Лекция 8. Метод наименьших квадратов

Постановка задачи

Опытным путем был получен ряд точек

$$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N).$$

Точки примерно выстраиваются в прямую линию. Как составить уравнение этой прямой? -- Уравнение прямой всегда имеет вид

$$y = kx + b.$$

Если бы точки точно лежали на этой прямой, то

$$kx_n + b - y_n = 0$$

и из двух таких уравнений можно было бы определить параметры k, b, полностью характеризующую прямую. Мы не можем сделать все величины

$$kx_n + b - y_n$$

равными нулю, но можем подобрать их таким образом, чтобы квадратичное уклонение

$$\sum_{n=1}^{N} (kx_n + b - y_n)^2$$

было минимально возможным. При этом говорят, что подбирают параметры уравнения прямой по методу наименьших квадратов (МНК).

Задача. Заданы N точек

$$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$$

Требуется подобрать числа k и b таким образом, чтобы величина

$$\sum_{n=1}^{N} (kx_n + b - y_n)^2$$

имела наименьшее значение.

Отыскание коэффициентов линейной зависимости по методу наименьших квадратов
Решению поставленной задачи предпошлем лемму.

Лемма (нер-во Коши-Буняковского). Всегда верно, что

$$\left(\sum_{n=1}^{N} x_n y_n\right)^2 \leq \left(\sum_{n=1}^{N} x_n^2\right) \cdot \left(\sum_{n=1}^{N} y_n^2\right),$$

причем неравенство строгое, если среди x_n и y_n имеются отличные от нуля.

Док-во. Рассмотрим вспомогательную функцию

$$\sum_{n=1}^{N} (x_n + t y_n)^2 = \sum_{n=1}^{N} x_n^2 + 2\left(\sum_{n=1}^{N} x_n y_n\right) t + \left(\sum_{n=1}^{N} y_n^2\right) t^2 > 0.$$

Следовательно дискриминант квадратного уравнения

$$\sum_{n=1}^{N} x_n^2 + 2\left(\sum_{n=1}^{N} x_n y_n\right) t + \left(\sum_{n=1}^{N} y_n^2\right) t^2 = 0$$

строго меньше нуля. Это означает, что

$$\frac{D}{4} = \left(\sum_{n=1}^{N} x_n y_n\right)^2 - \left(\sum_{n=1}^{N} x_n^2\right) \cdot \left(\sum_{n=1}^{N} y_n^2\right) < 0,$$

что и тр.д.

Решение задачи. Рассмотрим переменную

$$z = \sum_{n=1}^{N} (k x_n + b - y_n)^2$$

как функцию переменных k и b. Раскрыв скобки

$$(k x_n + b - y_n)^2 = k^2 \, {x_n}^2 + 2 \, b \, k \, x_n - 2 \, k \, x_n \, y_n + {y_n}^2 - 2 \, b \, y_n + b^2,$$

видим, что

$$z = \left(\sum_{n=1}^{N} x_n^2\right) k^2 + 2\left(\sum_{n=1}^{N} x_n\right) kb + N \, b^2 - 2\left(\sum_{n=1}^{N} x_n y_n\right) k - 2\left(\sum_{n=1}^{N} y_n\right) b + \left(\sum_{n=1}^{N} y_n^2\right),$$

то есть z является квадратичной функцией переменных k,b. Эта функция имеет только в стационарных точках, координаты (k,b) которых удовлетворяют системе

$$\begin{cases} \left(\sum_{n=1}^{N} x_n^2\right) k + \left(\sum_{n=1}^{N} x_n\right) b - \left(\sum_{n=1}^{N} x_n y_n\right) = 0 \\ \left(\sum_{n=1}^{N} x_n\right) k + Nb - \left(\sum_{n=1}^{N} y_n\right) = 0 \end{cases}$$

