Линейна регресия

Теоретична постановка

Линейна регресия

Линейната регресия е статистически метод за построяване на (евентуално) приемлива линейна връзка между група независими променливи x_1, x_2, ..., x_m и зависима променлива y (наричана *индикатор* или *контролна величина*). Т.е. построяване на линеен математически модел, с чиято помощ могат да се правят прогнози за състоянието на y при различни данни за x.

За тази цел се провеждат n на брой (случайни) експеримента – за различни стойности на x-овете се измерват стойностите на зависимата променлива y, т.е. за i-тия експеримент имаме съпоставянето:

$$y_i \leftrightarrow L\,(x_{i\,1},\ x_{i\,2},\ \ldots,\ x_{i\,m}) \tag{1}$$

Тогава можем да запишем резултатите от експериментите във вида:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1\,m} \\ x_{21} & x_{22} & \cdots & x_{2\,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n\,1} & x_{n\,2} & \cdots & x_{n\,m} \end{pmatrix} = (\,x_1\ \ x_2\ \ \cdots\ \ x_m\,) = \left(x_{i\,j}\right)_{n\times m} \tag{2}$$

Нашата задача е да съставим линейна зависимост между y и x_1, ..., x_m, която приемливо да удовлетворява данните от (2). Освен това, тази зависимост трябва да отчита евентуална неточност в данните при протичането на експериментите. Тогава търсим такива стойности a_k (които се наричат параметри на модела или регресионни коефициенти), че да е в сила зависимост от вида:

$$y = L\,(x_1,\ \ldots,\ x_m) + \epsilon = a_1\,x_1 + a_2\,x_2 + \ldots + a_m\,x_m + \epsilon,\ \epsilon = (\,\epsilon_1\ \ \epsilon_2\ \ \cdots\ \ \epsilon_m\,)^T \tag{3}$$

Тук ϵ_i съответсва на отклонението от истинската стойност от наблюдаваната в даден експеримент. Тогава имаме системата:

$$y_i = a_1\,x_{i\,1} + a_2\,x_{i\,2} + \ldots + a_m\,x_{i\,m} + \epsilon_i,\ i = 1,\ \ldots,\ n \tag{4}$$

Или записано матрично:

$$y = L\,(x_1,\ \ldots,\ x_m) + \epsilon = X\,a + \epsilon,\ a = (\,a_1\ \ a_2\ \ \cdots\ \ a_m\,)^T \tag{5}$$

Задачата е да намерим стойности на a, такива че отклонението ϵ да е минимално, т.е. по някакъв начин (тъй като подлежи на интерпретация) да минимизираме ϵ, т.е. да е изпълнено

$$\epsilon = \min\,(y - L\,(x_1,\ ..,\ x_m)) = \min\,(y - X\,a) \tag{6}$$

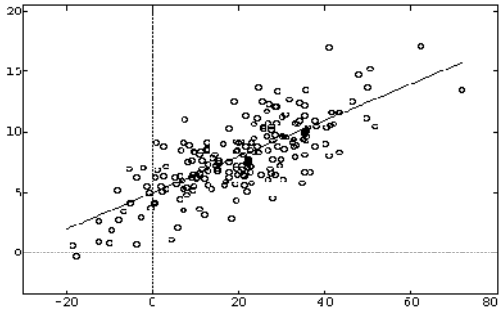
Желанието ни е при намерени регресионни коефициенти a да е приемливо приближението:

$$y \approx L\,(x_1,\ \ldots,\ x_m) = \hat{y} \tag{7}$$

Конкретното намиране на регресионните коефициенти ще покажем по-долу. Въпросът кога приближението (6) можем да считаме за приемливо също е разгледан по-долу.

Метод на най-малките квадрати (МНК)

МНК е метод широко употребяван за т.нар. *фитване* (т.е. приемливо апроксимиране) на линеен модел по дискретни данни. Например долната графика показва, как даденото множество от данни е *фитнато* с права.



