Метод наименьших квадратов

Метод наименьших квадратов (МНК) – метод, основанный на поиске минимума суммы квадратов отклонений значений некоторых функций от заданного множества значений.

МНК используется при решении различных задач. Например, при решении переопределенных систем линейных уравнений, решении систем нелинейных уравнений, при аппроксимации экспериментальных данных и т.д..

В том числе, МНК является одним из основных методов регрессионного анализа и применяется для оценки параметров регрессионных моделей на основе выборочных данных.

$$M(X / y) = q_1(y)$$
$$M(Y / x) = q_2(x) \qquad (6.1)$$

Метод наименьших квадратов

Рассмотрим, например, следующую ситуацию:

- задано m значений $y_i, i = 1,2,...,m$;
- задано m функций $f_i(x_1, x_2, ..., x_n), i = 1,2,...,m;\ n < m$
- требуется найти такие значения $x_1^*, x_2^*, ..., x_n^*$, при которых сумма квадратов отклонений значений функций $f_i(x_1, x_2, ..., x_n)$ от $y_i, i = 1,2,...,m$ была бы минимально возможной.

По сути требуется найти решение системы уравнений:

$$f_i(x_1, x_2, ..., x_n) = y_i,\ i = 1,2,...,m \qquad (6.2)$$

или (что эквивалентно) решить оптимизационную задачу:

$$\Phi(x_1^*, x_2^*, ..., x_n^*) = \min_{x \in X} \left\{ \sum_{i=1}^{m} \left(f_i(x_1, x_2, ...x_n) - y_i \right)^2 \right\} \qquad (6.3)$$

Метод наименьших квадратов

Рассмотрим другую ситуацию:

- задано m пар значений $(x_i, y_i), i = 1,2,...,m$;
- задана функция $f(x,a,b,c)$;
- требуется найти такие значения a^*, b^*, c^* , при которых сумма квадратов отклонений значений функции $f(x,a,b,c)$ от y_i, при $x = x_i, i = 1,2,...,m$ была бы минимально возможной.

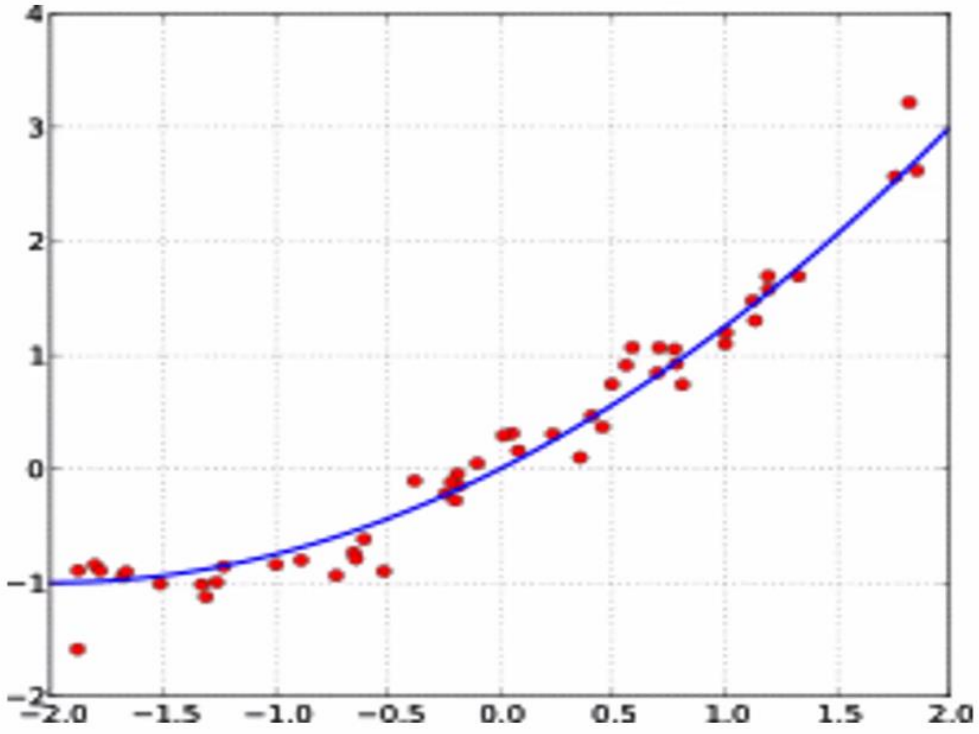
Здесь требуется найти решение системы уравнений:

$$f(x_i,a,b,c) = y_i,\ i = 1,2,...,m \qquad (6.4)$$

или (что эквивалентно) решить оптимизационную задачу:

$$\Phi(a^*,b^*,c^*) = \min_{a,b,c}\left\{\sum_{i=1}^{m}\left(f(x_i,a,b,c) - y_i\right)^2\right\} \qquad (6.5)$$

Метод наименьших квадратов



Линейная среднеквадратическая регрессия

Пусть имеется двумерная случайная величина $\{X, Y\}$, где X и Y зависимые случайные величины.

