Семинар 5

Линейна еднопроменлива регресия

(преговор)

Поради теоретичния характер на лекцията, то в този семинар ще бъдат решавани задачи, касаещи теорията и приложението на линейната еднопроменлива регресия, която е изучавана донякъде в дисциплината „Статистика и метрология в химията".

Задача 5.1. Един от критериите за "максимално доближаване" на кривата до експерименталните точки е сумата S, която се дава с уравнение (с5.1)

$$S = \Sigma (У_к - У'_к)^2, \qquad (с5.1)$$

където $У'_к$ е изчислената стойност на У при заместване на $Х_к$ в уравнението на кривата, а $У_к$ е експерименталната стойност, съответстваща на $Х_к$. При линейна регресия $У'_к = b_1 Х_к + b_0$. Тогава сумата в (с5.1) добива вида:

$$S = \sum_{k=1}^{M} (Y_k - b_1 X_k - b_0)^2$$

Това е функция на две променливи – $S = S(b_1, b_0)$: b_1 и b_0 са един вид променливи, понеже сумата от квадратите зависи от техните стойности. Диференцирайте сумата по b_1 и b_0.

Задача 5.2. Условието за локален екстремум (максимум или минимум) е първите производни по променливите да са равни на нула. Приравнете двете производни, които получихте в задача 5.1. Получава се система от две линейни уравнения с постоянни коефициенти с две неизвестни - b_1 и b_0. Напишете системата от уравнения.

Задача 5.3. По формулите на Крамер решете системата. Получават се т.н. формули на метода на най-малките квадрати.

Задача 5.4*. Че получените формули дават минимум на сумата в (с5.1) се прескача в почти всички учебници – аз лично не съм го срещал никъде. Но

ние с Вас ще разгледаме проблема. От висшата математика е известно, че в точките където първите производни са нули имаме минимум, ако т.н. Хесиан, H, е положителен и сумата от вторите производни S_{xx} и S_{yy} е положителна.

$$H = \begin{vmatrix} \frac{\partial^2 S}{\partial x^2} & \frac{\partial^2 S}{\partial x \partial y} \\ \frac{\partial^2 S}{\partial y \partial x} & \frac{\partial^2 S}{\partial y^2} \end{vmatrix}$$

Това е детерминанта от матрица 2 на 2, състояща се от вторите производни на функцията на две променливи по тях. Смесените частни производни са равни една на друга при непрекъснати функции, каквато е функцията в уравнение (с5.1).

Намерете вторите производни на функцията ***S = S(b₁, b₀)***. Заместете ги в Якобиана и проверете дали той винаги е положителен.

Решение на задачите

Задача 5.1. Един от критериите за "максимално доближаване" на кривата до експерименталните точки е сумата S, която се дава с уравнение (с5.1)

$$S = \Sigma (У_к - У'_к)^2, \qquad (с5.1)$$

където $У'_к$ е изчислената стойност на У при заместване на $Х_к$ в уравнението на кривата, а $У_к$ е експерименталната стойност, съответстваща на $Х_к$. При линейна регресия $У'_к = b_1 Х_к + b_0$. Тогава сумата в (8.3) добива вида:

$$S = \sum_{k=1}^{M} (Y_k - b_1 X_k - b_0)^2 \qquad (с5.2)$$

Това е функция на две променливи – ***S = S(b₁, b₀)***: b_1 и b_0 са един вид променливи, понеже сумата от квадратите зависи от техните стойности. Диференцирайте сумата по b_1 и b_0.

Решение: Използваме две свойства на производната: 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_1 е равна на:

$$\frac{\partial S}{\partial b_1} = \frac{\partial \sum_{k=1}^{M} (Y_k - b_1 X_k - b_0)^2}{\partial b_1} = \sum_{k=1}^{M} \frac{\partial (Y_k - b_1 X_k - b_0)^2}{\partial b_1} =$$

$$= \sum_{k=1}^{M} 2(Y_k - b_1 X_k - b_0) \frac{\partial (Y_k - b_1 X_k - b_0)}{\partial b_1} = \sum_{k=1}^{M} 2(Y_k - b_1 X_k - b_0)(-X_k) =$$

$$= (-2) \sum_{k=1}^{M} (Y_k X_k - b_1 X_k X_k - b_0 X_k) = (-2)(\sum_{k=1}^{M} Y_k X_k - \sum_{k=1}^{M} b_1 X_k X_k - \sum_{k=1}^{M} b_0 X_k) =$$

$$= (-2)(\sum_{k=1}^{M} Y_k X_k - b_1 \sum_{k=1}^{M} X_k X_k - b_0 \sum_{k=1}^{M} X_k)$$

