Криволинейная корреляция

Если линейная аппроксимация статистической зависимости между двумя величинами не отражает характер зависимости, используют модель *криволинейной корреляции*. Одной из распространенных является параболическая корреляция второго порядка, при которой уравнение регрессии *Y* на *X* имеет вид:

$$y(x) = a_0 + a_1 x + a_2 x^2.$$

На практике выборка совместного распределения случайных величин *X* и *Y* возникает как последовательность пар $(x_1, y_1), \dots, (x_n, y_n)$, перечисленных в порядке произведенных наблюдений, среди них могут быть и одинаковые. Для нахождения коэффициентов регрессии не обязательно группировать данные в корреляционную таблицу.

Как и в случае линейной корреляции, коэффициенты регрессии a_0, a_1, a_2 найдем из условия минимума функционала:

$$F(a_0, a_1, a_2) = \sum_{i=1}^{n} \left(a_0 + a_1 x_i + a_2 x_i^2 - y_i\right)^2 \to min.$$

Условием минимума является обращение в нуль частных производных:

$$\frac{\partial F}{\partial a_0} = 2 \sum_{i=1}^{n} \left(a_0 + a_1 x_j + a_2 x_j^2 - y_i\right) = 0,$$

$$\frac{\partial F}{\partial a_1} = 2 \sum_{i=1}^{n} x_i \left(a_0 + a_1 x_j + a_2 x_j^2 - y_i\right) = 0,$$

$$\frac{\partial F}{\partial a_2} = 2 \sum_{i=1}^{n} x_i^2 \left(a_0 + a_1 x_j + a_2 x_j^2 - y_i\right) = 0.$$

Это дает систему трех линейных уравнений относительно трех неизвестных a_0, a_1, a_2, которая называется системой *нормальных уравнений*:

$$\begin{cases} n\,a_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 + \left(\sum_{i=1}^{n} x_i^2\right) a_2 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^2\right) a_1 + \left(\sum_{i=1}^{n} x_i^3\right) a_2 = \sum_{i=1}^{n} y_i x_i \\ \left(\sum_{i=1}^{n} x_i^2\right) a_0 + \left(\sum_{i=1}^{n} x_i^3\right) a_1 + \left(\sum_{i=1}^{n} x_i^4\right) a_2 = \sum_{i=1}^{n} y_i x_i^2 \end{cases}$$

Решая ее, получаем уравнение регрессии.

Отметим, что если ввести матрицу A и векторы y и a:

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \qquad a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix},$$

то в матричном виде систему нормальных уравнений можно записать как

$$A'Aa = A'y,$$

где A' – матрица, получаемая из матрицы A транспонированием.

Такая запись системы нормальных уравнений облегчает ее запоминание. Она переносится и на рассматриваемую далее множественную корреляцию.

Пример. Желая установить цену на товар, обеспечивающую максимальную прибыль, магазин в течении 5 рабочих дней недели продавал получаемые от поставщика изделия с наценкой 1, 2, 3, 4 и 5 у.е. При этом в каждый из дней было продано соответственно 100, 80, 60, 30 и 10 единиц товара. С помощью модели параболической регрессии второго порядка выбрать надбавку, дающую максимальную прибыль.

Решение. Выпишем таблицу соответствия между наценкой и полученной прибылью, определяемой как произведение наценки на количество проданного товара.

наценка X	1	2	3	4	5
прибыль Y	100	160	180	120	50

Заметим, что устанавливаемая оценка по смыслу является величиной неслучайной. Прибыль, определяемая количеством проданного товара, напротив, величина случайная, среднее значение которой зависит от наценки. Уравнение регрессии Y на X ищем в виде:

$$\bar{y}_x = a_0 + a_1 x + a_2 x^2.$$

Из полученной таблицы находим коэффициенты системы нормальных уравнений:

$$n = 5;$$

$$\sum_{i=1}^{5} x_i = 15, \qquad \sum_{i=1}^{5} x_i^2 = 55, \qquad \sum_{i=1}^{5} x_i^3 = 225, \qquad \sum_{i=1}^{5} x_i^4 = 979;$$

$$\sum_{i=1}^{5} y_i = 610, \qquad \sum_{i=1}^{5} y_i x_i = 1690, \qquad \sum_{i=1}^{5} y_i x_i^2 = 5530.$$

Система нормальных уравнений запишется в виде:

$$\begin{cases} 5a_0 + 15a_1 + 55a_2 = 610 \\ 15a_0 + 55a_1 + 225a_2 = 1690 \\ 55a_0 + 225a_1 + 979a_2 = 5530. \end{cases}$$

Произведя сокращение на 5, получим систему:

$$\begin{cases} a_0 + 3a_1 + 11a_2 = 122 \\ 3a_0 + 11a_1 + 45a_2 = 338 \\ 11a_0 + 45a_1 + 195{,}8\,a_2 = 1106, \end{cases}$$

которую будем решать методом Гаусса.

$$\begin{cases} a_0 + 3a_1 + 11a_2 = 122 \\ 2a_1 + 12a_2 = -28 \\ 12a_1 + 74{,}8\,a_2 = -236; \end{cases}$$

$$\begin{cases} a_0 + 3a_1 + 11a_2 = 122 \\ 2a_1 + 12a_2 = -28 \\ 2{,}8\,a_2 = -68; \end{cases}$$

$$\begin{cases} a_0 \approx -5{,}8 \\ a_1 \approx 131{,}7 \\ a_2 \approx -24{,}3. \end{cases}$$

Выборочное уравнение регрессии примет вид:

$$\bar{y}_x = -5{,}8 + 131{,}7x - 24{,}3x^2$$

Даваемая моделью оптимальная наценка равна

$$x_{\text{опт}} = -a_1/2a_2 \approx 2{,}7$$

а получаемая при такой наценке средняя ежедневная прибыль

$$\bar{y}_{max} \approx -5{,}8 + 131{,}7 \cdot 2{,}7 - 24{,}3 \cdot (2{,}7)^2 \approx 173$$

Вычисленная по модели максимальная средняя ежедневная прибыль оказалась несколько меньше прибыли, полученной в день, когда наценка была равна 3. Это не должно вызывать недоумения. Согласно модели этот день был скорее случайной удачей, чем правилом.

На графике представлены значения полученных прибылей при различных наценках и полученная по ним параболическая линия регрессии.