Множественная корреляция

Пусть случайная величина Y зависит от величин $X_1, \dots, X_n$. Такую корреляцию называют *множественной*. Уравнение линейной множественной регрессии ищется в виде:

$$y = a_0 + a_1 x_1 + \cdots + a_l x_l.$$

Используемая выборка состоит из n наборов $x_{i1}, \dots, x_{il}, i = 1, \dots, n,$ величин $X_1, \dots X_l$ и соответствующих значений y_i величины Y, где $n \geq l + 1$. Коэффициенты $a_0, a_1, \dots, a_l$ находятся по выборке методом наименьших квадратов.

Как и в случае линейной парной регрессии, средние значения $\bar{y}, \bar{x}_1, \dots, \bar{x}_l$ должны удовлетворять этом уравнению:

$$\bar{y} = a_0 + a_1 \bar{x}_1 + \cdots + a_l \bar{x}_l.$$

Это позволяет, исключив коэффициент a_0, записать уравнение регрессии в виде:

$$y - \bar{y} = a_1(x_1 - \bar{x}_1) + \cdots + a_l(x_1 - \bar{x}_l).$$

Такая запись уравнения весьма удобна и позволяет понизить на единицу порядок системы нормальных уравнений.

Пример. В течение 7 месяцев фирма давала рекламу своего товара по телевидению и в печати. Ежемесячные расходы на рекламу ($X_1 - TV, X_2 - \text{печать}$), а также доход фирмы от продажи товара (Y) в тыс. у.е. сведены в таблице:

X_1	X_2	Y
100	100	500
140	100	550
100	140	570
120	120	570
140	100	560
100	140	580
140	140	590

Получить по таблице уравнение регрессии

$$y = a_0 + a_1 x_1 + a_2 x_2,$$

на основании которого предложить эффективную рекламную политику.

Решение. Уравнение регрессии будем искать в виде

$$y-\bar{y}=a_1(x_1-\bar{x}_1)+a_2(x_2-\bar{x}_2).$$

Из таблицы находим: $\bar{x}_1=120,\ \bar{x}_2=120,\ \bar{y}=560.$ Переопределенная система линейных уравнений, даваемая выборкой, примет вид:

$$\begin{cases}-20a_1-20a_2=-60\\ 20a_1-20a_2=-10\\ -20a_1+20a_2=10\\ 0a_1+0a_2=10\\ 20a_1-20a_2=0\\ -20a_1+20a_2=20\\ 20a_1+20a_2=30.\end{cases}$$

После сокращения и удаления уравнения, не содержащего неизвестных, получаем:

$$\begin{cases}-2a_1-2a_2=-6\\ 2a_1-2a_2=-1\\ -2a_1+2a_2=1\\ 2a_1-2a_2=0\\ -2a_1+2a_2=2\\ 2a_1+2a_2=3.\end{cases}$$

Соответствующая нормальная система запишется в виде:

$$\begin{cases}24a_1-8a_2=10\\ -8a_1+24a_2=26.\end{cases}$$

Ее решение: $a_1=7/8,\ a_2=17/16.$ Полученные значения коэффициентов регрессии свидетельствуют о том, что реклама по телевидению убыточна $(a_1<1)$, а реклама в печати, наоборот, приносит некоторый доход $(a_2>1)$. Поэтому относительно среднего уровня $(120\ \text{тыс.}\,у.е.)$ вложения в рекламу по телевидению следует снизить, направив освободившиеся средства на рекламу в печати.

4.Метод наименьших квадратов

Пусть величина Y является линейной комбинацией величин $X_1,\dots,X_l$:

$$Y=a_1X_1+\cdots+a_lX_l,$$

неизвестные коэффициенты $a_1, \dots, a_l$ которой нужно найти. Для этого величинам $X_1, \dots, X_n$ придается *n* наборов значений и измеряются соответствующие значения *Y*. Это дает для определения $a_1, \dots, a_l$ следующую систему линейных уравнений:

$$\begin{cases} x_{11}a_1 + x_{12}a_2 + \cdots + x_{1l}a_l = y_1 \\ x_{21}a_1 + x_{22}a_2 + \cdots + x_{2l}a_l = y_2 \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ x_{n1}a_1 + x_{n2}a_2 + \cdots + x_{nl}a_l = y_{n,} \end{cases}$$

где x_{ij} обозначает значение величины X_j в $i -$ ом опыте.

Минимальное число необходимых для этого уравнений *n* равно *l*. Если определитель системы отличен от нуля, что обычно и имеет место на практике, то система имеет при $n = l$ единственное решение. Если же число уравнений *n* больше числа неизвестных *l*, то так как любые *n* из уравнений системы являются независимыми, а остальные $n - l$ – их следствиями, теоретически можно выбрать любую подсистему из *l* уравнений и решить ее. На практике, однако, каждое измерение величины *Y* неизбежно связано с погрешностью. Это приводит к тому, что система при $n > l$ оказывается несовместной. Если же из нее выбрать подсистему из *l* уравнений, то полученные значения коэффициентов $a_1, \dots, a_l$ будут зависеть от этого выбора.

