Лекция 8

Многопроменлива линейна регресия

Линейните модели се използват повсевместно в науката. В някои от случаите зависимостта между величините е линейна, поради характера на закономерностите, които свързват величините. В други случаи линейната регресия е удобно и достоверно приближение на зависимости, чийто характер е неизвестен. Но и в двата случая тълкуването на модела е недвусмислено и директно, като позволява да се отчете влиянието на различните фактори (променливи). А това влияние е в основата на приближени теории за съответните (физико)химични процеси. Разбира се, когато нелинейността на връзките между величините не може да се пренебрегне, в ход влизат нелинейните модели, които обясняват количествените връзки много по-точно и имат по-добра предсказателна способност.

8.1. Линейна еднопроменлива регресия. С въпроса за вида на връзката между две величини се занимава регресионният анализ. Получената зависимост (на практика функция), изразена чрез уравнение (8.1) се нарича линия на регресия или само регресия.

$$У = b_1 X_к + b_0, \qquad (8.1)$$

където b_1 се нарича наклон (slope), а b_0 - отрез (intercept).

В общия случай зависимостта между величините X и У може да е крива, описвана с уравнение от степен n = 1 (различна от права) или трансцедентна крива (например $У = ae^{bX}$).

Задачата, която се поставя, е намирането на стойности на параметрите b_1 и b_0 в уравнение (8.1), така че кривата максимално да се доближава до експерименталните точки (Хк,Ук); k = 1, 2, … M. Параметрите, които трябва

да се определят, могат да бъдат повече от два, например при крива от четвърта степен те са пет b_0, b_1, b_2, b_3 и b_4 – уравнение (8.2).

$$У = b_0 + b_1X + b_2X^2 + b_3X^3 + b_4X^4 \quad (8.2)$$

Един от критериите за "максимално доближаване" на кривата до експерименталните точки е сумата S, която се дава с уравнение (8.3)

$$S = \Sigma (У_к - У'_к)^2, \quad (8.3)$$

където $У'_к$ е изчислената стойност на У при заместване на $X_к$ в уравнението на кривата, а $У_к$ е експерименталната стойност, съответстваща на $X_к$. Методът основаващ се на този критерий, се нарича метод на най-малките квадрати и изчислителната му страна е разгледана в много учебни пособия [1] - вижте задача 8.1.

Линейната еднопроменлива регресия се използва за построяване на калибрационни криви сигнал/концентрация с данните от количествени измервания на различни апарати. Ако сигналът е абсорбция на пробата при дадена дължина на вълната, то регресионното уравнение свързва абсорбцията на пробата и концентрацията на аналита. В този случай концентрацията на веществото е неслучайна величина, която се изменя съзнателно от експериментатора, с цел определяне на наличието на зависимост между величините C и A: на практика концентрацията на стандартните разтвори е случайна величина, но нейната невъзпроизводимост е много по-малка от тази на измерваните стойности на абсорбцията. Едно количествено характеризиране на тази зависимост, в смисъл на "ако се повиши C с dC, с колко ще се измени A", би имало практическа стойност, извън тази на теоретичните изводи от връзката между C и A. На практика се установява, че една линейна връзка $A = a_1.C + a_0$ достатъчно добре характеризира опитните данни те са точки с координати

($C_к$,$A_к$), разположени близо до споменатата права – отклоненията се дължат на случайните грешки при измерване на абсорбцията. При корекция (нулиране) на сигнала на празната проба отрезът b_0 в (8.1) е нула и се получава известното уравнение на Буге-Ламбер-Беер.

8.2. Многопроменлива регресия [2]. Нейният математически израз се дава с уравнение (8.4).

$$У = b_0 + \Sigma\, b_k X_k \qquad (8.4)$$

където b_0 се нарича отрез, а коефициентите b_k , k = 1, 2, … M, дават влиянието на отделните променливи X_k върху зависимата променлива У и се наричат регресионни коефициенти (отрезът също е регресионен коефициент, но се използва краткото име за да се отличава от b_k , k = 1, 2, … N).

При четири независими променливи уравнението има вида

$$У = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 \qquad (8.5)$$

Обърнете внимание на приликата между (8.2) и (8.5) – на практика, ако се използват матрици изчисленията на регресионните коефициенти и отреза ще бъдат подобни.

Както споменахме в предишната лекцшия, в науката се поставят задачи от вида на матричното уравнение (8.6), с които се цели да се изчислят величите b_m, m = 1,2,… M, чрез стойностите на други величини x_k, k = 1,2,… K.

$$\begin{pmatrix} a_{1,2} & a_{1,2} & \dots & a_{1,K} \\ a_{2,1} & a_{2,2} & \dots & a_{2,K} \\ \dots & \dots & \dots & \dots \\ a_{M,1} & a_{N,2} & \dots & a_{M,K} \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_K \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_M \end{pmatrix} \qquad (8.6)$$

Реално линейните уравнения в (8.6) описват някакъв ***линеен модел*** и този модел е необходимо да описва възможно най-голям брой експериментални

данни с малко на брой параметри (в случая M на брой), за да има този модел своята предсказателна и обяснителна ценност. Т.е. на практика N > M и това означава, че системата няма точно решение, в смисъл, че няма такава комбинация от M параметъра, които да описват с произволна точност тези N уравнения. Но е напълно възможно системата да има решение, което да удовлетворява приблизително N-те уравнения, като по този начин това решение представлява необходимите параметри в линейния модел, които обясняват възможно най-голям брой експериментални данни.