Първата производна по b_0 е равна на:

$$\frac{\partial S}{\partial b_0} = \frac{\partial \sum_{k=1}^{M} (Y_k - b_1 X_k - b_0)^2}{\partial b_0} = \sum_{k=1}^{M} \frac{\partial (Y_k - b_1 X_k - b_0)^2}{\partial b_0} =$$

$$= \sum_{k=1}^{M} 2(Y_k - b_1 X_k - b_0) \frac{\partial (Y_k - b_1 X_k - b_0)}{\partial b_0} = \sum_{k=1}^{M} 2(Y_k - b_1 X_k - b_0)(-1) =$$

$$= (-2) \sum_{k=1}^{M} (Y_k - b_1 X_k - b_0) = (-2)(\sum_{k=1}^{M} Y_k - \sum_{k=1}^{M} b_1 X_k - \sum_{k=1}^{M} b_0) =$$

$$= (-2)(\sum_{k=1}^{M} Y_k - b_1 \sum_{k=1}^{M} X_k - b_0 M)$$

Задача 5.2. Условието за локален екстремум (максимум или минимум) е първите производни по променливите да са равни на нула. Приравнете двете производни, които получихте в задача 8.1. Получава се система от две линейни уравнения с постоянни коефициенти с две неизвестни - b_1 и b_0. Напишете системата от уравнения.

Решение: Приравняваме двете производни, които получихме от задача 8.1, на нула:

$$\frac{\partial S}{\partial b_1} = (-2)(\sum_{k=1}^{M} Y_k X_k - b_1 \sum_{k=1}^{M} X_k X_k - b_0 \sum_{k=1}^{M} X_k) = 0$$

$$\frac{\partial S}{\partial b_0} = (-2)(\sum_{k=1}^{M} Y_k - b_1 \sum_{k=1}^{M} X_k - b_0 M) = 0$$

След съкращававане на –2 и прехвърляне от другата страна на членовете, в които няма b_1 и b_0, получаваме:

$$\begin{cases} b_1 \sum_{k=1}^{M} X_k X_k + b_0 \sum_{k=1}^{M} X_k = \sum_{k=1}^{M} Y_k X_k \\ b_1 \sum_{k=1}^{M} X_k + b_0 M = \sum_{k=1}^{M} Y_k \end{cases}$$

Задача 5.3. По формулите на Крамер решете системата. Получават се т.н. формули на метода на най-малките квадрати.

Решение: Формулите на Крамер в този случай дават:

$$b_1 = \frac{\begin{vmatrix} \sum_{k=1}^{M} Y_k X_k & \sum_{k=1}^{M} X_k \\ \sum_{k=1}^{M} Y_k & M \end{vmatrix}}{\begin{vmatrix} \sum_{k=1}^{M} X_k X_k & \sum_{k=1}^{M} X_k \\ \sum_{k=1}^{M} X_k & M \end{vmatrix}} \quad u \quad b_0 = \frac{\begin{vmatrix} \sum_{k=1}^{M} X_k X_k & \sum_{k=1}^{M} Y_k X_k \\ \sum_{k=1}^{M} X_k & \sum_{k=1}^{M} Y_k \end{vmatrix}}{\begin{vmatrix} \sum_{k=1}^{M} X_k X_k & \sum_{k=1}^{M} X_k \\ \sum_{k=1}^{M} X_k & M \end{vmatrix}}$$