Нека разгледаме нашия случай. За него търсим приближеното решение $\hat{y}$:

$$\hat{y} = X\,a,\; a = (a_1\;\; a_2\;\; \cdots\;\; a_m)^T \tag{8}$$

Нека означим сумата от квадратите от разстоянията $\|\hat{y}_i - y_i\|$ с $Q(x)$

$$Q(x) = \sum_{i=1}^{n} \left\| y_i - \hat{y}_i \right\|^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} a_i\, x_{ij} \right)^2 \tag{9}$$

Методът се състои в търсенето на a при условие, че сумата $Q(a)$ е минимална. Така всъщност ние минизираме квадрата на грешката (това, което искахме да направим при линейната регресия), т.е.:

$$Q(x) = (y - X\,a)^2 = \epsilon^2 \tag{10}$$

Приемаме, че съществува (единствен) минимум и на базата на това наше твърдение ще търсим стойностите на a. Съществуване и единствност на минимума може да се докаже, но тук това няма да правим.

Вече, за да намерим a, достатъчно е да решим системата, получена от факта, че НДУ за да имаме минимум е първите частни производни да са нули:

$$\frac{\partial \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} a_i\, x_{ij} \right)^2 \right)}{\partial a_k},\; k = 1,\; \ldots,\; m \tag{11}$$

- **Матричен запис на решението**

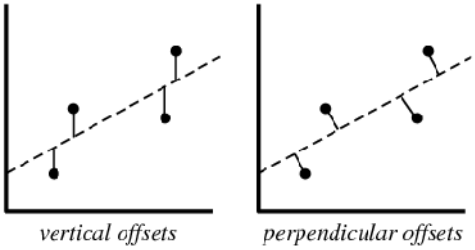
Може да се покаже, използвайки матричния запис на $Q(x)$, че за намерените в (11) регресионни коефициенти a е в сила матричното представяне:

$$a = \left(X^T X\right)^{-1} \left(X^T y\right) \tag{12}$$

Приближението $\hat{y}$ намираме от (8).

- **Забележка върху дефинирането на разстоянията**

Така дефинираните разстояния, са всъщност вертикалните такива между приближението и множеството дадени точки. Друга възможност е да разглеждаме проекциите на тези точки върху приближението. Първият подход обикновено (както и при нас) се използва, тъй като притежава известни математически улеснения, а и при достатъчен брой случайно подбрани начални данни, разликата между двата подхода е доста малка.



■ **Забележка върху линейността**

Първо, нека отбележим, че така търсеният модел (7), всъщност представлява линеен хомогенен такъв. Често обаче, поради наличие на константи, се използва нехомогенна линейна регресия. Тогава (3) добива вида

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + a_{m+1} + \epsilon \tag{13}$$

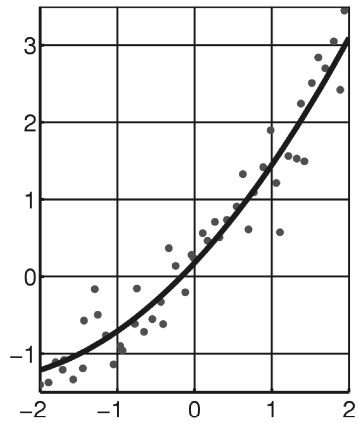
За да не променяме вида на връзка (3) и последвалите я разсъждения е необходимо да променим вида на X – добавяме единици в последната колона.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} & 1 \\ x_{21} & x_{22} & \cdots & x_{2m} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} & 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_m & 1 \end{pmatrix} \tag{14}$$

Тогава (13) може да се запише вече в познатия вид

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + a_{m+1} x_{m+1} + \epsilon \tag{15}$$

Горните разсъждения ни въвеждат на мисълта, че с промени в матрицата X можем да развием досега разглежданата регресия. Естесвено е *фитването* на данни не винаги да се извършва с права, а например с полином и дори с други математически обекти. Например долната графика показва *фитване* чрез парабола



Тук пак говорим за линейна регресия, тъй като искаме линейност не по независимите променливи x_k, а по параметрите a_k. За да *фитнем* с парабола проста (т.е. едномерна) регресия например, единствено нужно е да запишем (3) във вида:

$$y = a_1 + a_2 x_1 + a_3 x_1^2 + \epsilon \tag{16}$$

И пак с цел да запазим стария вид на (3), това ще значи да запишем X така:

$$X = \begin{pmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{21} & x_{21}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1}^2 \end{pmatrix} \tag{17}$$

Където стълбовете отново се очаква да са линейно-независими.

