4_3: МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Пусть в результате измерений получена таблица некоторых значений функции $f(x)$:

x	x_1	x_2	…	x_n
$f(x)$	y_1	y_2	…	y_n

Нужно найти формулу, выражающую эту зависимость аналитически. Можно применить метод интерполяции: построить интерполяционный многочлен, значения которого в точках $x_1, x_2, ..., x_n$, будут совпадать с соответствующими значениями функции $f(x)$ из таблицы. Однако совпадение значений в узлах может не означать совпадения характеров поведения исходной и интерполирующей функций.

Поставим задачу так, чтобы обязательно учитывался характер поведения табличной функции: найти функцию заданного вида

$$y = F(x), \qquad (1)$$

которая в точках $x_1, x_2, ..., x_n$ принимает значения **как можно более близкие** к табличным значениям $y_1, y_2, ..., y_n$.

Вид приближающей функции F можно определить, построив точечный график функции $f(x)$, и проведя плавную кривую, **наилучшим образом** отражающую характер расположения точек (рис. 1).

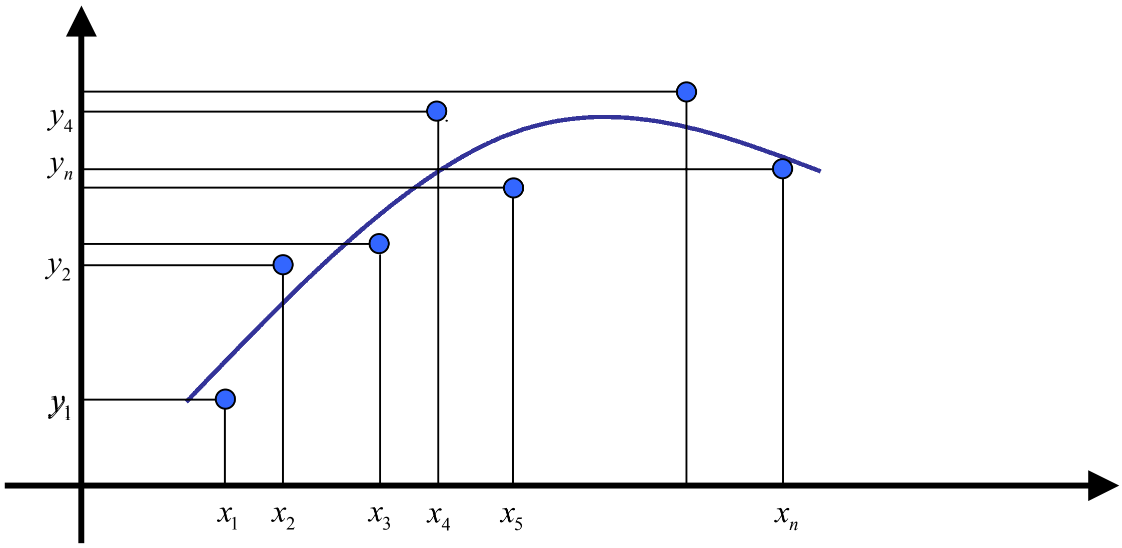

Рис. 1

Формула (1) называется эмпирической формулой или **уравнением регрессии** y на x. Формула позволяет находить значения функции $f(x)$ для нетабличных значений x, "сглаживая" результаты измерений величины y.

Рассмотрим один из распространенных способов нахождения формулы (1). Предположим, что приближающая функция F в точках $x_1, x_2, ..., x_n$ принимает значения

$$\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_n \qquad (2)$$

Требование близости табличных значений $y_1, y_2, ..., y_n$ и значений (2) можно истолковать следующим образом.

Будем рассматривать совокупность значений функции f из таблицы и совокупности (2), как координаты двух точек $n-$ мерного пространства. Тогда задача приближения функции может быть сформулирована так: найти такую функцию F заданного вида, чтобы **расстояние** между точками $M(y_1, y_2, ..., y_n)$ и $\overline{M}(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_n)$ **было наименьшим**. В Евклидовом пространстве приходим к требованию: величина

$$\sqrt{(y_1-\overline{y}_1)^2+(y_2-\overline{y}_2)^2+\ldots+(y_n-\overline{y}_n)^2} \qquad (3)$$

минимальна, что равносильно следующему: сумма квадратов

$$(y_1-\overline{y}_1)^2+(y_2-\overline{y}_2)^2+\ldots+(y_n-\overline{y}_n)^2 \qquad (4)$$

должна быть **наименьшей**.