Ако разкрием детерминантите се получава:

$$b_1 = \frac{M \sum_{k=1}^{M} X_k Y_k - \sum_{k=1}^{M} X_k \sum_{k=1}^{M} Y_k}{M \sum_{k=1}^{M} X_k{}^2 - (\sum_{k=1}^{M} X_k)^2} \quad u \quad b_0 = \frac{\sum_{k=1}^{M} X_k{}^2 \sum_{k=1}^{M} Y_k - \sum_{k=1}^{M} X_k \sum_{k=1}^{M} X_k Y_k}{M \sum_{k=1}^{M} X_k{}^2 - (\sum_{k=1}^{M} X_k)^2} \qquad (c5.3)$$

Задача 5.4*. Че получените формули дават минимум на сумата в (с5.2) се прескача в почти всички учебници – аз лично не съм го срещал никъде. Но ние с Вас ще разгледаме проблема. От висшата математика за функция на

две променливи, x и y, S = S(x,y) е известно[1], че в точките, където първите производни са нули имаме минимум, ако т.н. Хесиан, ***H***, е положителен и сумата от вторите производни S_{xx} и S_{yy} е положителна.

$$H = \begin{vmatrix} \frac{\partial^2 S}{\partial x^2} & \frac{\partial^2 S}{\partial x \partial y} \\ \frac{\partial^2 S}{\partial y \partial x} & \frac{\partial^2 S}{\partial y^2} \end{vmatrix}$$

Това е детерминанта на матрица 2 на 2, състояща се от вторите производни на функцията на две променливи по тях. Смесените частни производни са равни една на друга при непрекъснати функции, каквато е функцията в уравнение (c5.2).

Решение: Вторите производни са равни на производна от първите производни:

$$\frac{\partial^2 S}{\partial {b_0}^2} = \frac{\partial \frac{\partial S}{\partial b_0}}{\partial b_0} = \frac{\partial(-2)(\sum_{k=1}^{M} Y_k - b_1 \sum_{k=1}^{M} X_k - b_0 M)}{b_0} = \partial(-2)(-M) = 2M$$

$$\frac{\partial^2 S}{\partial {b_1}^2} = \frac{\partial \frac{\partial S}{\partial b_1}}{\partial b_1} = \frac{\partial[(-2)(\sum_{k=1}^{M} Y_k X_k - b_1 \sum_{k=1}^{M} X_k X_k - b_0 \sum_{k=1}^{M} X_k)]}{b_1} = (-2)(-\sum_{k=1}^{M} X_k X_k) = 2\sum_{k=1}^{M} {X_k}^2$$

$$\frac{\partial^2 S}{\partial b_1 \partial b_0} = \frac{\partial \frac{\partial S}{\partial b_1}}{\partial b_0} = \frac{\partial[(-2)(\sum_{k=1}^{M} Y_k X_k - b_1 \sum_{k=1}^{M} X_k X_k - b_0 \sum_{k=1}^{M} X_k)]}{b_0} = (-2)(-\sum_{k=1}^{M} X_k) = 2\sum_{k=1}^{M} X_k$$

За втората смесена производна бихме получили същия резултат, ако диференцираме първо по b_0, после по b_1:

$$\frac{\partial^2 S}{\partial b_0 b_1} = \frac{\partial \frac{\partial S}{\partial b_0}}{\partial b_1} = \frac{\partial(-2)(\sum_{k=1}^{M} Y_k - b_1 \sum_{k=1}^{M} X_k - b_0 M)}{b_1} = \partial(-2)(-\sum_{k=1}^{M} X_k) = 2\sum_{k=1}^{M} X_k$$

[1] Вижте тази страница: http://en.wikipedia.org/wiki/Second_partial_derivative_test

Вижда се, че вторите производни са положителни: $S_{b_0 b_0} = 2M$ и $S_{b_1 b_1}$ е сума от квадрати.

