Лекция 7

7. Условные распределения и регрессия

7.1. Дискретный случай

Пусть задан случайный вектор (ξ, η) дискретного типа с законом распределения:

$$\mathsf{P}(\,\xi = x_i, \eta = y_j\,) = p(x_i, y_j) = p_{ij}, \quad \sum_{i,\,j} p_{ij} = 1. \tag{1.1}$$

Введём понятие условного распределения величины η при условии, что задано значение ξ. Запишем

$$\mathsf{P}(\xi = x_i) = \sum_{j=1}^{\infty} p_{ij} = p_{i\cdot} > 0, \quad \mathsf{P}(\eta = y_j) = \sum_{i=1}^{\infty} p_{ij} = p_{\cdot j} > 0. \tag{1.2}$$

По определению условной вероятности

$$\mathsf{P}(\eta = y_j \,|\, \xi = x_i) = \frac{\mathsf{P}(\,\xi = x_i, \eta = y_j\,)}{\mathsf{P}(\xi = x_i)} = \frac{p_{ij}}{p_{i\cdot}}. \tag{1.3}$$

Если фиксировать x_i, то вероятности (1.3) можно рассматривать как *условное распределение* случайной величины η при условии, что $\xi = x_i$.

Условным математическим ожиданием случайной величины η при условии, что $\xi = x_i$, называется число

$$\mathsf{M}(\eta \,|\, \xi = x_i) = \sum_{j=1}^{\infty} y_j \,\mathsf{P}(\eta = y_j \,|\, \xi = x_i) = \sum_{j=1}^{\infty} y_j \,\frac{p_{ij}}{p_{i\cdot}}. \tag{1.4}$$

Если задана функция $g(y)$, то

$$\mathsf{M}(g(\eta) \,|\, \xi = x_i) = \sum_{j=1}^{\infty} g(y_j) \,\mathsf{P}(\eta = y_j \,|\, \xi = x_i) = \sum_{j=1}^{\infty} g(y_j) \,\frac{p_{ij}}{p_{i\cdot}}. \tag{1.5}$$

Аналогично определяется условное распределение случайной величины ξ при условии, что $\eta = y_j$. Имеем

$$\mathsf{P}(\xi = x_i \,|\, \eta = y_j) = \frac{\mathsf{P}(\eta = y_j,\, \xi = x_i)}{\mathsf{P}(\eta = y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \tag{1.6}$$

$$\mathsf{M}(\xi \,|\, \eta = y_j) = \sum_{i=1}^{\infty} x_i \,\mathsf{P}(\xi = x_i \,|\, \eta = y_j) = \sum_{i=1}^{\infty} x_i \,\frac{p_{ij}}{p_{\cdot j}}. \tag{1.7}$$

Если задана функция $f(x)$, то

$$\mathsf{M}(f(\xi) \,|\, \eta = y_j) = \sum_{i=1}^{\infty} f(x_i) \,\mathsf{P}(\xi = x_i \,|\, \eta = y_j) = \sum_{i=1}^{\infty} f(x_i) \,\frac{p_{ij}}{p_{\cdot j}}. \tag{1.8}$$

Часто вместо $M(\eta \,|\, \xi = x_i)$, $\mathsf{M}(\xi \,|\, \eta = y_j)$ пишут просто $M(\eta \,|\, x_i)$, $\mathsf{M}(\xi \,|\, y_j)$.

Пример 1. Пусть дан случайный вектор (ξ, η) дискретного типа с законом распределения

$\xi \setminus \eta$	0	2	5
1	0,1	0	0,2
2	0	0,3	0
4	0,1	0,3	0

Найдём одномерные законы распределения по формулам (1.2):

$$p_{1\cdot} = \mathsf{P}\{\,\xi = 1\,\} = 0{,}1 + 0 + 0{,}2 = 0{,}3,$$
$$p_{2\cdot} = \mathsf{P}\{\,\xi = 2\,\} = 0 + 0{,}3 + 0 = 0{,}3,$$
$$p_{3\cdot} = \mathsf{P}\{\,\xi = 4\,\} = 0{,}1 + 0{,}3 + 0 = 0{,}4.$$

