Глава 9. Регрессионный анализ
9.1. Задачи регрессионного анализа

Во время статистических наблюдений, как правило, получают значения нескольких признаков. Для простоты будем рассматривать в дальнейшем двумерные выборки X и Y. Результаты измерений записывают в таблицы, а затем их анализируют для того, чтобы установить связи между переменными. Связь между случайными величинами часто носит случайный характер. Такая связь называется **стохастической** или **статистической**, если изменение одной величины вызывает изменение распределения другой величины. Если среднее значение одной случайной величины функционально зависит от значений другой случайной величины, то такая статистическая зависимость называется регрессией:

$$M(Y|X = x) = f(x) \qquad \text{(регрессия } Y \text{ на } X\text{);}$$

$$M(X|Y = y) = g(y) \qquad \text{(регрессия } X \text{ на } Y\text{).}$$

Так как законы распределения случайных величин неизвестны, то находят их приближенные значения (оценки); например, в качестве оценки условного математического ожидания находят условное среднее.

Условным средним $\bar{y}_x$ называется среднее арифметическое наблюдаемых значений Y, полученных при одном и том же значении $X = x$.

Пример. Если в результате измерений получена таблица

X	1	2	3	2	0	2	1
Y	3	4	6	7	1	4	2

то условное среднее для $x = 2$

$$\bar{y}_{x=2} = \frac{4+7+4}{3} = 5\,.$$

Ответ. $\bar{y}_{x=2} = 5$.

Условные средние $\bar{y}_x$ и $\bar{x}_y$ являются функциями соответственно от x и y, т.е.

$$\bar{y}_x = f^*(x),$$

$$\bar{x}_y = g^*(y).$$

Первое уравнение называют выборочным уравнением регрессии *Y* на *X*, а второе уравнение называют выборочным уравнением регрессии *X* на *Y*.

Для того, чтобы найти фактический вид зависимости между случайными величинами результаты наблюдений, записанные в таблице, переносят на координатную плоскость в виде точек, координатами которых являются значения $(x_i; y_i)$, $i = 1, 2, \ldots, n$. (рис.31).

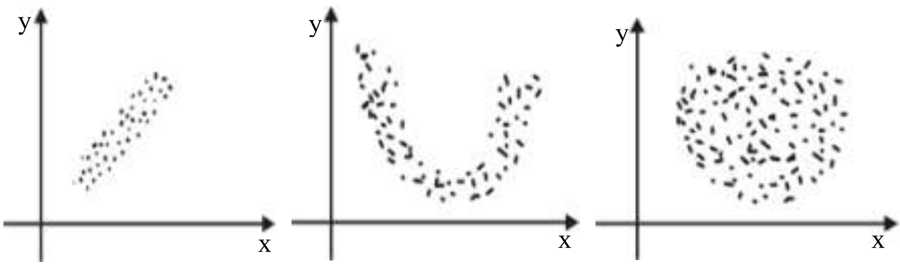

Рис.31

Из рис. 31 видно, что в случае а) зависимость между *X* и *Y* является линейной ($y = ax + b$); в случае б) зависимость между *X* и *Y* является квадратичной ($y = ax^2 + bx + c$); в случае в) между величинами *X* и *Y* зависимости не существует.

На практике часто встречаются следующие виды уравнений регрессии:

$y = ax + b$ линейное;

$y = a_k x^k + a_{k-1} x^{k-1} + \ldots + a_1 x + a_0$ полиномиальное;

$y = a_1 \cdot \frac{1}{x} + b_1$ гиперболическое;

$y = a \cdot e^{bx}$ экспоненциальное.

Оценка неизвестных параметров a, b по результатам выборки объемом n является основной задачей регрессионного анализа.

Для оценки неизвестных параметров уравнения регрессии чаще всего используют метод наименьших квадратов, который позволяет получить несмещенные оценки.

9.2. Определение параметров выборочного уравнения линейной регрессии по несгруппированным данным

Пусть в результате n независимых испытаний получены n пар независимых чисел (x_1; y_1), (x_2; y_2), …, (x_n; y_n).
Заметим, что вид зависимости $y = f(x)$ предполагается заранее известным или из теоретических соображений, или в результате анализа расположения точек (x_i; y_i) на координатной плоскости.