Хесианът на тази функция от двете променливи е:

$$H = \begin{vmatrix} \frac{\partial^2 S}{\partial b_1^{\,2}} & \frac{\partial^2 S}{\partial b_1 \partial b_0} \\ \frac{\partial^2 S}{\partial b_0 \partial b_1} & \frac{\partial^2 S}{\partial b_0^{\,2}} \end{vmatrix}$$

Като заместим изразите за вторите производни за Хесиана получаваме:

$$H = \begin{vmatrix} 2\sum_{k=1}^{M} X_k^{\,2} & 2\sum_{k=1}^{M} X_k \\ 2\sum_{k=1}^{M} X_k & 2M \end{vmatrix} = 4[M\sum_{k=1}^{M} X_k^{\,2} - \left(\sum_{k=1}^{M} X_k\right)^2]$$

Сумата в средните скоби може да се представи по следния начин:

$$M\sum_{k=1}^{M} X_k^{\,2} - \left(\sum_{k=1}^{M} X_k\right)^2 = M\sum_{k=1}^{M} X_k^{\,2} - M^2\left(\frac{\sum_{k=1}^{M} X_k}{M}\right)^2 = M\sum_{k=1}^{M} X_k^{\,2} - M^2\left(\overline{X}\right)^2 = M[\sum_{k=1}^{M} X_k^{\,2} - M\left(\overline{X}\right)^2]$$

В задача 2.2 на семинар 2 се доказва, че в последното равенство сумата в средните скоби е стандартното отклонение, умножено по M-1. Тогава за горния израз се получава:

$$M[\sum_{k=1}^{M} X_k^{\,2} - M\left(\overline{X}\right)^2] = M[\sum_{k=1}^{M} (X_k - \overline{X})^2$$

Което е положителна величина, защото е сума от квадрати на реални величини. Т.е. Хесианът е положителен и имаме минимум!

$$J = \begin{vmatrix} 2\sum_{k=1}^{M} X_k^{\,2} & 2\sum_{k=1}^{M} X_k \\ 2\sum_{k=1}^{M} X_k & 2M \end{vmatrix} = 4[M\sum_{k=1}^{M} X_k^{\,2} - \left(\sum_{k=1}^{M} X_k\right)^2] = 4M\sum_{k=1}^{M} (X_k - \overline{X})^2 > 0$$

Практически задачи

Задача C1. Студент от ХФ на ПУ решава да намери топлината на изпарение, $\Delta H_{изп}$, чрез уравнението на Клаузиус-Клапейрон[2]. При измерване на налягането на парите на азотната киселина **p** в зависимост от температурата **t** при постоянен обем е получил следните резултати:

t /°C	0	20	40	50	70	80	90	100
p / mm Hg ст.	14.4	47.9	133	208	467	670	937	1282

a) Отворете файла **seminar05_regression.xls**. Разгледайте таблицата (sheet) "**READY**", в която

- са дадени горните числови данни;

- температурата е превърната в единици Келвин, намерена е нейната обратна стойност, **1/T**, и налягането е логаритмувано, **ln(p)** – целта е да се получи зависимостта $\ln(p) = -(\Delta H_{изп}/R)(1/T) + const.$, която е линейна в координати **ln(p)** и **1/T**. Очевидно нейният наклон е равен на $b_1 = -\Delta H_{изп}/R$, т.е. $\Delta H_{изп} = -b_1 R$.

- намерени са четирите вида суми в (c5.3) с функциите на Excel **SUM()**, **SUMSQ()** и **SUMPRODUCT()**;

- намерени са коефициентите на регресия, b_1 и b_0" – регионът **G7:G8**;

- намерена е молярната топлина на изпарения в клетки **B21** и **B22**: $\Delta H_{изп} = 37.99\ KJ\ K^{-1}\ mol^{-1}$: интересно, че в книгата на Еткинс за отговор е дадено $\Delta H_{изп} = 34.9\ KJ\ K^{-1}\ mol^{-1}$;

- регресионните коефициенти са изчислени и с **Excel** функцията **LINEST()**, която познавате от дисциплината „Статистика и метрология в химията" – регионът **A14:B16.**

b) Отворете таблицата (sheet) "**WORK**" на файла **seminar05_regression.xls** и повторете изчисленията.

---

[2] $p = p^*\exp[-\Delta H_{изп}(1/RT - 1/RT^*)]$; П.Еткинс; Физическая химия. Издат. „Мир", Москва, 1980 г. Том 1, стр. 203. Данните в задачата са взети от същата книга, задача 7.18, стр. 223.