Для разрешения данной ситуации еще в начале XIX века немецким математиком Гауссом и французским математиком Лежандром был предложен прием, получивший название *метода наименьших квадратов*, который стал одним из основных способов обработки экспериментальных данных. Фактически, этот прием уже использовался нами при определении коэффициентов линейной и параболической парной корреляции. Теперь этот важный метод будет рассмотрен в общем виде.

Уравнения системы пытаются удовлетворить приближенно. В качестве меры близости берется сумма квадратичных уклонений левых частей от свободных членов. Решением по методу наименьших квадратов называется набор $a_1, \dots, a_l$, доставляющий минимум функционала

$$F(a_1, \dots, a_l) = \sum_{i=1}^{n} (x_{11}a_1 + x_{12}a_2 + x_{1l}a_l - y_1)^2 \to min.$$

Отметим, что если система допускает точное решение, то минимальное значение *F* оказывается равным нулю, и решение по методу наименьших квадратов является точным решением. Практически же для более точного нахождения неизвестных коэффициентов систему стараются переопределить как можно сильнее, увеличивая число уравнений *n*. Если ошибку в измерении величины Y считать, как обычно делается в теории ошибок, нормально распределенной случайной величиной с нулевым математическим ожиданием, то такой метод может быть обоснован теоретически как доставляющий значения $a_1, \dots, a_l$, наиболее близкие к их действительным значениям.

Условия минимума F является равенство нулю частных производных:

$$\frac{\partial F}{\partial a_j = 0}, \qquad j = 0,1, \dots, l,$$

что дает для определения $a_1, \dots, a_l$ систему l линейных уравнений с l неизвестными, которая называется *системой нормальных уравнений.*

Если ввести матрицу A исходной системы уравнений, вектор-столбец свободных членов y и вектор-столбец неизвестных a:

$$A = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \qquad a = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_l \end{pmatrix},$$

то в матричном виде систему нормальных уравнений можно записать как

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

где A' − матрица, получаемая из матрицы А транспонированием.

Матрица $A'A$ нормальной системы является квадратной симметрической $(l \times l)$ матрицей. Ее (ij) − ый и(ji) − ый элементы равны скалярному произведению i-го и j-го столбцов матрицы A.

Пример: Дана система точек, координаты которых указаны в таблице, число точек $n = 6$.

x	−1	0	1	2	3	4
y	0	2	3	3,5	3	4,5

Требуется построить прямую с уравнением $y = ax + b$.

Решение: Очевидно, что точки с данными координатами не могут быть расположены на одной прямой, а построить прямую как бы «сглаживающую» эти точки, можно. Для этого достаточно решить систему уравнений, приведенную в соответствующей теоретической части. Для удобства расчетов строим рабочую таблицу:

№	x_i	y_i	x_i^2	$x_i y_i$	ax_i+b	ax_i+b-y_i	$(ax_i+b-y_i)^2$
1	-1	0	1	0	0,81	0,81	0,6561
2	0	2	0	0	1,55	$-0{,}45$	0,2025
3	1	3	1	3	2,29	$-0{,}71$	0,5041
4	2	3,5	4	7	3,03	$-0{,}47$	0,2209
5	3	3	9	9	3,77	0,77	0,5929
6	4	4,5	16	18	4,51	0,01	0,0001
$\sum$	9 A_2, B_1	16 C_2	31 A_1	37 C_1			2,1766

Первый столбец обозначает номер по порядку записи точек (координат). Из сумм столбцов при $x_i, y_i, y_i^2, x_i y_i$ составляются коэффициенты системы

$$\begin{cases} A_1 a + B_1 b = C_1, \\ A_2 a + B_2 b = C_2, \end{cases}$$

где $A_1 = \sum_{i=1}^{n} x_i^2$ $, B_1 = \sum_{i=1}^{n} x_1$ $, C_1 = \sum_{i=1}^{n} x_1 y_1$ $, A_2 = B_1 = \sum_{i=1}^{n} x_1$ $, B_2 = n,$

$$C_2 = \sum_{i=1}^{n} y_i.$$

Для определения параметров a и b прямой $y = ax + b$. Система имеет вид:

$$\begin{cases} 31a + 9b = 37, \\ 9a + 6b = 16. \end{cases}$$

Решим ее методом определений:

$$\Delta = \begin{vmatrix} 31 & 9 \\ 9 & 6 \end{vmatrix} = 105, \qquad \Delta_1 = \begin{vmatrix} 37 & 9 \\ 16 & 6 \end{vmatrix} = 78,\ \Delta_2 = \begin{vmatrix} 31 & 37 \\ 9 & 16 \end{vmatrix} = 163,$$

$$a_1 = \frac{\Delta_1}{\Delta} = \frac{78}{105} = 0{,}74, b = \frac{\Delta_2}{\Delta} = \frac{163}{105} = 1{,}55.$$

Искомое уравнение $y = 0{,}74x + 1{,}55$.