Далее

$$p_{\cdot 1} = \mathsf{P}\{\,\eta = 0\,\} = 0{,}1 + 0 + 0{,}1 = 0{,}2,$$
$$p_{\cdot 2} = \mathsf{P}\{\,\eta = 2\,\} = 0 + 0{,}3 + 0{,}3 = 0{,}6,$$
$$p_{\cdot 3} = \mathsf{P}\{\,\eta = 5\,\} = 0{,}2 + 0 + 0 = 0{,}2.$$

Найдём условные законы распределения по формулам (1.3):

$$\mathsf{P}(\eta = y_1 \,|\, \xi = x_1) = 1/3, \quad \mathsf{P}(\eta = y_2 \,|\, \xi = x_1) = 0, \quad \mathsf{P}(\eta = y_3 \,|\, \xi = x_1) = 2/3,$$

$$\mathsf{P}(\eta = y_1 \,|\, \xi = x_2) = 0, \quad \mathsf{P}(\eta = y_2 \,|\, \xi = x_2) = 1, \quad \mathsf{P}(\eta = y_3 \,|\, \xi = x_2) = 0,$$

$$\mathsf{P}(\eta = y_1 \,|\, \xi = x_3) = 1/4, \quad \mathsf{P}(\eta = y_2 \,|\, \xi = x_3) = 3/4, \quad \mathsf{P}(\eta = y_3 \,|\, \xi = x_3) = 0.$$

Далее по формулам (1.6):

$$\mathsf{P}(\xi = x_1 \,|\, \eta = y_1) = 1/2, \quad \mathsf{P}(\xi = x_2 \,|\, \eta = y_1) = 0, \quad \mathsf{P}(\xi = x_3 \,|\, \eta = y_1) = 1/2,$$

$$\mathsf{P}(\xi = x_1 \,|\, \eta = y_2) = 0, \quad \mathsf{P}(\xi = x_2 \,|\, \eta = y_2) = 1/2, \quad \mathsf{P}(\xi = x_3 \,|\, \eta = y_2) = 1/2,$$

$$\mathsf{P}(\xi = x_1 \,|\, \eta = y_3) = 1, \quad \mathsf{P}(\xi = x_2 \,|\, \eta = y_3) = 0, \quad \mathsf{P}(\xi = x_3 \,|\, \eta = y_3) = 0.$$

Найдём условные математические ожидания по формулам (1.4):

$$\mathsf{M}(\eta \,|\, \xi = x_1) = 0 \cdot 1/3 + 2 \cdot 0 + 5 \cdot 2/3 = 10/3,$$
$$\mathsf{M}(\eta \,|\, \xi = x_2) = 0 \cdot 0 + 2 \cdot 1 + 5 \cdot 0 = 2,$$
$$\mathsf{M}(\eta \,|\, \xi = x_3) = 0 \cdot 1/4 + 2 \cdot 3/4 + 5 \cdot 0 = 3/2.$$

Далее по формулам (1.7):

$$\mathsf{M}(\xi \,|\, \eta = y_1) = 1 \cdot 1/2 + 2 \cdot 0 + 4 \cdot 1/2 = 5/2,$$
$$\mathsf{M}(\xi \,|\, \eta = y_2) = 1 \cdot 0 + 2 \cdot 1/2 + 4 \cdot 1/2 = 3,$$
$$\mathsf{M}(\xi \,|\, \eta = y_3) = 1 \cdot 1 + 2 \cdot 0 + 4 \cdot 0 = 1.$$

7.2. Непрерывный случай

Пусть задан случайный вектор (ξ, η) непрерывного типа с плотностью совместного распределения $p(x, y)$. Запишем одномерные законы распределения:

$$p_\xi(x) = \int_{-\infty}^{\infty} p(x, y)\, dy, \quad p_\eta(y) = \int_{-\infty}^{\infty} p(x, y)\, dx. \tag{2.1}$$

Плотностью условного распределения вероятностей случайной величины η при условии $\xi = x$ называется величина

$$p_{\eta\,|\,\xi}(y\,|\,\xi = x) = p_{\eta\,|\,\xi}(y\,|\,x) = \frac{p_(x, y)}{p_\xi(x)}, \quad p_\xi(x) > 0. \tag{2.2}$$

По определению считается, что $p_{\eta\,|\,\xi}(y\,|\,x) = 0$, если $p_\xi(x) = 0$.