Окончательно задачу приближения функции f сформулируем следующим образом: для функции f, заданной таблицей, найти функцию F определенного вида так, чтобы сумма квадратов (4) была наименьшей. Задача носит название приближения функции **методом наименьших квадратов**.

В качестве приближающих функций в зависимости от характера точечного графика функции f используют следующие функции:

1. $y = ax + b$
2. $y = ax^2 + bx + c$
3. $y = ax^m$
4. $y = ae^{mx}$
5. $y = \dfrac{1}{ax+b}$
6. $y = a\ln x + b$
7. $y = a\dfrac{1}{x} + b$
8. $y = \dfrac{x}{ax+b}$

Здесь a, b, c, m – параметры. Когда вид приближающей функции установлен, задача сводится к отысканию значений параметров.

Рассмотрим метод нахождения параметров приближающей функции в общем виде на примере функции с тремя параметрами:

$$y = F(x, a, b, c) \qquad (5)$$

Составим сумму квадратов разностей $\sum_{i=1}^{n}(y_i - F(x_i, a, b, c))^2 = \Phi(a, b, c)$

Эта сумма является функцией $\Phi(a,b,c)$ трех переменных a, b, c. Задача сводится к отысканию ее минимума.

Необходимые условия экстремума:

$$\frac{\partial \Phi}{\partial a} = 0 \qquad \frac{\partial \Phi}{\partial b} = 0 \qquad \frac{\partial \Phi}{\partial c} = 0, \qquad \text{т. е.}$$

$$\begin{cases} \sum_{i=1}^{n}[y_i - F(x_i, a, b, c)] \cdot F'_a(x_i, a, b, c) = 0 \\ \sum_{i=1}^{n}[y_i - F(x_i, a, b, c)] \cdot F'_b(x_i, a, b, c) = 0 \\ \sum_{i=1}^{n}[y_i - F(x_i, a, b, c)] \cdot F'_c(x_i, a, b, c) = 0 \end{cases} \qquad (6)$$

Решив эту систему трех уравнений с тремя неизвестными относительно параметров a, b, c, получим конкретный вид искомой функции $y = F(x, a, b, c)$.

Значения найденной функции $y = F(x,a,b,c)$ в точках $x_1, x_2, ..., x_n$ будут отличаться от табличных значений $y_1, y_2, ..., y_n$.

Разности $y_i - F(x_i, a, b, c) = \varepsilon_i \qquad (i = 1, 2, ..., n) \qquad (7)$

называются отклонениями измеренных значений от вычисленных по (5).

Для найденной эмпирической формулы (5) в соответствии с исходной таблицей можно найти сумму квадратов отклонений:

$$\sigma = \sum_{i=1}^{n}\varepsilon_i^2, \qquad (8)$$

которая в соответствии с принципом наименьших квадратов для заданного вида приближающей функции и найденных параметров a,b,c должна быть наименьшей. Из двух разных приближений одной и той же табличной функции, следуя принципу наименьших квадратов, лучшим нужно считать то, для которого сумма (8) имеет наименьшее значение.

Линейная регрессия

Будем искать приближающую функцию в виде

$$F(x,a,b) = ax + b \qquad (9)$$

Частные производные по параметрам равны: $\dfrac{\partial F}{\partial a} = x$; $\quad \dfrac{\partial F}{\partial b} = 1$

Составим систему:

$$\begin{cases} \sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases}.$$

Преобразуем

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0 \end{cases} \qquad (10)$$

Разделим каждое уравнение на n:

$$\begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)a + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)b = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)a + b = \frac{1}{n}\sum_{i=1}^{n} y_i \end{cases}$$

Обозначим:

$$M_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad M_y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$M_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad M_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad (11)$$

Тогда система (10) будет иметь вид:

$$\begin{cases} M_{x^2}\cdot a + M_x \cdot b = M_{xy} \\ M_x \cdot a + \qquad b = M_y \end{cases} \qquad (12)$$

Коэффициенты этой системы M_x, M_y, M_{x^2}, M_{xy} – числа, которые в каждой конкретной задаче приближения могут быть легко выражены по формулам (11), где x_i, y_i – значения из таблицы. Решив систему (12) (метод Крамера), получим значения параметров a и b, а, следовательно, и конкретный вид линейной функции.

Можно показать, что нахождение приближающей функции в виде функции с двумя параметрами из приведенного списка сводится к линейной функции.