

РЕГРЕСІЙНИЙ АНАЛІЗ

Змістовий модуль 6

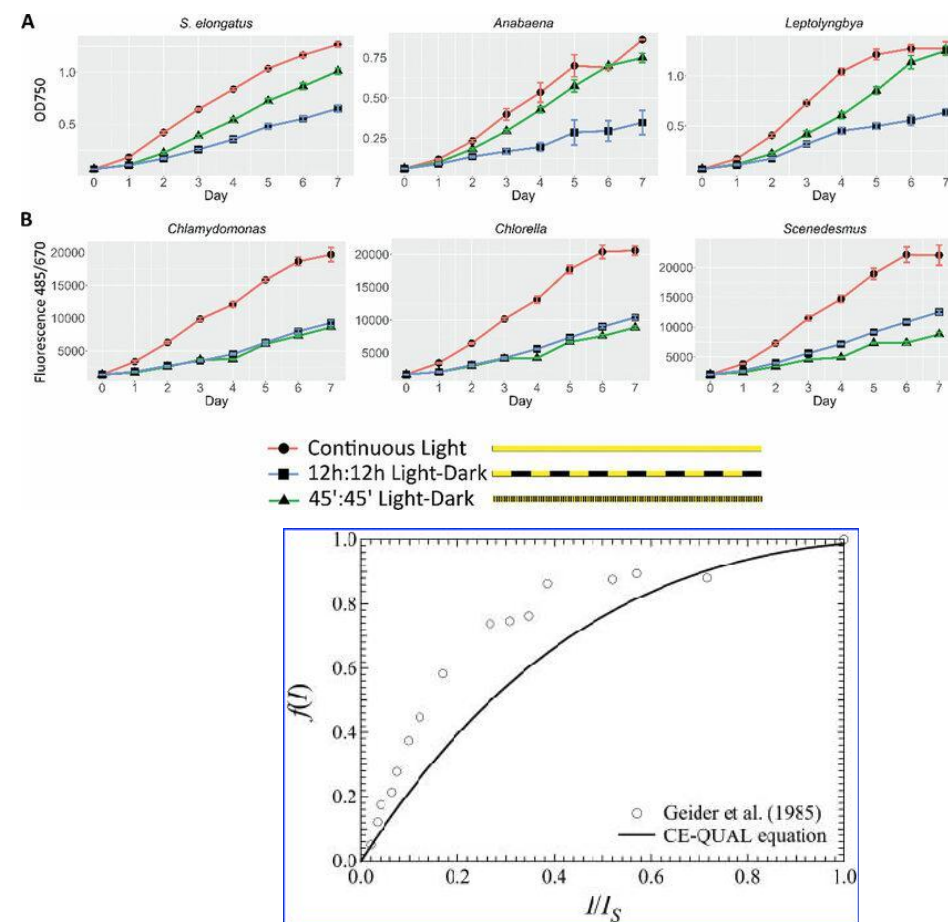
1. Застосування регресійного аналізу

2. Регресійний аналіз: основні положення

Регресійний аналіз — це математичний метод, який дозволяє зрозуміти, як одна змінна (наприклад, ріст рослини) залежить від інших змінних (наприклад, кількості світла, води чи добрив). Це як лінійка для вимірювання зв'язку між речами: він наочно демонструє, наскільки одна змінна змінюється, коли змінюється інша.

Наприклад, регресійний аналіз може дозволити визначити, як зміна освітлення впливає на кількість одноклітинних водоростей у водоймі. Зібравши дані про інтенсивність освітлення та кількість одноклітинних водоростей, можна застосувати регресійний аналіз, щоб побачити, чи є між ними зв'язок і наскільки він сильний.

Результатом може бути, наприклад, пряма лінія на графіку, яка покаже, як зазвичай кількість водоростей змінюється зі зростанням чи зниженням інтенсивності освітлення.



Кореляційний і регресійний аналізи мають схожість у тому, що обидва використовуються для оцінки зв'язку між змінними, але різняться за цілями та тим, яку інформацію вони надають:

Кореляційний аналіз:

Мета: визначити, чи існує зв'язок між двома змінними, і оцінити його силу та напрямок.

Результат: коефіцієнт кореляції (значення від -1 до 1), який показує, наскільки змінні "рухаються разом". Наприклад, якщо коефіцієнт кореляції близький до 1, це означає сильний прямий зв'язок (зростання однієї змінної супроводжується зростанням іншої), а якщо близький до -1 — сильний зворотний зв'язок (зростання однієї змінної супроводжується зниженням іншої).

Обмеження: кореляція не показує причинно-наслідкових зв'язків і не враховує, яка змінна є залежною, а яка незалежною. Вона лише вказує на зв'язок, але не пояснює його.

Регресійний аналіз:

Мета: дослідити, як одна або кілька незалежних змінних впливають на залежну змінну та побудувати математичну модель цього впливу.