или

$$\begin{cases} \left(\sum_{n=1}^{N} x_n^2\right)k + \left(\sum_{n=1}^{N} x_n\right)b = \sum_{n=1}^{N} x_n y_n \\ \left(\sum_{n=1}^{N} x_n\right)k + Nb = \sum_{n=1}^{N} y_n \end{cases}$$

Определитель этой системы равен

$$\det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix} = N\sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n\right)^2.$$

В силу неравенства Коши-Буняковского

$$\left(\sum_{n=1}^{N} 1 \cdot x_n\right)^2 < \left(\sum_{n=1}^{N} x_n^2\right)(1 + \cdots + 1) = N\sum_{n=1}^{N} x_n^2$$

поэтому

$$\det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix} > 0.$$

Это означает, что координаты k и b стационарной точки можно найти по формам Крамара:

$$k = \det\begin{pmatrix} \sum_{n=1}^{N} x_n y_n & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} y_n & N \end{pmatrix} : \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix},$$

$$b = \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} y_n \end{pmatrix} : \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix}$$

Поскольку определитель матрицы

$$\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix}$$

квадратичной формы

$$\left(\sum_{n=1}^{N} x_n^2\right) k^2 + 2\left(\sum_{n=1}^{N} x_n\right) kb + N\, b^2$$

строго больше нуля, равно как и ее след

$$\mathrm{Tr}\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix} = \sum_{n=1}^{N} x_n^2 + N > 0,$$

по критерию Сильвестра квадратичная форма строго положительно определена, а наша квадратичная функция имеет в стационарной точке строгий глобальный минимум.

Ответ: минимум достигается при

$$k = \det\begin{pmatrix} \sum_{n=1}^{N} x_n y_n & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} y_n & N \end{pmatrix} : \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix},$$

$$b = \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} y_n \end{pmatrix} : \det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix}.$$

Числовой пример

Задача. Данные о росте безработицы x, %, и росте преступности y, %, приведены в таблице.

Год	1991	1992	1993	1994	1995	1996	1997
Уровень безработицы	0.5	1.2	2	3.1	4	5.2	5.9
Уровень преступности	4.2	4.3	4.4	4.6	4.6	4.7	4.9

Табл. 1

Допустим, что уровень преступности растет линейно с ростом безработицы, то есть что примерно $y = kx + b$. Требуется подобрать параметры k и b по МНК.

Решение.

Величина	Вычисление	Значение
$\sum_{n=1}^{N} x_n$	$0.5 + 1.2 + 2 + 3.1 + 4 + 5.2 + 5.9$	21.9
$\sum_{n=1}^{N} x_n^2$	$0.5^2 + 1.2^2 + 2^2 + 3.1^2 + 4^2 + 5.2^2 + 5.9^2$	93.15
$\sum_{n=1}^{N} y_n$	$4.2 + 4.3 + 4.4 + 4.6 + 4.6 + 4.7 + 4.9$	31.7
$\sum_{n=1}^{N} x_n y_n$	$0.5 \cdot 4.2 + 1.2 \cdot 4.3 + 2 \cdot 4.4 + 3.1 \cdot 4.6 + 4 \cdot 4.6 + 5.2 \cdot 4.7 + 5.9 \cdot 4.9$	102.07

$$\det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & N \end{pmatrix} = \det\begin{pmatrix} 93.15 & 21.9 \\ 21.9 & 7 \end{pmatrix} = 172.44$$

$$\det\begin{pmatrix} \sum_{n=1}^{N} x_n y_n & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} y_n & N \end{pmatrix} = \det\begin{pmatrix} 102.07 & 21.9 \\ 31.7 & 7 \end{pmatrix} = 20.26$$

$$\det\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} y_n \end{pmatrix} = \det\begin{pmatrix} 93.15 & 102.07 \\ 21.9 & 31.7 \end{pmatrix} = 717.522$$

$$k = \frac{20.26}{172.44} = 0.1174901414985$$

$$b = \frac{717.522}{172.44} = 4.1609951287404$$

Ответ: $y = 0.117\,x + 4.160$.