Аналогично, *плотностью условного распределения вероятностей* случайной величины ξ при условии $\eta = y$ называется величина

$$p_{\xi\,|\,\eta}(x\,|\,\eta = y) = p_{\xi\,|\,\eta}(x\,|\,y) = \frac{p(x, y)}{p_\eta(y)}, \quad p_\eta(y) > 0. \tag{2.3}$$

Условным математическим ожиданием случайной величины η при условии $\xi = x$ называется величина

$$\mathsf{M}(\eta\,|\,\xi = x) = \int_{-\infty}^{\infty} y\, p_{\eta\,|\,\xi}(y\,|\,x)\, dy = \int_{-\infty}^{\infty} y\, \frac{p(x, y)}{p_\xi(x)}\, dy. \tag{2.4}$$

Аналогично,

$$\mathsf{M}(\xi\,|\,\eta = y) = \int_{-\infty}^{\infty} x\, p_{\xi\,|\,\eta}(x\,|\,y)\, dx = \int_{-\infty}^{\infty} x\, \frac{p(x, y)}{p_\eta(y)}\, dx. \tag{2.5}$$

Вместо $\mathsf{M}(\eta\,|\,\xi = x)$ и $\mathsf{M}(\xi\,|\,\eta = y)$ обычно пишут просто $\mathsf{M}(\eta\,|\,x)$ и $\mathsf{M}(\xi\,|\,y)$.

Пример 1. Пусть случайный вектор (ξ, η) равномерно распределён в треугольнике с вершинами $(0, 0)$, $(0, 2)$, $(5, 0)$. Из геометрических соображений получаем

$$p_{\eta\,|\,\xi}(y\,|\,x) = \frac{5}{2(5-x)} = \frac{1}{\beta}, \quad 0 < y < \frac{2(5-x)}{5} = \beta,\ 0 < x < 5.$$

$$p_{\xi\,|\,\eta}(x\,|\,y) = \frac{2}{5(2-y)} = \frac{1}{\alpha}, \quad 0 < x < \frac{5(2-y)}{2} = \alpha,\ 0 < y < 2,$$

Следовательно,

$$\mathsf{M}(\eta\,|\,x) = \int_0^\beta x\, \frac{dx}{\beta} = \frac{x^2}{2\beta}\bigg|_0^\beta = \frac{\beta}{2} = \frac{5-x}{5}.$$

$$\mathsf{M}(\xi\,|\,y) = \int_0^\alpha x\, \frac{dx}{\alpha} = \frac{x^2}{2\alpha}\bigg|_0^\alpha = \frac{\alpha}{2} = \frac{5(2-y)}{4},$$

7.3. Регрессия

Пусть случайные величины ξ и η имеют совместное распределение абсолютно непрерывного типа. Величина $\mathsf{M}(\eta\,|\,\xi = x)$ называется *регрессией случайной величины* η *на случайную величину* ξ. Аналогично $\mathsf{M}(\xi\,|\,\eta = y)$ называется *регрессией величины* ξ *на величину* η.

