Рассмотрим зависимые X и Y.

Представим Y как функцию X в виде

$$Y \approx g(X) = \beta X + \alpha, \quad \text{где}$$

α, β – параметры, подлежащие определению.

Функция $Y = g(X)$ – ***среднеквадратическая регрессия*** Y на X.

Теорема 3 (уравнение среднеквадратической регрессии)

$$\boxed{Y - m_Y = r_{XY}\frac{\sigma_Y}{\sigma_X}(X - m_X)}$$

— уравнение регрессии Y на X;

$$\boxed{X - m_X = r_{XY}\frac{\sigma_X}{\sigma_Y}(Y - m_Y)}$$

— уравнение регрессии X на Y.

Доказательство

Рассмотрим вспомогательную функцию:

$$F(\alpha,\beta)=M\left[(Y-\alpha-\beta X)^2\right]$$

Цель: $F(\alpha,\beta)\to\min.$

$$F(\alpha,\beta)=M\left[(Y-\alpha-\beta X+m_Y-m_Y+\beta m_X-\beta m_X)^2\right]=$$

$$=M\left[\left((Y-m_Y)+m_Y-\alpha-\beta(X-m_X)-\beta m_X\right)^2\right]=$$

$$= M\left[\left(\left(Y - m_Y\right) - \beta\left(X - m_X\right) + \left(m_Y - \alpha - \beta m_X\right)\right)^2\right] =$$

$= \Big[$ возводим сумму в квадрат; при этом

$$M\left[X - m_X\right] = M\left[Y - m_Y\right] = 0;$$

$$M\left[\left(X - m_X\right)\left(Y - m_Y\right)\right] =$$

$$= M\Big[(Y-m_Y)^2 + \beta^2 (X-m_X)^2 + (m_Y-\alpha-\beta m_X)^2 +$$

$$+2(Y-m_Y)(m_Y-\alpha-\beta m_X) - 2(Y-m_Y)\cdot\beta\cdot(X-m_X) -$$

$$-2\beta\cdot(X-m_X)(m_Y-\alpha-\beta m_X)\Big] =$$

$$= M\Big[(Y-m_Y)^2\Big] + \beta^2\cdot M\Big[(X-m_X)^2\Big] + (m_Y-\alpha-\beta m_X)^2 +$$

$$+2M\big[Y-m_Y\big]\cdot(m_Y-\alpha-\beta m_X) - 2\beta\cdot M\Big[(Y-m_Y)(X-m_X)\Big] -$$

$$= D[Y] + \beta^2 D[X] + (m_Y - \alpha - \beta m_X)^2 -$$

$$-2\beta \cdot M[(Y - m_Y)(X - m_X)] = \sigma_Y^2 + \beta^2 \sigma_X^2 +$$

$$+(m_Y - \alpha - \beta m_X)^2 - 2\beta r_{XY} \sigma_X \sigma_Y .$$

Исследуем $F(\alpha,\beta)$ на экстремум:

$$\begin{cases} \dfrac{\partial F}{\partial \alpha} = -2\left(m_Y - \alpha - \beta m_X\right) = 0 \\ \dfrac{\partial F}{\partial \beta} = 2\beta\,\sigma_X^2 + 2\left(m_Y - \alpha - \beta m_X\right)\left(-m_X\right) - 2r_{XY}\sigma_X\sigma_Y = 0. \end{cases}$$

Решая системy, получаем: $\beta = r_{XY}\dfrac{\sigma_Y}{\sigma_X}$;

$$\alpha = m_Y - r_{XY}\frac{\sigma_Y}{\sigma_X}m_X.$$

Легко проверить, что при найденных α и β функция $F(\alpha,\beta)$ имеет минимум. Подставляя в уравнение $Y = \beta X + \alpha$ найденные параметры, получаем требуемое. □

Пример задачи на линейную корреляцию СВ

Известен закон распределения случайного вектора

Y/X	25	30	35
120	0,05		
125	0,15	0,3	0,05
130	0,05	0,25	0,1
135			0,05

1) Найти законы распределения составляющих. Вычислить числовые характеристики X и Y.

2) Записать условные законы распределения Y|X и X|Y. Построить на координатной плоскости эмпирические линии регрессии.

3) Вычислить коэффициент корреляции r_{XY}.

4) Записать уравнения теоретических линий регрессии

Y по X и X по Y.

Построить теоретические линии регрессии.

Решение.

1)

Y/X	25	30	35	$\Sigma \rightarrow$
120	0,05			0,05
125	0,15	0,3	0,05	0,5
130	0,05	0,25	0,1	0,4
135			0,05	0,05
$\Sigma \downarrow$	0,25	0,55	0,2	1

X	25	30	35
P	0,25	0,55	0,2

Y	120	125	130	135
P	0,05	0,5	0,4	0,05

Законы распределения X и Y.