Будем искать линейное выборочное уравнение регрессии Y на X в виде

$$\bar{y}_x = ax + b.$$

По выборочным данным можно получить только приближенные значения (оценки) параметров a и b. Подставим в формулу $y = ax + b$ значения величины $x = x_i$, получим «теоретические» результаты

$$y_i^{\mathrm{T}} = ax_i + b, \ (i = 1, 2, \ldots, n).$$

Коэффициенты a и b найдем методом наименьших квадратов из предположения, что опытные и теоретические результаты мало отличаются между собой. В методе наименьших квадратов условие близости опытных и теоретических данных записывается в виде:

$$\sum_{i=1}^{n} \left(y_i - y_i^T\right)^2 = \min,$$

или, более подробно

$$\sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2 = \min.$$

Введем в рассмотрение функцию

$$\Phi(a, b) = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2.$$

Эта функция достигает минимума при тех значениях a и b, при которых обращаются в нуль частные производные

$$\frac{\partial \Phi}{\partial a} = 0, \ \frac{\partial \Phi}{\partial b} = 0.$$

Продифференцируем функцию $\Phi(a, b)$ по каждой переменной a и b, и приравняем производные нулю. В результате получим систему двух уравнений:

$$\begin{cases} 2\sum\limits_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0 \\ 2\sum\limits_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

Преобразуем эту систему уравнений и запишем ее в виде

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

Определитель этой системы отличен от нуля и, значит, решение всегда существует и единственно.

Решив эту систему, получим значение параметров a и b и можем записать выборочное уравнение линейной регрессии *Y* на *X*.

Аналогично находится выборочное уравнение линейной регрессии *X* на *Y*

$$\bar{x}_y = a_1 y + b_1,$$

где параметры a_1 и b_1 находят из системы уравнений

$$\begin{cases} a_1\left(\sum\limits_{i=1}^{n} y_i^2\right) + b_1\left(\sum\limits_{i=1}^{n} y_i\right) = \sum\limits_{i=1}^{n} x_i y_i \\ a_1\left(\sum\limits_{i=1}^{n} y_i\right) + b_1 \cdot n = \sum\limits_{i=1}^{n} x_i \end{cases}.$$

<u>Пример</u>. В результате некоторых испытаний получены значения (x_i; y_i), которые представлены в таблице

x_i	3	5	6	9	10
y_i	2	5	9	16	14

Найти выборочное уравнение линейной регрессии *Y* на *X*.
Решение. Результаты вычислений записаны в следующей таблице

i	x_i	y_i	x_i^2	$x_i y_i$
1	3	2	9	6
2	5	5	25	25
3	6	9	36	54
4	9	16	81	144
5	10	14	100	140
$\sum$	33	46	251	369

Таким образом, для нахождения коэффициентов a и b уравнения $\bar{y}_X = ax + b$ необходимо решить систему уравнений

$$\begin{cases} 251a + 33b = 369 \\ 33a + 5b = 46 \end{cases}.$$

Отсюда,

$$a = \frac{365 \cdot 9 - 33 \cdot 46}{251 \cdot 5 - 33 \cdot 33} = \frac{1845 - 1518}{1255 - 1089} = \frac{327}{166} = 1{,}97$$

$$b = \frac{251 \cdot 46 - 33 \cdot 369}{251 \cdot 5 - 33 \cdot 33} = \frac{11546 - 12177}{1255 - 1089} = -\frac{631}{166} = -3{,}80 .$$

Ответ. $\bar{y}_X = 1{,}97 - 3{,}80 \cdot x$.

9.3. Определение параметров выборочного уравнения линейной регрессии по сгруппированным данным

При большом числе испытаний пара значений (x_i; y_j) может наблюдаться несколько раз. Значения частот n_{ij} подсчитывают и записывают в двумерную таблицу

x_i \ y_i	y_1	…	y_j	…	y_m
x_1	n_{11}	…	n_{1i}	…	n_{1m}
…	…	…	…	…	…
x_i	n_{i1}	…	n_{ij}	…	n_{im}
…	…	…	…	…	…
x_k	n_{k1}	…	n_{kj}	…	n_{km}

Эту таблицу называют **корреляционной**.
В этом случае объем выборки n находят по формуле

$$n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} .$$

Пусть уравнение регрессии Y на X имеет вид $\bar{y}_x = ax + b$.

