноз для значення факторної ознаки :



Побудуємо в одній системі координат графік емпіричної залежності між факторною та результативною ознаками, теоретичну лінію регресії, а також межі довірчого інтервалу для індивідуальних значень  .



Рис 2.

**Приклад 2**. Агент з продажу будинків вивчає залежність між ціною будинку *y* (в $ 1000) і загальною його площею *х* (у сотнях квадратних футів). З цією метою він склав вибірку з 15 будинків і зафіксував такі результати:

Таблиця 1 – Тестова вибірка

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***i*** | ***xi*** | ***yi*** | ***i*** | ***xi*** | ***yi*** |
| **1** | 20.0 | 89.5 | **9** | 24.3 | 119.9 |
| **2** | 14.8 | 79.9 | **10** | 20.2 | 87.6 |
| **3** | 20.5 | 83.1 | **11** | 22.0 | 112.6 |
| **4** | 12.5 | 56.9 | **12** | 19.0 | 120.8 |
| **5** | 18.0 | 66.6 | **13** | 12.3 | 78.5 |
| **6** | 14.3 | 82.5 | **14** | 14.0 | 74.3 |
| **7** | 27.5 | 126.3 | **15** | 16.7 | 74.8 |
| **8** | 16.5 | 79.3 |   |   |   |

Позначивши пари *(xi, yi)* на координатній площині, він отримує так звану *кореляційну хмару*, вигляд якої дозволяє передбачити, що наявна лінійна залежність між змінними. Вона наведена на рис. 1.



Рисунок 1 – Кореляційне поле

Прийнявши цю гіпотезу, він обчислює

а потім за отриманими вище формулами оцінки:

; ; ; ; .

; .

Тепер пряма регресії описується рівнянням:

.

Її графік нанесемо на кореляційне поле (рис. 2.2):



Рис. 2 – Регресійне рівняння