Представим приближенно случайную величину Y как линейную функцию случайной величины X:

$$Y \simeq g(x) = ax + b \qquad (6.6)$$

Значения коэффициентов a и b определим с помощью МНК из условия минимума функции

$$F(a,b) = m[(Y - aX - b)^2] \qquad (6.7)$$

В этом случае функцию $g(x)$ называют линейной функцией среднеквадратической регрессии Y на X.

$$g(x) = m(Y / x) = m(Y) + r_{xy} \frac{\sigma_y}{\sigma_x}[x - m(X)] \qquad (6.8)$$

Линейная среднеквадратическая регрессия

Коэффициент $\rho_{yx} = r_{xy} \frac{\sigma_y}{\sigma_x}$ называют коэффициентом регрессии Y на X.

Минимальное значение функции $F(a,b)$ равно $\sigma_y^2(1 - r_{xy}^2)$ и называется остаточной дисперсией случайной величины Y относительно случайной величины X.

Линейная функция среднеквадратической регрессии X на Y строится аналогично:

$$q(y) = m(X / y) = m(X) + r_{xy} \frac{\sigma_x}{\sigma_y}[y - m(Y)] \qquad (6.9)$$

$\rho_{xy} = r_{xy} \frac{\sigma_x}{\sigma_y}$ - коэффициент регрессии X на Y, остаточная дисперсия равна $\sigma_x^2(1 - r_{xy}^2)$

Выборочные прямые среднеквадратической регрессии

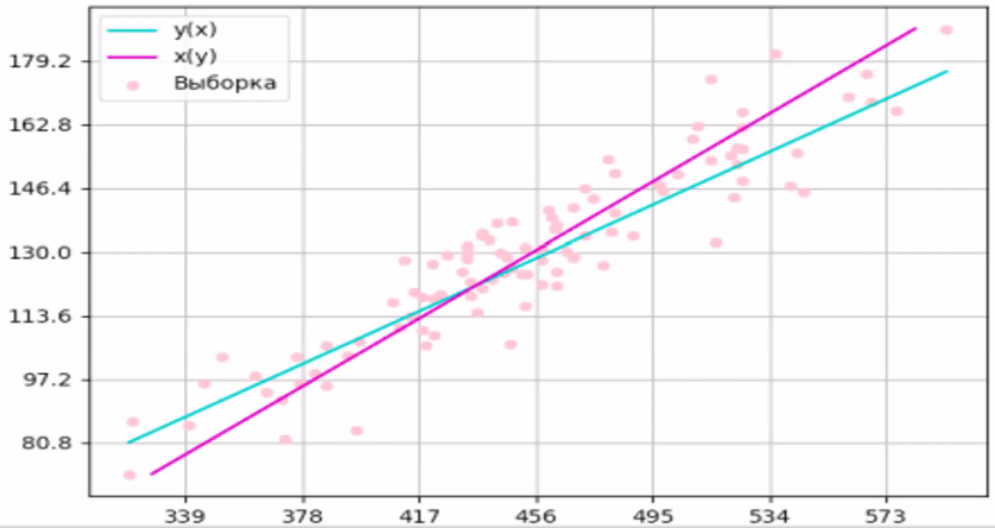
В случае, когда известны только выборочные данные - двумерная выборка значений случайных величин X и Y, возможно построение только выборочных прямых среднеквадратической регрессии.
Уравнения выборочных прямых среднеквадратической регрессии получаются на основе уравнений линейных среднеквадратических регрессий (6,8), (6.9), в которых все параметры распределений заменяются их статистическими оценками:

$$\overline{y}_x = \overline{y}_в + \overline{r}_{xy}\frac{S_y}{S_x}(x - \overline{x}_в) \qquad (6.10)$$

Выборочные прямые среднеквадратической регрессии

$$\overline{x}_y = \overline{x}_в + \overline{r}_{xy}\frac{S_x}{S_y}(y - \overline{y}_в) \qquad (6.11)$$

Аналогично вычисляются статистические оценки остаточной дисперсии.



Корреляционное отношение

Характер корреляционной зависимости может отличаться от линейной.

Для оценки корреляционной зависимости между случайными величинами в общем, а не только линейной, может быть использовано так называемое **корреляционное отношение**.

Пусть по-прежнему имеется выборка, сформированная для двумерной случайной величины $\{X, Y\}$, представленная соответствующей корреляционной таблицей:

Корреляционное отношение

	X			
Y	3	7	n_{x_i}	$\overline{x}_{y_i}$
1	5	7	12	5,3
5	3	5	7	6,3
8	4	2	6	4,3
n_{x_j}	12	13	25	
$\overline{y}_{x_j}$	4,3	3,7		

В этой таблице выборочные данные естественным образом разбиты на несколько групп (сгруппированы).

В таблице ошибки арифметические, на да ладно.