Числовые характеристики:

$M[X] = 29{,}75$ $\qquad$ $M[Y] = 127{,}25$

$D[X] = 11{,}2$ $\qquad$ $D[Y] = 11{,}2$

$\sigma[X] = 3{,}34$ $\qquad$ $\sigma[Y] = 3{,}34$

Составим условные законы распределения.

Y|X=25	120	125	130
P	0,2=0,05/0,25	0,6=0,15/0,25	0,2=0,05/0,25

$M[Y|X=25] = 125$ – условное м.о.

$A_1(25; 125)$

Y|X=30	125	130
P	0,54	0,46

$M[Y|X=30] = 127,3$

$A_2(30; 127,3)$

Y|X=35	125	130	135
P	0,25	0,5	0,25

$A_3\,(35;\,130)$

$\parallel$

$M[Y|X=35]$

$A_1 A_2 A_3$ — эмпирическая ломаная Y по X.

Аналогично находим $B_1 B_2 B_3 B_4$ — эмпирическую ломаную X по Y:

$B_1\,(25;\,120)$; $B_2\,(29;\,125)$; $B_3\,(30{,}625;\,130)$; $B_4\,(35;\,135)$

$\parallel$ $M[X|Y=125]$ $\qquad$ $\parallel$ $M[X|Y=135]$

Вычисление коэффициента корреляции.

Переходим к системе центрированных СВ $(X - M[X], Y - M[Y])$.

$y-m_y$ \ $x-m_x$	$-4{,}75$	$0{,}25$	$5{,}25$
$-7{,}75$	0,05		
$-2{,}75$	0,15	0,3	0,05
$2{,}25$	0,05	0,25	0,1
$7{,}75$			0,05

$$r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}$$

$$K_{XY} = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - m_X) \cdot (y_j - m_Y) \cdot p_{ij}$$

$$K_{xy} = 5{,}69$$

$$r_{xy} = \frac{5{,}69}{11{,}2} \approx 0{,}51$$

Уравнения линий регрессии.

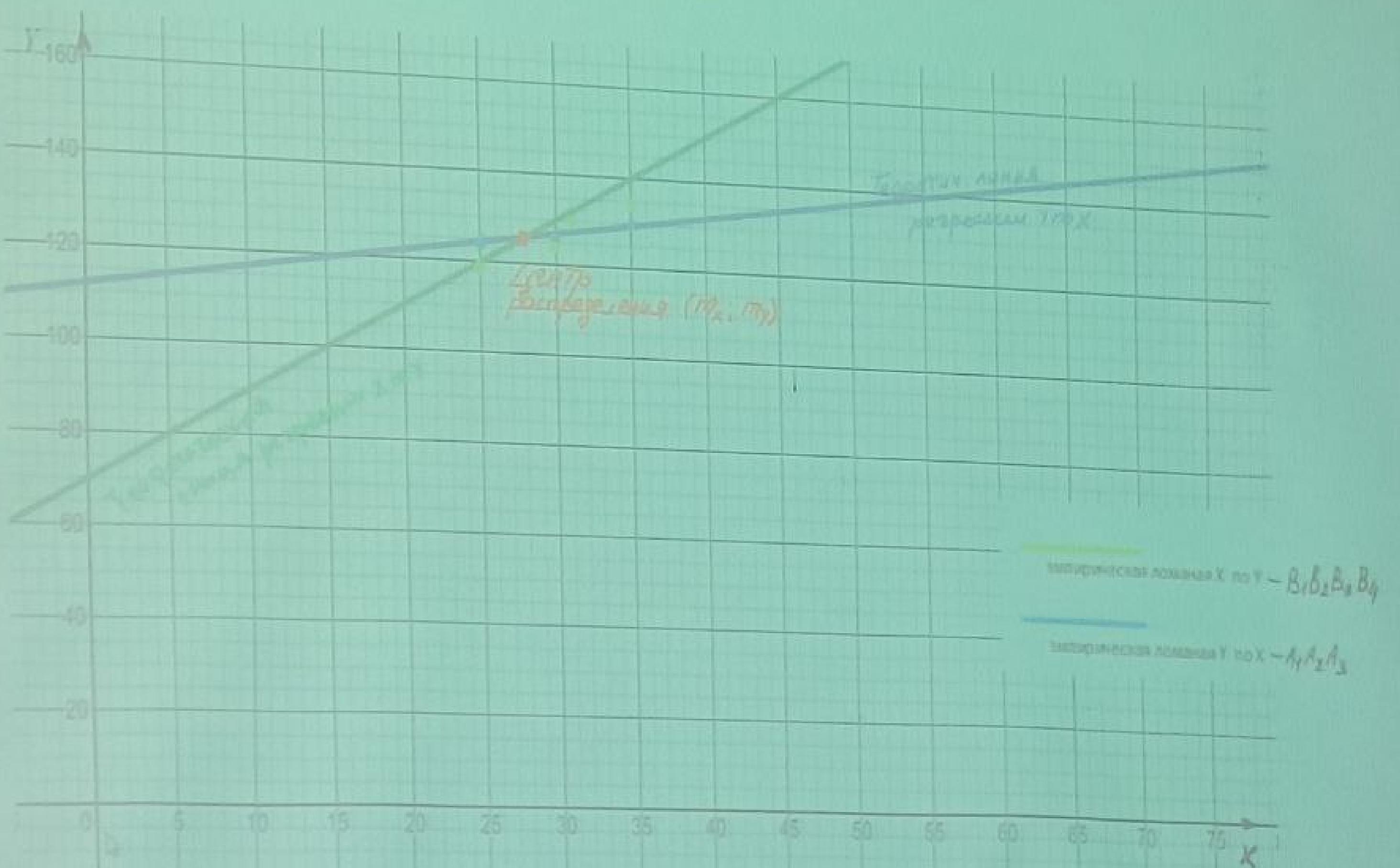
$$Y - 127{,}25 = 0{,}51 \cdot \frac{3{,}34}{3{,}34}(X - 29{,}75)$$