Результат: рівняння регресії, яке показує, наскільки змінюється залежна змінна при зміні незалежних змінних. Наприклад, регресійне рівняння може показати, що збільшення кількості годин освітлення призводить до зростання висоти рослини на певну кількість сантиметрів.

Особливість: регресія дозволяє робити прогнози. Вона показує, наскільки зміниться залежна змінна при зміні незалежної змінної, а також враховує напрямок впливу, що робить її корисною для аналізу причинно-наслідкових зв'язків.

Підсумок: кореляція показує тільки силу та напрямок зв'язку між змінними, а регресія — залежність і можливий вплив однієї змінної на іншу, дозволяючи робити прогнози.

Регресійний аналіз у біології широко застосовується для визначення залежностей між біологічними змінними та передбачення результатів на основі цих залежностей. Наприклад:

Аналіз впливу факторів навколишнього середовища на ріст рослин. Регресійний аналіз допомагає оцінити, як змінюються параметри росту (наприклад, висота, біомаса) залежно від таких факторів, як температура, вологість ґрунту, освітлення тощо.

Вивчення зв'язку між біомаркерами та ризиком захворювань. У дослідженнях біомедичних даних часто використовують регресійні моделі для передбачення ризику захворювань на основі концентрацій певних речовин у крові чи інших біологічних зразках.

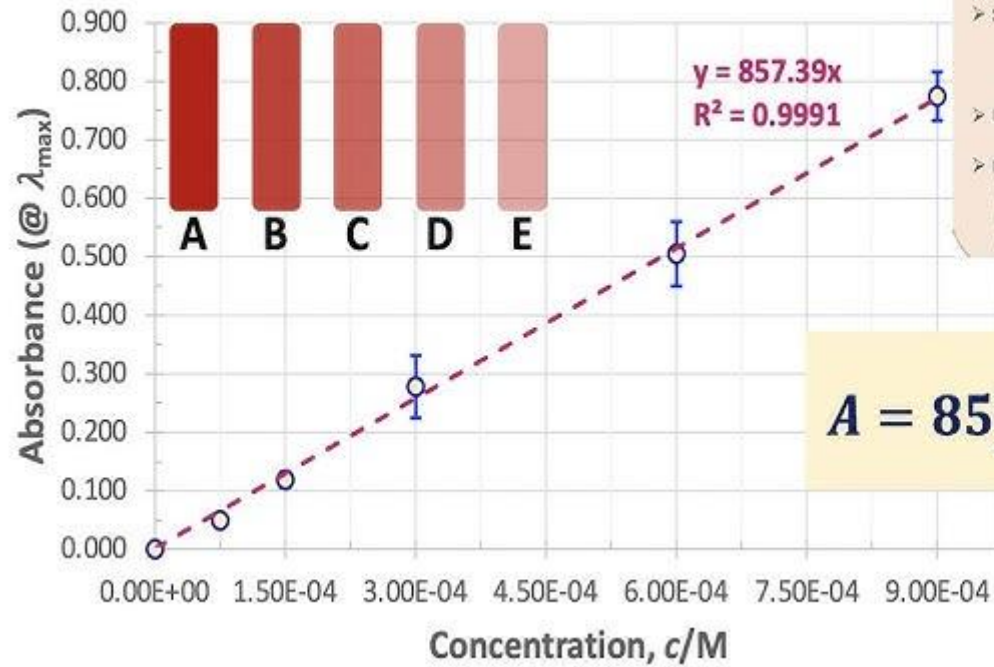
Моделювання популяційної динаміки. У популяційній екології регресія допомагає виявити фактори, які впливають на чисельність і виживання популяцій різних видів (наприклад, вплив харчових ресурсів або хижаків на чисельність).

Оцінка впливу генетичних факторів на фенотипові характеристики. Регресійний аналіз дозволяє визначити, які гени або їхні варіанти найбільше асоціюються з певними фенотиповими ознаками (наприклад, висота у людей, колір шерсті у тварин).

Визначення залежності між метаболічними показниками та фізичною активністю. У фізіології та спортивній біології часто аналізують залежність між рівнем фізичної активності й такими показниками, як частота серцевих скорочень, рівень глюкози, рівень ліпідів та ін.

Прогнозування розвитку захворювань у тварин. Для ветеринарної медицини регресійні моделі використовуються, наприклад, для оцінки ризику інфекційних або хронічних захворювань у тварин залежно від умов утримання, раціону або спадкових факторів.