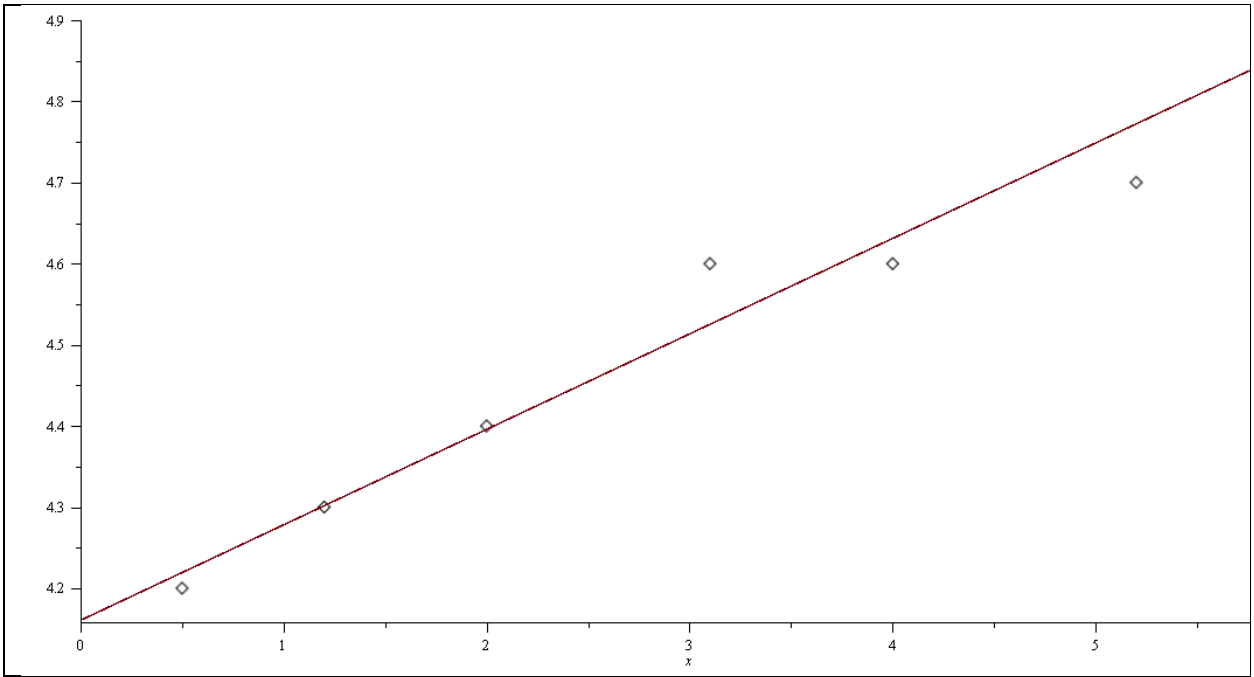

Рис. 9. Зависимость уровня преступности от уровня безработицы. Точками отмечены данные наблюдений.

Замечание. Рутинные вычисления коэффициентов МНК можно препоручить системе компьютерной алгебры. Решение задачи в системе компьютерной алгебры Maple[1] выглядит так.

```
[> with(CurveFitting):

[> P := [[.5, 4.2], [1.2, 4.3], [2, 4.4], [3.1, 4.6], [4, 4.6], [5.2, 4.7], [5.9, 4.9]];

   [[0.5, 4.2], [1.2, 4.3], [2, 4.4], [3.1, 4.6], [4, 4.6], [5.2, 4.7], [5.9, 4.9]]

[> y := LeastSquares(P, x);

   4.160995128740432 + 0.11749014149849218 x
```

Построение приведенного выше графика выполняется так.

```
[> with(plots):

[> display(pointplot(P), plot(y, x = 0 .. 6));
```

Домашнее задание

По территориям региона приводятся данные за 199X г.

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y

[1] Url: www.maplesoft.com.

1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Подберите по МНК коэффициенты зависимости $y = kx + b$.