Уравнение

$$y = \mathsf{M}(\eta \,|\, \xi = x) \tag{3.1}$$

называется *уравнением регрессии случайной величины* η *на случайную величину* ξ. Аналогично уравнение

$$x = \mathsf{M}(\xi \,|\, \eta = y) \tag{3.2}$$

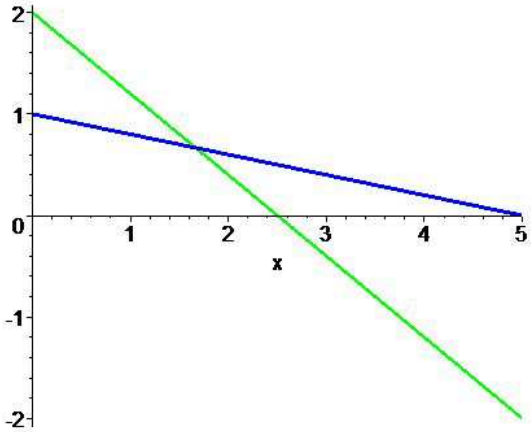
называется *уравнением регрессии величины* ξ *на величину* η. Графики уравнений (3.1) и (3.2) называются *линиями регрессии*.

Пример 1. Пусть случайный вектор (ξ, η) равномерно распределён в треугольнике с вершинами $(0,0)$, $(0,2)$, $(5,0)$. В примере 2.1 было найдено, что

$$\mathsf{M}(\eta \,|\, \xi = x) = \frac{5-x}{5}, \quad \mathsf{M}(\xi \,|\, \eta = y) = \frac{5(2-y)}{4}.$$

Следовательно, линии регрессии имеют вид:

$$y = \frac{5-x}{5}, \quad x = \frac{5(2-y)}{4}.$$



Теорема 3.1. (Основное свойство регрессии) *Если* $f(x)$ *есть функция регрессии случайной величины* η *на случайную величину* ξ*, то для любой функции* $h(x)$ *справедливо неравенство*:

$$\mathsf{M}\,[\,\eta - f(\xi)\,]^2 \le \mathsf{M}\,[\,\eta - h(\xi)\,]^2\,. \tag{3.3}$$

Доказательство. Сначала докажем, что для любой функции $u(\xi)$

$$\mathsf{M}\,[u(\xi)\eta] = \mathsf{M}\,[u(\xi) f(\xi)]\,. \tag{3.4}$$

По формуле (6.3.5) для математического ожидания функции от случайных величин имеем

$$\mathsf{M}\,[u(\xi)\eta] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x) y p(x,y)\,dx\,dy.$$

Заменяя здесь по формуле (2.2)

$$p(x,y) = p_\xi(x) p_{\eta\,|\,\xi}(y\,|\,x),$$

получаем

$$\mathsf{M}\,[u(\xi)\eta] = \int_{-\infty}^{\infty} u(x) p_\xi(x)\,dx \int_{-\infty}^{\infty} y p_{\eta\,|\,\xi}(y\,|\,x)\,dy.$$

Так как второй интеграл равен $\mathsf{M}(\eta \,|\, \xi = x) = f(x)$, то

$$\mathsf{M}\,[u(\xi)\eta] = \int_{-\infty}^{\infty} u(x) f(x) p_\xi(x)\, dx = \mathsf{M}\,[u(\xi) f(\xi)]\,.$$

Равенство (3.4) доказано. Отметим, в частности, что

$$\mathsf{M}\eta = \mathsf{M} f(\xi),\ \ \mathsf{M}(\xi\eta) = \mathsf{M}[\xi f(\xi)]. \tag{3.5}$$

Теперь преобразуем правую часть неравенства (3.3). Запишем

$$\mathsf{M}\,[\,\eta - h(\xi)\,]^2 = \mathsf{M}\,[\,(\eta - f(\xi)) + (f(\xi) - h(\xi))\,]^2 =$$

$$= \mathsf{M}\,[\,\eta - f(\xi)\,]^2 + \mathsf{M}\,[\,f(\xi) - h(\xi)\,]^2 + 2\mathsf{M}\,[\,(\eta - f(\xi))(f(\xi) - h(\xi))\,]\,.$$

Обозначим $u(\xi) = f(\xi) - h(\xi)$ и применим формулу (3.4). Получим

$$\mathsf{M}\,[\,(\eta - f(\xi))(f(\xi) - h(\xi))\,] = \mathsf{M}\,[\,\eta u(\xi)\,] - \mathsf{M}\,[\,f(\xi)u(\xi)\,] = 0.$$