■ **Оценка за линейност**

Както видяхме, поради естеството на приложения подход намерената линейна зависимост е приближение на истинската. Т.е. дори и първоначалната зависимост да няма линеен характер, приближение пак ще бъде намерено. Тогава от изключително значение е да имаме способи за оценка линейността на първоначалната зависимост и *качеството* на направеното приближение. За тази цел ще направим няколко практически оценки върху т.нар. коефициенти на корелация.

▫ *Предварителен тест*

Преди да ползваме МНК нека проверим дали връзката между y и x-овете е приемливо линейна, т.е. употребата на линейна регресия е оправдана.

1. Система корелационни коефициенти

1.1. *базисни* корелационни коефициенти, корелационно средно

$$r_k^* = \frac{\langle y, x_k \rangle}{\|y\| \, \|x_k\|}, \text{ т.е. } r_k^* = \frac{\sum_{i=1}^n y_i \, x_{i\,k}}{\sqrt{\sum_{i=1}^n y_i^2} \, \sqrt{\sum_{i=1}^n x_{i\,k}^2}}, \; k = 1, \; \ldots, m \tag{18}$$

Ясно е, че тези величини имат смисъл на косинуси на ъглите между вектора y и векторите x_k. От този геометричен смисъл е изведен следният *практически принцип*: считаме, че има (приемлива) линейност на връзката $y \leftrightarrow \{x_k, k = 1, \ldots, m\}$, ако $|r_k^*| \geq 0.75, \; \forall k$. В такъв случай казваме също, че величините $\{x_1, x_2, \ldots, x_m\}$ и y корелират (приемливо добре) или, че има (приемливо добра) *корелация* между тях. Като пресметнем аритметичното средно от абсолютните стойностите на параметрите r_k^* получаваме коефициент, който можем да наричаме (базисно) корелационно средно или среден базисен корелационен коефициент:

$$r^* = \frac{\sum_{k=1}^m |r_k^*|}{m} \tag{19}$$

И в случая на предварителни тестове чрез r^* се изисква условието $|r_k^*| \geq 0.75, \; \forall k$ (за да се преме, че има добра корелация).

1.2. *относителни* корелационни коефициенти, среден коефициент

Да въведем средните стойности $\overline{x}_k = \frac{\sum_{i=1}^n x_{ik}}{n}$, $k = 1, \ldots, m$, $\overline{y} = \frac{\sum_{i=1}^n y_i}{n}$. В статистическия анализ системно се работи с т.нар. отклонения (от съответното средно) $y_i - \overline{y}$, $x_{ik} - \overline{x}$, $i = 1, \ldots, n$, $k = 1, \ldots, m$ и с изчислените посредством тях корелационни коефициенти (които можем да наричаме *относителни*):

$$r_k = \frac{\langle y - \overline{y}\, e_n, \; x_k - \overline{x}_k \, e_n \rangle}{\|y - \overline{y}\, e_n\| \, \|x_k - \overline{x}_k \, e_n\|} \tag{20}$$

Горе e_n е единичният вектор от размерност n. Среден корелационен коефициент получаваме по формулата:

$$r = \frac{\sum_{k=1}^m |r_k|}{m} \tag{21}$$

Отново се има предвид практическото условие, съответно $|r_k| \geq 0.75, \; \forall k$ или $r \geq 0.75$.

▫ *Пост-оценка на приближението*

След като по горните тестове сме се убедили, че имаме приемлива линейност и сме намерили регресионните коефициенти a и приближението $\hat{y}$, е необходимо отново постфактум да го оценим. Това ще направим, като разгледаме стойността на общия корелационен коефициент.