Standard Curve Dye Solutions



Preparation of a Standard Curve

- > Background Corrected measurements
- > Statistics
 - Avg, Std Deviation
 - Error bars
 - Linear regression – {slope, intercept} errors
- > Use linear eqn to determine trends
- > Extra tips (Excel)
 - Functions: Absolute reference (\$); LINEST()
 - Use of tables (ease calculations)
 - Adding error bars to your plot

$$A = 857.4 \left[\frac{1}{M} \right] \cdot c [M]$$

Цілі регресійного аналізу

1. Прогнозування значення залежної змінної за допомогою незалежних змінних.
2. Визначення внеску окремих незалежних змінних у варіацію залежної змінної.
3. Регресійний аналіз не можна використовувати для визначення наявності зв'язку між змінними, оскільки наявність такого зв'язку і є передумовою для застосування цього виду аналізу.

Для того щоб застосувати регресійний аналіз і отримати достовірні результати, важливо дотримуватися кількох ключових умов. Основні умови для використання регресійного аналізу включають:

Лінійність:

Зв'язок між незалежною та залежною змінними повинен бути лінійним (тобто можна приблизно описати прямою або іншою простою функцією). Це можна перевірити за допомогою графіка (наприклад, побудувати діаграму розсіювання) або використовувати інші статистичні тести.

Нормальність залишків:

Різниця між фактичними значеннями залежної змінної та прогнозованими значеннями (залишки) має розподілятися за нормальним законом. Це можна перевірити через графік (наприклад, гістограму) або за допомогою тестів на нормальність.

Гомоскедастичність:

Залишки повинні мати постійну дисперсію по всьому діапазону незалежної змінної (тобто, варіативність залишків не повинна змінюватися зі зміною рівня незалежної змінної). Це можна перевірити за допомогою діаграми залишків: якщо дисперсія залишків стає більшою або меншою для різних значень незалежної змінної, то це порушує гомоскедастичність.

Відсутність автокореляції:

Значення залишків не повинні бути пов'язані між собою. Це означає, що залишки мають бути випадковими, без послідовності чи закономірностей. Для перевірки цієї умови використовують, наприклад, тест Дарбіна-Ватсона.

Відсутність мультиколінеарності (у випадку множинної регресії):

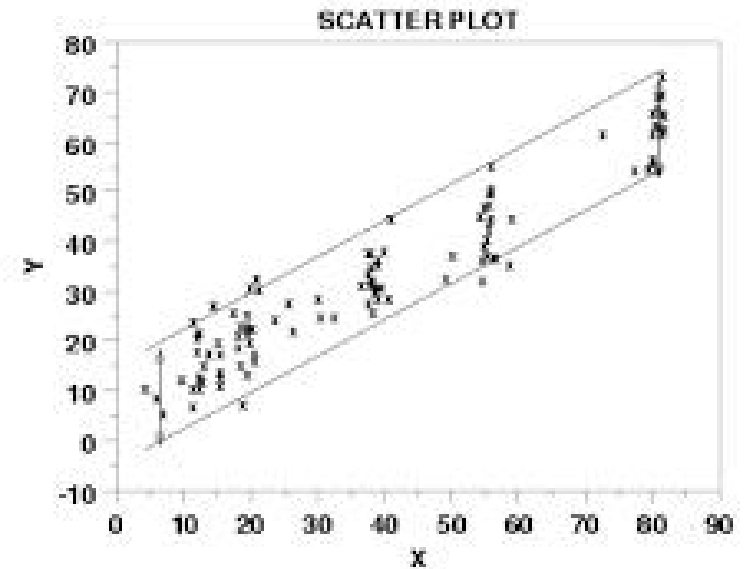
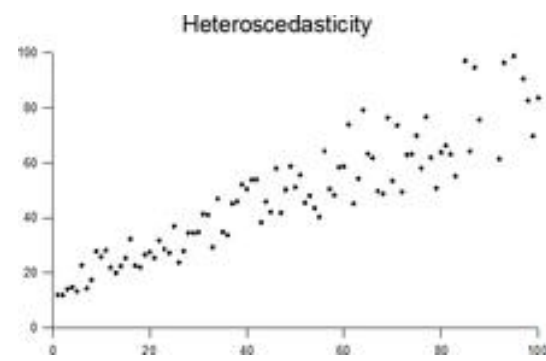
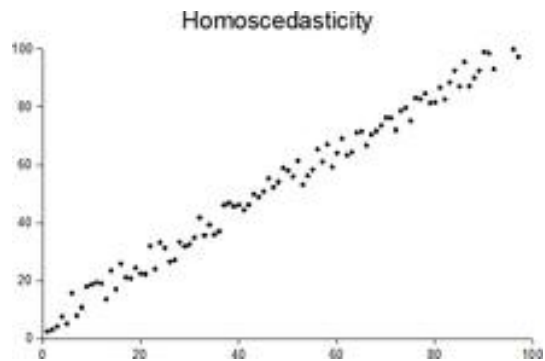
У моделі з декількома незалежними змінними важливо, щоб ці змінні не були сильно взаємопов'язані, оскільки це може призвести до нестабільності оцінок коефіцієнтів. Мультиколінеарність можна перевірити за допомогою коефіцієнта кореляції або фактору інфляції дисперсії (VIF).