Можно показать, что параметры a и b этого уравнения являются решениями системы уравнений

$$\begin{cases} a \cdot \overline{x}^2 + b\overline{x} = \overline{xy} \\ a\overline{x} + b = \overline{y}, \end{cases}$$

где

$$\overline{x} = \frac{1}{n} \cdot \left(\sum_{i=1}^{k} x_i \cdot n_i \right),$$

$$n_i = \sum_{j=1}^{m} n_{ij},$$

$$\overline{y} = \frac{1}{n} \cdot \left(\sum_{j=1}^{k} y_i \cdot n_i \right),$$

$$n_j = \sum_{i=1}^{K} n_{ij},$$

$$\overline{xy} = \frac{1}{n} \cdot \left(\sum_{i=1}^{k} \sum_{j=1}^{m} x_i \cdot y_j \cdot n_{ij} \right),$$

$$\overline{x}^2 = \frac{1}{n} \cdot \left(\sum_{i=1}^{k} x^2{}_i \cdot n_i \right),$$

$$n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} \ .$$

<u>Пример</u>. В результате некоторых испытаний получены значения (x_i; y_i), частоты для которых представлены в следующей таблице:

y_i \ x_i	4	6	8
10	5	-	3
20	-	6	15
30	7	6	-
40	14	4	-

Найти выборочное уравнение линейной регрессии *Y* на *X*.
<u>Решение</u>. На пересечении строк и столбцов данной таблицы записаны частоты n_{ij} для пары (x_i; y_j), (i = 1, 2, 3, 4; j = 1, 2, 3).
Промежуточные вычисления представлены в следующей таблице:

y_i \ x_i	y_i			n_i	$x_i n_i$	$x_i^2 n_i$
	4	6	8			
10	200 5	-	240 3	8	80	800
20	-	720 6	2400 15	21	420	8400
30	840 7	1080 6	-	13	390	11700
40	2240 14	960 4	-	18	720	28800
n_j	26	16	18	Σ = 60	Σ = 16100	Σ = 49700
$y_j n_j$	104	96	144	Σ = 344		

В верхние части клеток с частотами n_{ij} записаны произведения $x_i\ y_i\ n_{ij}$, например, $x_i\ y_i\ n_{ij} = 10\cdot 4\cdot 5 = 200$.
В столбце n_i записаны суммы частот n_{ij} для каждой строки, например, для первой строки 5 + 3 = 8.
В строке n_j записаны суммы частот n_{ij} для каждого столбца, например для первого столбца 5 + 7 + 14 = 26.
В столбце $x_i n_i$ записаны произведения чисел из столбцов $x_i n_i$. В столбце $x^2{}_i n_i$ записаны произведения квадратов чисел из столбца x_i на числа n_i.
В строке $y_j n_j$ записаны произведения чисел из строк $y_j n_j$, например $4\cdot 26 = 104$.
Найдем теперь средние арифметические

$$\bar{x} = \frac{1}{n}\cdot\left(\sum_{i=1}^{k} x_i \cdot n_i\right) = \frac{1610}{60} = 26{,}8333\,,$$

$$\bar{y} = \frac{1}{n}\cdot\left(\sum_{j=1}^{k} y_i \cdot n_i\right) = \frac{344}{60} = 5{,}7333\,,$$

$$\bar{x}^2 = \frac{1}{n}\cdot\left(\sum_{i=1}^{k} x_i^2 \cdot n_i\right) = \frac{49700}{60} = 828{,}3333\,.$$

Для определения величины $\overline{xy}$ сложим числа из верхних частей клеток и разделим сумму на $n = 60$:

$$\overline{xy} = \frac{200+240+720+2400+840+1080+2240+960}{60} = \frac{8680}{60} = 144{,}666\,.$$

Запишем систему уравнений для нахождения коэффициентов a и b.

$$\begin{cases} 828{,}3333a + 26{,}8333b = 144{,}6667 \\ 26{,}8333a + b = 5{,}7333 \end{cases}.$$

Решение системы найдем по формулам Крамера:

$$a = \frac{\Delta a}{\Delta}\,,\ b = \frac{\Delta b}{\Delta}\,,$$

где

$$\Delta = \begin{vmatrix} 828{,}3333 & 26{,}8333 \\ 26{,}8333 & 1 \end{vmatrix} = 108{,}3056$$

$$\Delta_a = \begin{vmatrix} 144{,}6667 & 26{,}8333 \\ 5{,}7333 & 1 \end{vmatrix} = -9{,}1778$$

$$\Delta_b = \begin{vmatrix} 828{,}3333 & 144{,}6667 \\ 26{,}8333 & 1 \end{vmatrix} = 867{,}2222 .$$

В результате получим $a = -0{,}085$; $b = 8{,}007$.