Решението на матричното уравнение (1), AX = B, може да се получи ако уравнението се умножи отляво с израза $(A^TA)^{-1}A^T$

$$(A^TA)^{-1}(A^TA)X = (A^TA)^{-1}A^TB,$$

което води до

$$I\,X = (A^TA)^{-1}A^TB,$$

т.е.

$$X = (A^T*A)^{-1}A^TB \qquad (8.7)$$

Така получения набор от параметри – матрицата-колона X (т.е. наборът от неизвестните x_k) параметризира линейният модел, с който се описва съответното природно явление. Тук означенията са малко объркващи за начинаещия читател – коефициентите на регресия, които се търсят са означени с матрицата колона X, данните които се измерват – с матрицата A, и променливата, която се предсказва с линейния модел – с матрицата колона B.

Обърнете внимание, че в уравнения (8.6) не фигурира отрез на регресията. При отрез изчисленията са подобни – просто в матрицата A първата колона е само от единици и, разбира се, броят на колоните в нея е с единица по-голям от броя на независимите променливи.

8.3. Работа със статистически програми. Статистическите професионални програми като **STATISTICA**, **SPSS** и **SAS** поддържат изчисляването на линейна многопроменлива регресия, както и на различни нелинейни модели. Работа с тях е изключително лесна, ако човек е запознат с ръководството на програмата и знае статистика на средно ниво, но ако това не е така, потребителят може да допусне редица грешки.

1. Включване/изключване на отрез в/от регресията: това е най-критичния момент при задаване на модела. Ако вашите данни изискват отрез, то изчисляването с модел без отрез ще доведе до грешни коефициенти на регресия. И обратно – ако няма отрез, както е задачата за многокомпонентния анализ (лекция 14), то изборът на работа с отрез ще доведе до неверни резултати. В програмата **STATISTICA** отрезът е наименован **intercept** и в менюто за **multiple regression** има checkbox „**No intercept**".

2. Избор на променливи. При грешен избор на променливите, програмата ще извърши изчисленията без да „възрази", но резултати ще са за друга, а не желаната регресия. В различните програми се появяват различни прозорци, с различни наименования. Потребителят е длъжен внимателно да прочете, кой прозорец за кои променливи е. В регресията има независими променливи, $\mathbf{x_k}$, които на български понякога се наричат и входни променливи. Променливата **y** се нарича зависима променлива, а понякога изходна променлива. В програмата **STATISTICA** тези променливи се наричат, съответно, **dependent variable** (**y**) и **predictor variables** ($\mathbf{x_k}$).

Има и други „подводни камъни", но те са извън материала на тази лекция. В заключение само ще спомена, че *образованият потребител* е този потребител, който минимум е (1) запознат с основите на статистиката, (2) чете научна литература по проблема, който решава, (3) чете частта от ръководството на статистическата програма, която засяга неговия проблем,

и (4) критично осмисля и проверява получените резултати. Вероятността образованият потребител на статистически софтуер да получи грешни резултати е много, ама много по-малка от тази, свързана с работата на полуграмотен специалист.

Литература

1. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michote, L. Kaufman; *Chemometrics: A Textbook*. Elsevier, Amsterdam, 1988.
2. М. А. Шараф, Л. Иллмэн, Б.Р. Ковальски; *Хемометрика.* Химия, Ленинград, 1989.

Въпроси и задачи

Задача 8.5*. Матричните уравнения (8.7) могат да се изведат по метода на най-малките квадрати, подобно на изчисленията в задачи 8.1, 8.2 и 8.3. Сумата, която се минимизира е:

$$S = \Sigma (Y_m - b_0 - \Sigma b_k X_{m,k})^2, \qquad (8.5\text{-I})$$

където първата сума е по m (номерът на измерването) от 1 до M (броят измервания, т.е. броят точки, по които се строи регресията), втората сума – по к (номерът на променливата) от 1 до K (броят на променливите). В (8.5-I) $X_{m,k}$ е стойността на к-тата променлива при m-тото измерване.

Нека допуснем, че отрезът е нула: при такова условие са получени формули (8.7). Тогава сумата, която трябва да се минимизира става

$$S = \Sigma (Y_m - \Sigma b_k X_{m,k})^2, \qquad (8.5\text{-II})$$

8.5.1. Намерете първата производна на тази сума по b_s. (Нарочно избираме индекс с друга буква, за да го различаваме от индекса **k** във втората сума!)

8.5.2. Приравнете първите производни на тази сума по различните коефициенти b_s на нула. Напишете системата от K уравнения в матричен вид.