Семинар 8

Многопроменлива линейна регресия

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) $X_{k,m}$ 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 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.

Решение на 8.5.1: Отново използваме, че производна на сума е сума от производните и производна от функция на втора степен е две, умножено по функцията и по нейната производна. Първата производна по b_s е равна на:

$$\frac{\partial S}{\partial b_s} = \frac{\partial \sum_{m=1}^{M} (Y_m - \sum_{k=1}^{K} b_k X_{m,k})^2}{\partial b_s} = \sum_{m=1}^{M} \frac{\partial (Y_m - \sum_{k=1}^{K} b_k X_{m,k})^2}{\partial b_s} =$$

$$= \sum_{m=1}^{M} 2(Y_m - \sum_{k=1}^{K} b_k X_{m,k}) \frac{\partial (Y_m - \sum_{k=1}^{K} b_k X_{m,k})}{\partial b_s} = \sum_{m=1}^{M} 2(Y_m - \sum_{k=1}^{K} b_k X_{m,k})(-X_{m,s}) =$$

$$(-2)[\sum_{m=1}^{M} Y_m X_{m,k} - \sum_{m=1}^{M} \sum_{k=1}^{K} b_k X_{m,k} X_{m,s}] = 2[\sum_{m=1}^{M} \sum_{k=1}^{K} b_k X_{m,k} X_{s,m} - \sum_{k=1}^{M} Y_m X_{m,s}] =$$

$$2[\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,s} - \sum_{k=1}^{M} Y_m X_{m,s}]$$

В горното равенство при последното преобразуване сменихме местата на сумите по k и тази по m – това е така, защото няма значение дали една правоъгълна таблица от числа ще я сумирате първо по колони, после по редове или първо по редове, после по колони.

Решение на 8.5.2: Използвайки резултатите от предната точка получаваме:

$$\frac{\partial S}{\partial b_1} = 2[\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,1} - \sum_{k=1}^{M} Y_m X_{m,1}] = 0$$

$$\frac{\partial S}{\partial b_1} = 2[\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,2} - \sum_{k=1}^{M} Y_m X_{m,2}] = 0$$

. . .

$$\frac{\partial S}{\partial b_1} = 2[\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,K} - \sum_{k=1}^{M} Y_m X_{m,K}] = 0$$

Ако разделим всяко уравнение на 2 и пренесем сумите, в които няма коефициенти b_k отдясно ще получим:

$$\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,1} = \sum_{k=1}^{M} Y_m X_{m,1}$$

$$\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,2} = \sum_{k=1}^{M} Y_m X_{m,2} \qquad \text{(8.5-III)}$$

. . .

$$\sum_{k=1}^{K} b_k \sum_{m=1}^{M} X_{m,k} X_{m,K} = \sum_{k=1}^{M} Y_m X_{m,K}$$

Нека разгледаме коефициентите пред b_k и свободните членове. В уравнение (8.6) на лекция 8 първите са записани с означение $a_{m,k}$ (m = 1,2,… M и k = 1,2,… K) и b_m. Ако сравним матричния запис в (8.6) с последната система от

уравнения (8.5-III), ще видим, че (8.5-III) е система от **K** уравнения с **K** неизвестни – това е „класическият случай", в който броят на неизвестните е равен на броя на уравненията: не така стои въпросът в (8.6), където имахме повече уравнения (M на брой), отколкото неизвестни (K на брой). Тази система от уравнения (8.5-III) подобно на (8.6) се записва матрично като $\mathbf{A'X'} = \mathbf{B'}$. За коефициентите пред неизвестните $\mathbf{A'}$, свободните членове $\mathbf{B'}$ и самите неизвестни $\mathbf{X'}$ са равни[1] на:

$$a'_{l,k} = \sum_{m=1}^{M} X_{m,k} X_{m,l} \quad \text{и} \quad b'_{l} = \sum_{k=1}^{M} Y_m X_{m,l} \quad \text{и} \quad X'_{k} = b_k \qquad \text{(8-IV)}$$

Първото уравнение показва, че матрицата $\mathbf{A'}$ от коефициентите пред неизвестните е $\mathbf{A'} = \mathbf{X^T X}$, а матрицата-стълб $\mathbf{B'}$ от свободните членове е $\mathbf{B'} = \mathbf{X^T Y}$, докато неизвестните X' са коефициентите на регресия B. Понеже броят на неизвестните е равен на броя на уравненията и матрицата на коефициентите пред неизвестните е квадратна, тази система може да бъде решена, чрез обръщане на матрицата пред неизвестните (ако тя има ранг, равен на K):

$$\mathbf{X'} = \mathbf{A'}^{-1}\mathbf{B'} \qquad \text{(8-V)}$$

Ако заместим (8-IV) в (8-V) получаваме

$$\mathbf{B} = \mathbf{X'} = \mathbf{A'}^{-1}\mathbf{B'} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y},$$

т.е.

$$\mathbf{B} = \mathbf{X'} = \mathbf{A'}^{-1}\mathbf{B'} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} \qquad \text{(8-VI)}$$

Ако сравним решенията (8-VI) с (8.7) от лекция 8, то ще видим, че те са еднакви. Така доказахме, че матричното решение (8.7) на системата уравнения (8.6) отговаря на изискването на метода на най-малките квадрати – сумата (8.3) да е минимална. Понеже броят на уравненията е по-голям от

[1] В случая е малко объркващо като спазваме традицията с означенията на матриците с буквите **A**, **B** и **X**, затова сме ги означили с прим.