Ответ. $y_x = -0{,}09x + 8{,}01$.

9.4. Выборочный коэффициент корреляции. Выборочное уравнение регрессии для таблиц с постоянной разностью между вариантами

Предположим, что в результате некоторых испытаний получена корреляционная таблица со значениями частот n_{ij} для каждой пары значений (x_i; y_i) случайных величин X и Y. Для оценки связи между случайными величинами обычно используется выборочный коэффициент корреляции

$$r_b = \frac{\sum\limits_{i=1}^{k} \sum\limits_{j=1}^{m} n_{ij} x_i y_j - \bar{x} \cdot n \cdot \bar{y}}{n\sigma_{x_b} \cdot \sigma_{y_b}},$$

где выборочные дисперсии

$$\sigma_{x_b}^2 = \overline{x^2} - (\bar{x})^2, \quad \sigma_{y_b}^2 = \overline{y^2} - (\bar{y})^2 .$$

При этом выборочное уравнение линейной регрессии Y на X можно получать по формуле

$$y_x - \bar{y} = r_b \frac{\sigma_{y_b}}{\sigma_{x_b}} (x - \bar{x}).$$

$$u_i = \frac{x_i - c_1}{\Delta x}; \quad v_j = \frac{y_j - c_2}{\Delta y},$$

где c_1 и c_2 – ложные нули (выбираемые числа, расположенные вблизи середины интервалов, в которых находится все значения выборки). Таким образом, значения u_i и v_i будут малыми по абсолютной величине.

Например, для корреляционной таблицы

Y \ X	15	25	35	45	55	n_x
10	5	-	-	-	-	5
20	7	20	-	-	-	27
30	-	23	30	10		63
40	-	-	47	11	9	67
50	-	-	2	20	7	29
60	-	-	-	6	3	9
n_y	12	43	79	47	19	n = 200

значения $\Delta x = 10$, $\Delta y = 10$; можно также выбрать $c_1 = 40$, $c_2 = 35$.
при этом мы получим таблицу с условными вариантами

u_i \ v_i	-2	-1	0	1	2	n_i
-3	5	-	-	-	-	5
-2	7	20	-	-	-	27
-1	-	23	30	10		63
0	-	-	47	11	9	67
1	-	-	2	20	7	29
2	-	-	-	6	3	9
n_j	12	43	79	47	19	n = 200

При этом имеют место формулы

$$\bar{x} = \Delta x \cdot \bar{u} + c_1,$$

где

$$\bar{u} = \frac{\left(\sum_{i=1}^{k} n_i u_i\right)}{n};$$

$$\bar{y} = \Delta y \cdot \bar{v} + c_2,$$

где

$$\bar{v} = \frac{\left(\sum_{j=1}^{m} n_j v_j\right)}{n};$$

$$\sigma^2_{x_в} = \Delta x^2 \cdot \sigma^2_u = (\Delta x)^2 \cdot \left(\overline{u^2} - (\bar{u})^2\right);$$

где

$$\overline{u^2} = \frac{\left(\sum_{i=1}^{k} n_i u_i^2\right)}{n};$$

$$\sigma^2_{y_в} = \Delta y^2 \cdot \sigma^2_u = (\Delta y)^2 \cdot \left(\overline{v^2} - (\bar{v})^2\right);$$

где

$$\overline{v^2} = \frac{\left(\sum_{j=1}^{m} n_j v_j^2\right)}{n}.$$

<u>Пример</u>. Получим уравнение регрессии *Y* на *X* для корреляционной таблицы, рассмотренной выше. Промежуточные вычисления представлены в следующей таблице:

v_j / u_i	-2	-1	0	1	2	n_i	$n_i u_i$	$n_i u_i^2$
-3	30 5	-	-	-	-	5	-15	45
-2	28 7	40 20	-	-	-	27	-54	108
-1	-	23 23	0 30	10 10		63	-63	63
0	-	-	0 47	0 11	0 9	67	0	0
1	-	-	0 2	20 20	14 6	29	29	29
2	-	-	-	12 0	12 3	9	18	36
n_j	12	43	79	47	19	n = 200	$\sum n_i u_i$= -85	$\sum n_i u_i^2$= = 281
$n_j v_j$	-24	-43	0	47	38	$\sum n_j v_j$=18		
$n_j v_j^2$	48	43	0	47	76	$\sum n_j v_j^2$=214		