2. Общ корелационен коефициент

Това название всъщност е естествена последица от очевидния геометричен смисъл на величината r_0. Известно е, че векторът $\hat{y}$ е ортогоналната проекция на y върху хиперравнината, породена от векторите x_k ($k = 1, \ldots, m$) и тогава представлява косинуса на ъгъла на наклона на вектора y спрямо тази хиперравнина. Ясно е, че когато споменатия ъгъл е „достатъчно" малък, можем да считаме, че векторът y е почти успореден на (лежи в) хиперравнината и значи е фактически линейна комбинация на векторите x_k ($k = 1, \ldots, m$).

При вече установени приемливи стойности на корелационните коефициенти, пълната практическа процедура по МНК включва (като финална) и следната стъпка.

Пресмятаме отклоненията $y_i - \hat{y}_i,\ i = 1,\ ..., n$
Като окончателна практическа оценка на съответния линеен регресионен модел проверяваме дали за поне 75 % от отклоненията $y_i - \hat{y}_i$ имаме, че $|y_i - \hat{y}_i|$ не надминава 10 % от средната стойност $\frac{y_i+\hat{y}_i}{2}$. В такъв случай за останалите (най-много 25 %) отклонения приемаме, че големината им се дължи на неточности в отчитането (измерването) на съответните експериментални стойности y_i.

■ Качества на линейните модели

Трябва да отбележим, че широката употреба на линейни модели се дължи не само на чисто линейният характер на изследваните зависимости, но и на факта, че всеки процес в достатъчно малки граници на независимите променливи може да бъде разглеждан като линеен. От друга страна, ако искаме да използваме модела екстраполирано, то трябва да сме убедени в линейността на началната зависимост.

■ Пример: пречиствателна станция

■ Постановка

Разглеждаме действието на пречиствателна станция. В даден област (където се намира станцията) на даден речен басейн се наблюдава определено количество вещество замърсител, обозначено с B_1. Използвайки константно количество от вещество пречиствател B_2 и наблюдавайки експеримента в променлива среда, характеризирана с температура T и наличие на катализатор K, ние получаваме продукт на пречиствателната реакция – вещество B_3. По наличието на вещество B_3 можем да съдим за степента на пречистване на района. Т.е. имаме зависимостта:

$$B_3 \leftrightarrow (B_1, T, K) \qquad (22)$$

Физическото отчитане на вещество B_3 обаче е трудоемко и ние искаме да построим модел, реализиращ зависимост (22), така че отчитането на B_3 да става на базата на B_1, T и K.

■ Експериментални данни

За да намерим явния вид на зависимост (22) са проведени 15 на брой случайни експеримента. Резултатите са въведени в съответно вектора myY и матрицата myX.

```
myY = {140.28, 142.02, 149.90, 147.12, 163.62, 173.40, 178.86,
    186.26, 183.53, 198.76, 205.30, 206.77, 198.42, 216.48, 221.45};

      252.36  96.67  8.37  1
      262.54 100.07  9.07  1
      285.70  96.78  9.35  1
      277.52 101.30  9.67  1
      307.95 100.35  9.45  1
      322.44 104.80 10.12  1
      334.88 106.17 10.35  1
myX = 350.11 109.20 11.03  1;
      346.10 104.48 10.38  1
      374.91 106.88 12.15  1
      378.49 113.14 12.98  1
      397.48 112.38 11.34  1
      378.39 109.07 10.95  1
      393.44 114.45 12.89  1
      403.84 115.23 13.71  1

{n, m} = Dimensions@myX;
```

В myY сме въвели резултатите (количеството (в килограми) B_3) от всеки един от петнадесетте експеримента. В myX сме въвели условията, при които са протекли експериментите:

- Първата колона (x_1) са количествата (в килограми) B_1

x_2 показва температурата на реакциите (в C°)

- x_3 са количествата на катализатора K (в грамове)
- Четвъртата колона показва, че ще разглеждаме нехомогенен модел. Тя отчита присъствието на константата B_2 и други възможни константи.