Следовательно,

$$\mathsf{M}\,[\,\eta - h(\xi)\,]^2 = \mathsf{M}\,[\,\eta - f(\xi)\,]^2 + \mathsf{M}\,[\,f(\xi) - h(\xi)\,]^2\,. \tag{3.6}$$

Полученное соотношение доказывает теорему. ■

Аналогично формулируется и основное свойство регрессии величины ξ на величину η:

$$\mathsf{M}\,[\,\xi - g(\eta)\,]^2 \le \mathsf{M}\,[\,\xi - h(\eta)\,]^2\,, \tag{3.7}$$

где $g(y) = \mathsf{M}(\xi \,|\, \eta = y)$, а $h(y)$ — произвольная функция.

Пусть задан случайный вектор (ξ, η) дискретного типа с законом распределения (1.1). В этом случае регрессией случайной величины η на случайную величину ξ называется последовательность

$$(x_1, \mathsf{M}(\eta \,|\, \xi = x_1)), (x_2, \mathsf{M}(\eta \,|\, \xi = x_2)), \ldots \tag{3.8}$$

Регрессией величины ξ на величину η называется последовательность

$$(\mathsf{M}(\xi \,|\, \eta = y_1), y_1), (\mathsf{M}(\xi \,|\, \eta = y_2), y_2), \ldots \tag{3.9}$$

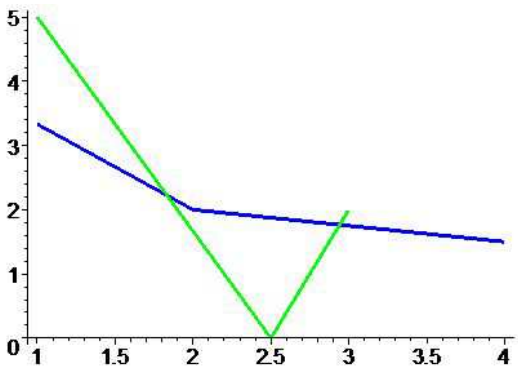
В обоих случаях можно соединить соответствующие точки прямолинейными отрезками и рассматривать таким образом построенные ломаные как кривые регрессии дискретного распределения.

Пример 2. Пусть дан случайный вектор (ξ, η) дискретного типа с законом распределения из примера 2.1. Тогда регрессией величины η на величину ξ является последовательность

$$(1, 10/3), (2, 2), (4, 3/2).$$

Регрессией величины ξ на величину η является последовательность

$$(5/2, 0), (3, 2), (1, 5).$$



7.4. Линейная регрессия

Рассмотрим случай, когда обе функции регрессии

$$f(x) = \mathsf{M}(\eta \,|\, \xi = x),\ g(y) = \mathsf{M}(\xi \,|\, \eta = y)$$

линейны. В этом случае обе линии регрессии являются прямыми линиями. Они называются *прямыми регрессии*.

Выведем уравнения прямых регрессии. Введём обозначения:

$$m_x = \mathsf{M}\xi,\ m_y = \mathsf{M}\eta,\ \sigma_x^2 = \mathsf{D}\xi,\ \sigma_y^2 = \mathsf{D}\eta,\ \mu_{xy} = \operatorname{cov}(\xi, \eta).$$

Так как $f(x)$ линейна, то можно записать

$$f(x) = A(x - m_x) + B,$$

где A, B — константы. Подставляя функцию

$$f(\xi) = A(\xi - m_x) + B$$

в формулы (3.5), получим

$$m_y = \mathsf{M}\eta = \mathsf{M}f(\xi) = B,$$

$$\mu_{xy} = \mathsf{M}[(\xi - m_x)(\eta - m_y)] = \mathsf{M}[(\xi - m_x)(f(\xi) - m_y)] =$$

$$= A\mathsf{M}(\xi - m_x)^2 = A\sigma_x^2.$$

В результате функция регрессии η на ξ имеет вид:

$$f(x) = \frac{\mu_{xy}}{\sigma_x^2}\,(x - m_x) + m_y. \tag{4.1}$$

Аналогично, функция регрессии ξ на η имеет вид:

$$g(y) = \frac{\mu_{xy}}{\sigma_y^2}\,(y - m_y) + m_x. \tag{4.2}$$

Величины

$$\frac{\mu_{xy}}{\sigma_x^2}, \quad \frac{\mu_{xy}}{\sigma_y^2} \tag{4.3}$$

называются *коэффициентами регрессии*.