Корреляционное отношение

Каждую из этих групп данных можно (при определенных условиях) рассматривать как отдельную выборку и определить для каждой группы групповую выборочную среднюю и групповую выборочную дисперсию.

Определим $D_{внгр}$ - внутригрупповую дисперсию, как взвешенную по объемам групп среднюю арифметическую групповых дисперсий.

Определим $D_{межгр}$ - межгрупповую дисперсию, как дисперсию условных (групповых) средних $\overline{x}_{y_i}$ относительно выборочной средней $\overline{x}_в$.

Оценку общей дисперсии X можно представить, как сумму внутригрупповой и межгрупповой дисперсий:

$$D_{общ} = D_{межгр} + D_{внгр} \qquad (6.12)$$

Корреляционное отношение

Пусть $D_{межгр} = 0$. Это означает что все условные (групповые) средние равны и не зависят от выборочных значений y. В этом случае выборочные данные не дают оснований предположить, что между исследуемыми случайными величинами имеет место корреляционная зависимость.

Пусть $D_{внгр} = 0$. Это означает, что в каждой из групп (в данном случае в каждой строке корреляционной таблицы) имеется только одно значение X. В этом случае выборочные данные позволяют предположить наличие функциональной зависимости между исследуемыми случайными величинами.

Корреляционное отношение

Определим $\overline{\eta}_{xy}$ - выборочное корреляционное отношение X к Y в соответствии с выражением:

$$\overline{\eta}_{xy} = \frac{\overline{\sigma}_{\overline{x}_y}}{\overline{\sigma}_x} \qquad (6.13)$$

где $\overline{\sigma}_{\overline{x}_y} = \sqrt{D_{межгр}}$; $\overline{\sigma}_x = \sqrt{D_{общ}}$ - выборочные значения

СКВО $\overline{x}_y$ и X соответственно.

Аналогично определяется выборочное корреляционное отношение Y к X:

$$\eta_{yx} = \frac{\overline{\sigma}_{\overline{y}_x}}{\overline{\sigma}_y} \qquad (6.14)$$

как для $\overline{\eta}_{xy}$, так и для $\overline{\eta}_{yx}$ имеет место равенство:

$$D_{внгр} = D_{общ}\left(1 - \frac{D_{межгр}}{D_{общ}}\right) = D_{общ}\left(1 - \eta^2\right) \qquad (*)$$

Корреляционное отношение

Свойства выборочного корреляционного отношения (имеют место как для $\overline{\eta}_{yx}$, так и для $\overline{\eta}_{xy}$):

1. $0 \le \overline{\eta} \le 1$;
2. $\overline{\eta} = 0$ - выборочные данные согласованы с предположением, что случайные величины X и Y не связаны корреляционной зависимостью;
3. $\overline{\eta} = 1$ - выборочные данные согласованы с предположением, что случайные величины X и Y связаны функциональной зависимостью;
4. $\overline{\eta} \ge |\overline{r}_{xy}|$;
5. $\overline{\eta} = |\overline{r}_{xy}|$ - выборочные данные согласованы с предположением, что случайные величины X и Y связаны **линейной** корреляционной зависимостью;
6. Корреляционное отношение является количественной мерой тесноты корреляционной зависимости между случайными величинами X и Y. Вместе с тем, кроме п.5, оно не позволяет определить характер корреляционной зависимости

Построение уравнений выборочных кривых для параболической среднеквадратической регрессии

Запишем выборочное уравнение регрессии Y на X в виде:

$$\overline{y}_x = ax^2 + bx + c \qquad (6.15)$$

Значения коэффициентов a, b и c определим с помощью МНК, что приводит к необходимости решения системы линейных уравнений третьего порядка:

$$\begin{cases} \left(\sum_{i=1}^{m} n_{x_i} x_i^4\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) b + \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) c = \sum_{i=1}^{m} n_{x_i} \overline{y}_{x_i} x_i^2 \\ \left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) b + \left(\sum_{i=1}^{m} n_{x_i} x_i\right) c = \sum_{i=1}^{m} n_{x_i} \overline{y}_{x_i} x_i \\ \left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) a + \left(\sum_{i=1}^{m} n_{x_i} x_i\right) b + Nc = \sum_{i=1}^{m} n_{x_i} \overline{y}_{x_i} \end{cases} \qquad (6.16)$$

Построение уравнений выборочных кривых для параболической среднеквадратической регрессии

Y	X			n_y	$\overline{x}_y$
	2	3	5		
25	20	—	—	20	20
45	—	30	1	31	3.06
110	—	1	48	49	4.96
n_x	20	31	49	$n = 100$	
$\overline{y}_x$	25	47.1	108.67		

Построение уравнений выборочных кривых для параболической среднеквадратической регрессии

x	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
2	20	25	40	80	160	320	500	1 000	2 000
3	31	47,1	93	279	837	2 511	1460	4 380	13 141
5	49	108,67	245	1225	6125	30 625	5325	26 624	133 121
Σ	100		378	1584	7122	33 456	7285	32 004	148 262