■ Предварителен тест

Първоначално нека анализираме данните за линейност по тестовете, разгледани по-горе. Ако получим задоволителни резултати, ще има смисъл да пробваме да изградим модела линейно, използвайки линейна регресия.

▫ *Базисни корелационни коефициенти*

Първо да сметнем базисните корелационни коефициенти:

```
Table[
 myRBasic[k] = myY.myX[[All, k]] / (Norm[myY] Norm[myX[[All, k]]]),
 {k, 1, m}
]
```

```
{0.999827, 0.995695, 0.998723, 0.989504}
```

Както виждаме, тестът дава добри резултати – изключително близки до единица. Максималният ъгълът (в градуси) е:

```
ArcCos@Min@% Degree^-1
```

```
8.3088
```

Т.е. изглежда, че имаме добра линейност.

Среден базисен коефициент:

```
myRBasicAvg = (∑_{k=1}^{m} Abs[myRBasic[k]]) / m
```

```
0.995937
```

И тук получаваме добър резултат:

```
ArcCos@% Degree^-1
```

```
5.16653
```

▫ *Относителни корелационни коефициенти*

Първо да сметнем средните стойности:

```
myYAvg = (∑_{k=1}^{n} myY[[k]]) / n
```

```
180.811
```

Или изполвайки Apply (@@) върху списъка

```
Table[
 myXAvg[i] = Plus @@ myX[[All, i]] / n,
 {i, 1, m}
]
```

```
{337.743, 106.065, 10.7873, 1}
```

Тогава за относителните корелационни коефициенти имаме:

```
Table[
 myRRel[k] = (myY - myYAvg) . (myX[[All, k]] - myXAvg[k]) / (Norm[myY - myYAvg] Norm[myX[[All, k]] - myXAvg[k]]),
 {k, 1, m}
]
```

```
Power::infy : Infinite expression 1/0 encountered. ≫

∞::indet : Indeterminate expression 0. ComplexInfinity encountered. ≫

{0.991695, 0.955469, 0.93732, Indeterminate}
```

Забележката от страна на *Mathematica* е по отношение на последния относителен корелационен коефициент, където средната стойност на x_4 и тези на x_4 съвпадат. Това не бива да ни тревожи.

```
ArcCos@Min@%[[1 ;; 3]] Degree^-1
```

```
20.3937
```

Тук ъгълът вече е малко по-голям но също така приемлив. За средният корелационен коефициент имаме:

```
myRRelAvg = (∑_{k=1}^{m-1} Abs[myRRel[k]]) / (m - 1)
```

```
0.961495
```

```
ArcCos@% Degree^-1
```

```
15.9514
```

Отново добър резултат.

След проведените тестове стигаме до извода, че построяване на линеен модел за задачата има смисъл.

■ Намиране на регресионните коефициенти

▫ *Пресмятане чрез МНК*

In[55]:=
$$myQ = \sum_{i=1}^{n}\left(myY[[i]] - \sum_{j=1}^{m} a[j]\, myX[[i, j]]\right)^2$$

Out[55]=
$$\begin{aligned}&(221.45 - 403.84\, a[1] - 115.23\, a[2] - 13.71\, a[3] - a[4])^2 + \\ &(205.3 - 378.49\, a[1] - 113.14\, a[2] - 12.98\, a[3] - a[4])^2 + \\ &(216.48 - 393.44\, a[1] - 114.45\, a[2] - 12.89\, a[3] - a[4])^2 + \\ &(198.76 - 374.91\, a[1] - 106.88\, a[2] - 12.15\, a[3] - a[4])^2 + \\ &(206.77 - 397.48\, a[1] - 112.38\, a[2] - 11.34\, a[3] - a[4])^2 + \\ &(186.26 - 350.11\, a[1] - 109.2\, a[2] - 11.03\, a[3] - a[4])^2 + \\ &(198.42 - 378.39\, a[1] - 109.07\, a[2] - 10.95\, a[3] - a[4])^2 + \\ &(183.53 - 346.1\, a[1] - 104.48\, a[2] - 10.38\, a[3] - a[4])^2 + \\ &(178.86 - 334.88\, a[1] - 106.17\, a[2] - 10.35\, a[3] - a[4])^2 + \\ &(173.4 - 322.44\, a[1] - 104.8\, a[2] - 10.12\, a[3] - a[4])^2 + \\ &(147.12 - 277.52\, a[1] - 101.3\, a[2] - 9.67\, a[3] - a[4])^2 + \\ &(163.62 - 307.95\, a[1] - 100.35\, a[2] - 9.45\, a[3] - a[4])^2 + \\ &(149.9 - 285.7\, a[1] - 96.78\, a[2] - 9.35\, a[3] - a[4])^2 + \\ &(142.02 - 262.54\, a[1] - 100.07\, a[2] - 9.07\, a[3] - a[4])^2 + \\ &(140.28 - 252.36\, a[1] - 96.67\, a[2] - 8.37\, a[3] - a[4])^2\end{aligned}$$