Незалежність спостережень:

Кожне спостереження має бути незалежним від інших. Це означає, що значення однієї спостережуваної одиниці не повинні залежати від значень іншої. Незалежність спостережень є особливо важливою у вибірках, які збираються у часовому або просторовому контекстах.

Дотримання цих умов забезпечує коректність результатів регресійного аналізу і робить можливим його використання для обґрунтованих висновків та прогнозів.

Одним із важливих припущень у регресійному аналізі є припущення про **гомоскедактичність** - тобто, для кожного значення незалежної змінної розкид значень залежної має бути однаковим і описуватися нормальним розподілом.



Регресійни аналіз: основні положення

Регресійний аналіз - це інструмент для кількісного визначення значення однієї змінної на підставі іншої.

Парна (проста) лінійна регресія дає нам правила, що визначають лінію регресії, яка краще за інших прогнозує найімовірніші значення однієї змінної на підставі іншої (змінних усього дві).

Множинна регресія є розширенням простої лінійної регресії.

По осі Y розташовують змінну, яку необхідно передбачити (залежну), а по осі X - змінну, на основі якої буде здійснюватися передбачення (незалежну).

Залежна змінна - це змінна в регресії, яку не можна змінювати, її зміна є наслідком впливу незалежної змінної (змінних).

Незалежна змінна - це та змінна в регресії, яку можна змінювати.

Коефіцієнти регресії (β) - це коефіцієнти, які розраховуються в результаті виконання регресійного аналізу.

Обчислюються величини для кожної незалежної змінної, які представляють силу і тип взаємозв'язку незалежної змінної по відношенню до залежної.

Регресійний аналіз застосовують у кількісному аналізі при побудові градування, коли виникає необхідність знаходження градувальної функції $y = f(x)$ експериментальним шляхом.

Для цього вимірюють **аналітичний сигнал (y)** серії зразків порівняння з вмістом **досліджуваного компонента (x)**, унаслідок чого отримують масив даних (x_i, y_j) . Визначення градувальної функції у регресійному аналізі зводиться до знаходження рівняння регресії, тобто математичної форми залежності вимірюваної фізичної величини від факторів, що впливають на неї.

Вид рівняння вибирається із зовнішньої інформації та загальних міркувань відносно фізичних та хімічних законів, які поєднують аналітичний сигнал (y) з вмістом досліджуваного компонента (x). Найчастіше використовується лінійна залежність:

$$y = ax + b$$

$$y = ax + b$$

У цьому разі

x — незалежна змінна,

y — залежна змінна,

величини a та b — коефіцієнти рівняння регресії,

Якщо похибки невеликі, а діапазон, у якому знаходяться результати вимірювань достатньо широкий, застосовують графічний метод. Його основний недолік — неможливість оцінити похибки коефіцієнтів a та b .

Для оцінки похибок величин a та b найбільш простим є метод Тейлора. У цьому разі знаходять коефіцієнти b_{ij} прямої, яку можна провести через кожну пару експериментальних точок (x_i, y_i) , та розраховують медіану значень b_{ij} прямої — b , а потім визначають для кожної прямої коефіцієнт a_i і розраховують медіану a .

Частіше у регресійному аналізі застосовують **метод найменших квадратів** для знаходження коефіцієнтів a і b та розрахунків їх довірчих інтервалів. Після визначення коефіцієнтів a і b та оцінки їх похибок перевіряють адекватність самого рівняння регресії, потім проводять оцінку значущості відхилень експериментальних точок від прямої лінії Δy і ділянку допустимих значень y (регресійна смуга).

Суть методу найменших квадратів

Нехай є n спостережень ознак x і y . Причому відомий вид рівняння регресії - $f(x)$, наприклад, прямолінійна залежність:

$$f(x_i) = a + b \cdot x_i$$

Необхідно підібрати такі значення параметрів (a і b), які зможуть мінімізувати суму квадратів відхилень фактичних значень ознаки-результату y_i від розрахункових (теоретичних) значень $f(x_i)$ для всіх спостережень $i=1:n$

$$S = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 \Rightarrow \min_{a,b}$$

Оцінка якості рівняння регресії і коефіцієнт детермінації

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (Y_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

Коефіцієнт множинної детермінації R-квадрат показує, яку частку мінливості (можна виразити у відсотках) залежної змінної (Y) пояснює незалежна змінна (регресійна модель).

- **Під якістю рівняння регресії** розуміється ступінь близькості (відповідності) розрахованих за цим рівнянням значень ознаки результату $f(x)$ фактичним (спостережуваним) значенням y .
- Чим ближче R-квадрат до 1, тим вища якість регресійної моделі.