В верхние части клеток с частотами n_{ij} записаны произведения $u_i v_j n_{ij}$, например,

$$u_1 \cdot v_1 \cdot n_{11} = (-3) \cdot (-2) \cdot 5 = 30.$$

В столбце n_i записаны суммы частот n_{ij} для каждой строки, например, для третьей строки 23 + 30 + 10 = 63.
В столбце $n_i u_i$ записаны произведения чисел из столбцов n_i и u_i, затем найдена сумма этих произведений.
В столбце $n_i u_i^2$ записаны произведения квадратов чисел из столбца u_i на числа n_i, а затем подсчитана сумма этих произведений.
В строке n_j записаны суммы частот n_{ij} для каждого столбца, например, для первого столбца 5 + 7 = 12.
В строке $n_j v_j$ записаны произведения чисел из строк v_i и n_j, затем найдена сумма этих произведений.
В строке $n_j v_j^2$ записаны произведения квадратов чисел из строк v_i на числа из строки n_j, а затем подсчитана сумма этих произведений.
Таким образом,

$$\bar{u} = \frac{\left(\sum_{i=1}^{6} n_i u_i\right)}{n} = \frac{-85}{200} = -0{,}425,$$

$$\bar{v} = \frac{\left(\sum_{j=1}^{5} n_j v_j\right)}{n} = \frac{18}{200} = 0{,}09,$$

$$\overline{u^2} = \frac{\left(\sum_{i=1}^{6} n_i u_i^2\right)}{n} = \frac{281}{200} = 1{,}405,$$

$$\overline{v^2} = \frac{\left(\sum_{j=1}^{5} n_j v_j^2\right)}{n} = \frac{214}{200} = 1{,}07.$$

По формулам $\sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2}$, $\sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2}$ посчитаем средние квадратические отклонения:

$$\sigma_u = \sqrt{1{,}405 - (-0{,}425)^2} = 1{,}107$$

$$\sigma_v = \sqrt{1{,}07 - (0{,}09)^2} = 1{,}030 .$$

Просуммировав содержимое верхних частей клеток, найдем

$$\sum_{i=1}^{6}\sum_{j=1}^{5} u_i \cdot v_j \cdot n_{ij} = 30 + 28 + 40 + 23 - 10 + 20 + 12 + 14 + 12 = 169 .$$

Вычислим выборочный коэффициент корреляции, подставив полученные данные в формулу для

$$r_b = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}u_i v_j - \bar{u}\cdot\bar{v}\cdot n}{n\sigma_u \cdot \sigma_v}$$

$$r_в = \frac{169 - (-0{,}425)\cdot 0{,}09 \cdot 200}{200 \cdot 1{,}107 \cdot 1{,}030} = 0{,}775 .$$

Получим теперь уравнение регрессии, вычислив предварительно

$$\sigma_{x_В} = \sigma_u \cdot \Delta x = 1{,}107 \cdot 10 = 11{,}07 ,$$

$$\sigma_{y_в} = \sigma_u \cdot \Delta y = 1{,}03 \cdot 10 = 10{,}3 ,$$

$$\bar{x} = \bar{u} \cdot \Delta x + c_1 = -0{,}425 \cdot 10 + 40 = 35{,}75 ,$$

$$\bar{y} = \bar{v} \cdot \Delta y + c_2 = 0{,}09 \cdot 10 + 35 = 35{,}9 .$$

Подставим численное значение в формулу для уравнения регрессии:

$$y_x - \bar{y} = r_в \cdot \frac{\sigma_{y_в}}{\sigma_{x_в}}(x - \bar{x});$$

Получим,

$$y_x - 35{,}9 = 0{,}775 \cdot \frac{10{,}3}{11{,}07}(x - 35{,}75).$$

Окончательно имеем

$$y_x = 0{,}72x + 10{,}12 .$$

<u>Ответ</u>. $y_x = 0{,}72x + 10{,}12$.