In[56]:=
```
Solve[
 Table[
  ∂a[i] myQ == 0
  , {i, 1, m}
 ]
]
```

Out[56]= {{a[1] → 0.397982, a[2] → 0.45342, a[3] → 2.94118, a[4] → -33.4237}}

In[57]:=
```
myAMNK = Array[a, m] /. %
```

Out[57]= {{0.397982, 0.45342, 2.94118, -33.4237}}

▫ *Матрично пресмятане*

In[67]:=
```
myAMatrix = Inverse[Transpose@myX.myX].Transpose@myX.myY
```

Out[67]= {0.397982, 0.45342, 2.94118, -33.4237}

▫ *Пресмятане с Mathematica*

In[82]:=
```
myModel = LinearModelFit[{myX, myY}]
```

Out[82]= FittedModel[<<1>>]

In[83]:=
```
myModel["BestFit"]
```

Out[83]= 0.397982 #1 + 0.45342 #2 + 2.94118 #3 - 33.4237 #4

- **Сравнителна графика**

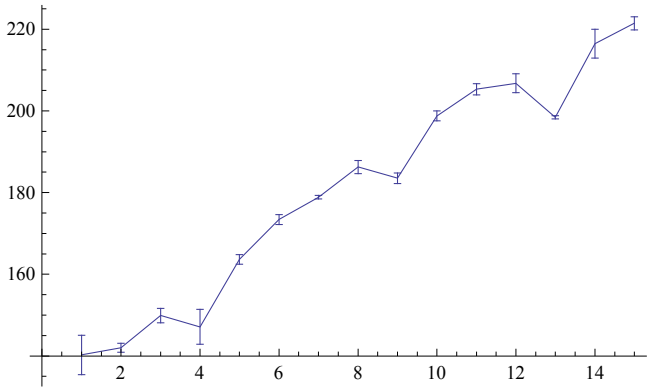
In[89]:=
```
myYApprox = myX.myAMatrix
```

Out[89]=
```
{135.461, 143.113, 151.662, 151.397, 162.43, 172.185, 178.433,
 187.868, 182.221, 199.981, 206.685, 209.074, 198.829, 212.964, 219.868}
```

In[90]:=
```
<< ErrorBarPlots`
```

In[93]:=
```
ErrorListPlot[
 Table[
  {myY[[i]], myY[[i]] - myYApprox[[i]]},
  {i, 1, n}
 ],
 Joined → True
]
```

Out[93]=



- **Пост-тестове**

  - ***Общ корелационен коефициент***

In[96]:=
$$\text{myR} = \frac{\text{myY.myYApprox}}{\text{Norm@myY Norm@myYApprox}}$$

Out[96]=
```
0.999923
```

  - ***Отклонение на грешката от средната стойност***

In[106]:=
$$\text{myW} = \text{Abs[myY - myYApprox]} - 0.1\ \frac{\text{Abs[myY + myYApprox]}}{2}$$

Out[106]=
```
{-8.96782, -13.1639, -13.3164, -10.649, -15.1121, -16.0639, -17.4379, -17.098,
 -16.9781, -18.7165, -19.2144, -18.4878, -19.4534, -17.9562, -20.4844}
```