Уравнения прямых регрессии можно записать в более симметричном виде, если воспользоваться безразмерным коэффициентом корреляции $\rho = \rho(\xi, \eta)$. Уравнения прямых регрессии примут вид

$$(y - m_y) = \rho\,\frac{\sigma_y}{\sigma_x}\,(x - m_x) \tag{4.4}$$

(прямая регрессии η на ξ),

$$(x - m_x) = \rho\,\frac{\sigma_x}{\sigma_y}\,(y - m_y) \tag{4.5}$$

(прямая регрессии ξ на η) или

$$\frac{y - m_y}{\sigma_y} = \rho\,\frac{x - m_x}{\sigma_x}, \quad \frac{x - m_x}{\sigma_x} = \rho\,\frac{y - m_y}{\sigma_y}. \tag{4.6}$$

Пример 1. Пусть случайный вектор (ξ, η) равномерно распределён в треугольнике с вершинами $(0,0)$, $(0,2)$, $(5,0)$. Из примера 3.1 известно, что функции регрессии являются линейными функциями. Найдём их по формулам (4.1) и (4.2).

Запишем одномерные плотности распределений случайных величин (ξ, η):

$$p_\xi(x) = \frac{1}{5}\int_0^{2(5-x)/5} dy = \frac{2(5-x)}{25}, \quad 0 < x < 5.$$

$$p_\eta(y) = \frac{1}{5}\int_0^{5(2-y)/2} dx = \frac{2-y}{2}, \quad 0 < y < 2.$$

Ищем математические ожидания $\mathsf{M}\xi$, $\mathsf{M}\xi^2$, $\mathsf{M}\eta$, $\mathsf{M}\eta^2$ и $\mathsf{M}(\xi\eta)$:

$$\mathsf{M}\xi = \int_0^5 x\,\frac{2(5-x)}{25}\,dx = \frac{5}{3}, \quad \mathsf{M}\eta = \int_0^2 y\,\frac{2-y}{2}\,dy = \frac{2}{3},$$

$$\mathsf{M}\xi^2 = \int_0^5 x^2\,\frac{2(5-x)}{25}\,dx = \frac{25}{6}, \quad \mathsf{M}\eta^2 = \int_0^2 y^2\,\frac{2-y}{2}\,dy = \frac{2}{3}.$$

$$\mathsf{M}(\xi\eta) = \frac{1}{5}\int_0^2 \left(\int_0^{5(2-y)/2} x\,dx\right) y\,dy = \frac{5}{6}.$$

Следовательно,

$$\mathsf{D}\xi = \frac{25}{6} - \left(\frac{5}{3}\right)^2 = \frac{25}{18}, \quad \mathsf{D}\eta = \frac{2}{3} - \left(\frac{2}{3}\right)^2 = \frac{2}{9}, \quad \operatorname{cov}(\xi,\eta) = \frac{5}{6} - \frac{5}{3}\cdot\frac{2}{3} = -\frac{5}{18}.$$

В результате

$$m_x = \frac{5}{3}, \quad m_y = \frac{2}{3}, \quad \sigma_x^2 = \frac{25}{18}, \quad \sigma_y^2 = \frac{2}{9}, \quad \mu_{xy} = -\frac{5}{18}.$$

Подставляя эти значения в формулы (4.1) и (4.2), получим функции регрессии

$$f(x) = \frac{5-x}{5}, \quad g(y) = \frac{5(2-y)}{4}.$$