МНОЖИННИЙ РЕГРЕСІЙНИЙ АНАЛІЗ

ДИСПЕРСІЙНИЙ АНАЛІЗ

Множинний регресійний аналіз

Якщо парна регресія досліджує вплив одного фактора (x) на результат (y), то **множинна регресія** дозволяє моделювати залежність від кількох незалежних змінних одночасно.

Загальний вигляд рівняння множинної лінійної регресії:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

де:

- y — залежна змінна (результат)
- $x_1, x_2, \ldots, x_n$ — незалежні змінні
- β_0 — вільний член
- $\beta_1, \ldots, \beta_n$ — коефіцієнти регресії (показують вплив кожного фактора)
- ε — випадкова помилка

Задача: Аналіз продуктивності вебсервера

Проводиться навантажувальне тестування системи. Які фізичні параметри впливають на затримку роботи системи; **прогнозування часу відгуку сервера** залежно від навантаження

y (результат) - час відгуку сервера (мс);
x_1 (навантаження) - кількість активних запитів на секунду;
x_2 (ресурси) - відсоток використання оперативної пам'яті

Тест	Час відгуку (y)	Запити (x1)	Пам'ять (x2)
1	150 мс	100	40%
2	210 мс	200	45%
3	320 мс	300	60%
4	450 мс	500	70%
5	600 мс	800	85%

Рівняння регресії (після розрахунків):

Припустимо, ми розрахували коефіцієнти і отримали:

$$y = 0{,}5 \cdot x_1 + 2{,}0 \cdot x_2 + 50$$

Розшифровка моделі:

50 мс (β_0): базова затримка мережі та ОС навіть при нульовому навантаженні

0,5 (β_1): кожен новий запит на секунду додає в середньому **0,5 мс** до часу відгуку

2,0 (β_2): кожен відсоток завантаження RAM додає **2,0 мс** до затримки

Прогноз. Практичний розрахунок

Яким буде час відгуку, якщо на сервер прийде **600 запитів/сек**, а пам'ять буде завантажена на **75%**?

$$y = 0,5*600 + 2,0*75 + 50$$

$$\boldsymbol{y = 500\ мс.}$$

Дисперсійний аналіз (ANOVA)

Дисперсійний аналіз використовується для перевірки гіпотез про рівність середніх значень у кількох групах.

Дисперсійний аналіз застосовується для дослідження впливу однієї або декількох якісних змінних (факторів) на одну залежну кількісну змінну

Порівнюючи компоненти дисперсії за допомогою F-критерію Фішера, можна визначити, наскільки результативна ознака зумовлена дією регульованих факторів

В комп'ютерних науках це часто застосовується для порівняння ефективності різних алгоритмів або конфігурацій систем

Задача: дослідження впливу технології віртуалізації на продуктивність бази даних

Дослідити вплив **технології віртуалізації** (фактор А) на **швидкість виконання складних SQL-запитів** (числовий показник у, вимірюється в секундах).

Є три різні технології (рівні фактору):

Технологія 1: Bare Metal (без віртуалізації, прямий доступ до заліза).

Технологія 2: Docker-контейнеризація.

Технологія 3: Повна віртуалізація (VMware).

Мета дисперсійного аналізу (ANOVA):

Визначити, чи є різниця в швидкості обробки даних між цими технологіями статистично значущою, чи спостережувані відхилення є результатом випадкових шумів (наприклад, мережевих затримок під час конкретного тесту).

Алгоритм ANOVA

Технологія 1 (Bare Metal)	Технологія 2 (Docker)	Технологія 3 (VMware)
10,2 сек	11,5 сек	14,1 сек
9,8 сек	12,1 сек	13,8 сек
10,5 сек	11,8 сек	14,5 сек
10,1 сек	12,0 сек	13,9 сек
9,9 сек	11,6 сек	14,2 сек

1) Обчислимо загальне середнє, середній час для всіх 15 тестів.

2) Обчислимо внутрішньогрупову дисперсію, наскільки дані «розпилюються» всередині одного Docker-сегмента («шум»).

3) Обчислимо міжгрупову дисперсію, наскільки сильно середнє значення Docker відрізняється від Bare Metal чи VMware

4) Знайдемо F-критерій:

F = Дисперсія між технологіями / Випадкова дисперсія (помилка)

Якщо **F_обчислений > F_критичний**, то **Технологія обробки (віртуалізація)** ***суттєво*** **впливає на продуктивність системи.**

Якщо ні - різниця несуттєва, і можна обирати будь-яку технологію, виходячи з інших міркувань

$n_1 := 5$ $x_{1,1} := 10.2$ $x_{1,2} := 9.8$ $x_{1,3} := 10.5$ $x_{1,4} := 10.1$ $x_{1,5} := 9.9$

$n_2 := 5$ $x_{2,1} := 11.5$ $x_{2,2} := 12.1$ $x_{2,3} := 11.8$ $x_{2,4} := 12.0$ $x_{2,5} := 11.6$

$n_3 := 5$ $x_{3,1} := 14.1$ $x_{3,2} := 13.8$ $x_{3,3} := 14.5$ $x_{3,4} := 13.9$ $x_{3,5} := 14.2$

Кількість технологій

$m := 3$ $N := \sum_{i=1}^{m} n_i$ $N = 15$ $i := 1..m$

$$X_i := \frac{\sum_{j=1}^{n_i} x_{i,j}}{n_i}$$

$$s1 := \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_{i,j} - X_i\right)^2 \qquad s1 = 0.86$$

$$XN := \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{i,j}}{N}$$

$$s := \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_{i,j} - XN\right)^2 \qquad s = 41.16$$

$\alpha := 0.05$ $x\alpha := qF(1 - \alpha, N - 1, N - m)$ $x\alpha = 2.637$ F критичне (табличне)

$$FN := \frac{s \cdot (N - m)}{s1 \cdot (N - 1)}$$

$FN = 41.023$ F теоретичне (обчислене)

Коэфициент детерминации

$$r2 := \frac{s - s1}{s} = 0.979$$

$a_i := X_i$ $a_i =$

10.1
11.8
14.1

$$\sigma := \frac{s1}{N - m} = 0.072$$

$\sqrt{\sigma} = 0.268$

Результат

В результаті обчислень **Fобчислений = 41.023**, а **Fкритичний = 2.637**, тобто

Fобчислений > Fкритичний

Таким чином, **гіпотеза H0 відхиляється** і можна зробити висновок що з **ймовірністю 95%** встановлено, що фактор (технологія віртуалізації) має статистично значущий вплив на продуктивність системи.

В розглянутому прикладі коефіцієнт детермінації **r = 0.979**, тобто можна сказати, що на **97.9%** результат ознаки X (продуктивність системи) обумовлений саме фактором, який на неї впливає (технологія віртуалізації), а не випадковими чинниками. Лише **2.1%** - це частка випадкових похибок (шуму)

Найбільший вплив зафіксовано у третьої технології (**середнє значення 14.1**), що дозволяє рекомендувати її для використання в першу чергу