$Y = 0{,}51X + 112{,}1$ – теоретическая линия регрессии Y по X

$$X - 29{,}75 = 0{,}51 \cdot \frac{3{,}34}{3{,}34}(Y - 127{,}25)$$

$$X = 0{,}51Y - 35{,}14$$

Построение эмпирических ломаных и теоретических линий регрессии.



n-мерные случайные величины

Определение

***n*-мерной случайной величиной** называется совокупность n измеримых функций

$$\left(\xi_1(\omega), \xi_2(\omega), \xi_3(\omega), \dots, \xi_n(\omega)\right),$$

отображающих

$$\left(\Omega, \mathcal{A}\right) \to \left(R^n, \mathcal{B}\right) \text{,где}$$

$\mathcal{B}$ – борелевская σ–алгебра в R^n.

Необходимым и достаточным условием измеримости системы функций является условие:

$$\forall x_1, x_2, \dots, x_n \quad \{\omega : \xi_1(\omega) < x_1, \xi_2(\omega) < x_2, \dots, \xi_n(\omega) < x_n\} \in \mathfrak{A}.$$

Поэтому основной характеристикой n-мерной СВ $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ является функция распределения

$$F(x_1, x_2, \dots, x_n) = P(\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n).$$

Свойства функции распределения аналогичны свойствам функции распределения одномерной СВ.

Свойства функции распределения $F(x_1, x_2, \ldots, x_n)$.

- $F(x_1, x_2, \ldots, x_n)$ – неубывающая функция по каждому из своих аргументов

- $\lim\limits_{x_i \to -\infty} F(x_1, x_2, \ldots, x_n) = 0, \forall i \in \overline{1,n}, \quad \lim\limits_{x_1, x_2, \ldots x_n \to +\infty} F(x_1, x_2, \ldots, x_n) = 1$

В этом случае функцию распределения можно представить в виде

$$F(x_1, x_2, \ldots, x_n) = \int\limits_{-\infty}^{x_1} \int\limits_{-\infty}^{x_2} \ldots \int\limits_{-\infty}^{x_n} f(t_1, t_2, \ldots, t_n)\, dt_1 dt_2 \ldots dt_n,$$

где почти всюду выполняется равенство

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}.$$

Свойства плотности вероятности $f(x_1, x_2, \ldots, x_n)$

случайного вектора $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$.

- Неотрицательность: $f(x_1, x_2, \ldots, x_n) \geq 0.$

- Условие нормированности:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_n)\,dx_1 dx_2 \ldots dx_n = 1.$$

- Если $\mathfrak{D} \in \mathfrak{B}$, то $P(\xi \in D) = \iint\limits_{D}\ldots\int f(x_1, x_2, \ldots, x_n)\,dx_1 dx_2 \ldots dx_n.$

Независимость n случайных величин

СВ $\xi_1, \xi_2, \ldots, \xi_n$ ***независимы в совокупности***, если события $\xi_1 < x_1, \xi_2 < x_2, \ldots, \xi_n < x_n$ независимы в совокупности для любых $x_1, x_2, \ldots, x_n \in R$.

СВ $\xi_1, \xi_2, \ldots, \xi_n$ ***независимы в совокупности***, если для любого набора борелевских подмножеств $B_1, B_2, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ имеет место равенство

$$P(\xi_1 \in B_1, \xi_2 \in B_2, \ldots, \xi_n \in B_n) = P(\xi_1 \in B_1) P(\xi_2 \in B_2) \cdots P(\xi_n \in B_n).$$

$(\xi_1, \xi_2, \dots, \xi_n$ независимы в совокупности $) \Leftrightarrow$

$\Leftrightarrow \left(F(x_1, x_2, \dots, x_n) = F_1(x_1) F_2(x_2) \cdots F_n(x_n) \right)$

$(\xi_1, \xi_2, \dots, \xi_n$ непрерывны, независимы

в совокупности $) \Leftrightarrow \left(f(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n) \right)$

Плотность распределения определяется с точностью до её значений на множестве нулевой меры Лебега, а значит последнее равенство понимается как «почти всюду».