броя на неизвестните, то системата няма точно решение[2]. В случая, когато уравненията са повече от неизвестните, решението е „най-доброто" в смисъла на това, че ако заместим намерените неизвестни в лявата част на системата от уравнения, то ще получим различна дясната част, но с много близки стойности до експерименталните, които са толкова близки, че сумата от квадратите на отклоненията им (8.3) е минимална – т.е. те са „най-близки".

Практически задачи

Задача C1. Стартирайте програмата **STATISTICA**. Копирайте данните от задача **C1** на семинар 5 в програмата и приложете към тях линейна регресия на една променлива. Не забравяйте да проверите, че checkbox-а **No Intercept** не е избран – т.е. да се прилага уравнение с отрез. Съвпадат ли изчислените стойности с тези от програмата **Excel**?

Ако не можете да се справите с изчисленията, на поддиректорията **stats** са дадени два файла - **sem05c01.sta** и **sem05c01.stw**, в които са данните и резултатите от регресията.

Задача C2. Дипломант от катедра „Физикохимия" в ХФ на ПУ изследва във воден разтвор скоростта на реакцията:

$$A + 2\ B \rightarrow P$$

Той подозира, че скоростта зависи по следния начин $v = k[A][B]^2$. Според теорията, която е учил, порядъкът на реакцията по двете вещества (степените на концентрацията) и скоростната константа се определят по следния начин:

[2] Точно решение означава, че ако заместим неизвестните в лявата част на системата от уравнения, то ще получим точните стойности на дясната част (свободните членове): при по-голям брой уравнения, отколкото неизвестни, при това заместване се получават стойности на свободните членове, близки до оригиналните.

1. Провежда се реакцията при излишък на веществото **B** (много голяма начална концентрация $[B]_0$), и тогава неговата концентрация се променя незначително и на практика скоростта на реакцията е $V = k_A[A]$; $k_A = k[B]_0^2$. От получените данни се определя порядъкът на реакцията по веществото **A**.

2. След това се провежда реакцията при излишък на веществото **A** (много голяма начална концентрация $[A]_0$), и тогава неговата концентрация се променя незначително и на практика скоростта на реакцията е $V = k_B[B]$; $k_B = k[A]_0$. От получените данни се определя порядъкът на реакцията по веществото **B**.

3. От k_A и $[B]_0$ или k_B и $[A]_0$ може да се изчисли k:

$$k = k_B / [A]_0 \text{ или } k = k_A / [B]_0^2$$

Проблемът, с който се сблъсква дипломантът е, че веществата **A** и **B** са малко разтворими във вода и не могат да се приготвят разтвори с техни големи начални концентрации и да се получи излишък на едно от веществата в разтвора. Затова той решава да приложи многопроменлива линейна регресия към зависимостта на началните скорости от концентрацията на двете вещества.

Ако скоростта на една реакция се дава като $V = k[A]^x[B]^y$, то при логаритмуване на това уравнение се получава:

$$\ln(V) = \ln(k) + x\ln([A]) + y\ln([B])$$

Това е линейна зависимост на променливата $\ln(V)$ от двете променливи $\ln([A])$ и $\ln([B]$ с коефициенти на регресия x и y и отрез $\ln(k)$.

Дипломантът избира начални концентрации, смесва веществата и след 10 секунди мери абсорбцията на веществото **P** на спектрометър. От закона на Буге-Ламберт-Беер определя молярната концентрация на **P** и като раздели на

10 намира началната скорост на реакцията. Данните, които е получил, са дадени в таблица I.

Таблица I. Начални концентрации на веществата и начална скорост на рекцията.

A mol/l	B mol/l	V mol/(ls)
0.1	0.1	0.00123
0.2	0.1	0.00247
0.3	0.2	0.01482
0.3	0.3	0.03334
0.4	0.5	0.12349

a) Отворете файла **sem05_multreg.xls**. Разгледайте таблицата (sheet) "**READY**", в която

- са дадени горните числови данни;

- началните концентрации и началната скорост са логаритмувани;

- регресионните коефициенти са изчислени с **Excel** функцията **LINEST()**, която познавате от дисциплината „Статистика и метрология в химията" - регионът **A12:C14**.

! Отворете Help-а и прочетете за функцията LINEST(). Обърнете внимание, че изходът от нея на първия ред започва с последния коефициент на регресия, b_m, в следващата клетка е по-предният коефициент, b_{m-1}, и т.н. докато редът завърши с отреза!

b) Отворете таблицата (sheet) "**WORK**" на файла **sem05_multreg.xls** и повторете изчисленията.

c) Стартирайте програмата **STATISTICA**. Копирайте данните и изчислете регресията. Не забравяйте да проверите, че checkbox-а **No Intercept** не е избран - т.е. да се прилага уравнение с отрез. Съвпадат ли изчислените стойности с тези от програмата **Excel**?

Ако не можете да се справите с изчисленията, на поддиректорията **stats** са дадени два файла - **sem08c01.sta** и **sem08c01.stw**, в които са данните и резултатите от регресията.