Лекция – 5 (2ч)

Тема: Регрессионный анализ

План:

1. Корреляционная зависимость между двумя признаками.
2. Определение уравнениями прямолинейной регрессии.

Цель: представление корреляционной зависимости между признаками в виде формулы, позволяющей прогнозировать значения одного показателя по конкретному значению другого.

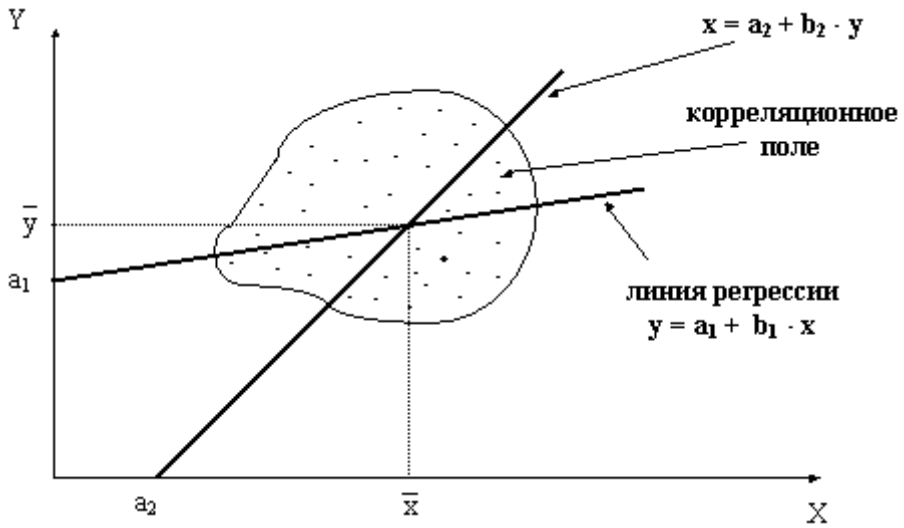
Теоретические сведения

1. Корреляционную зависимость между двумя признаками.

В практических исследованиях возникает необходимость аппроксимировать (математически описать приблизительно) корреляционную зависимость между двумя признаками уравнением. Для линейной зависимости вытянутое корреляционное поле заменить усредненной прямой линией и найти ее уравнение по статистическим данным коррелируемых признаков. В прямоугольной системе координат уравнение прямой линии записывается в виде:

$$y = a + b \cdot x$$

Это математическое выражение корреляционной зависимости называется уравнением регрессии. Коэффициенты a и b называются параметрами уравнения регрессии. Параметр a определяет на графике отрезок, отсекаемый прямой линией на оси Y. Параметр b показывает, как изменяется признак Y при изменении признака X. Это "b " еще называют коэффициентом регрессии.



Уравнение регрессии тем лучше описывает корреляционную зависимость, чем ближе она к линейной и чем больше ее достоверность. В

случае нелинейной зависимости математически запись может выражаться в виде более сложных уравнений различных кривых линий (экспоненциальной кривой, параболы, гиперболы и т.д.).

При наличии достоверной криволинейной корреляционной зависимости можно подобрать уравнение, хорошо ее описывающее. Особенно эта возможность становится реальной при наличии электронно-вычислительной техники.

2. Определение уравнения прямолинейной регрессии

Как уже было сказано, в случае линейной зависимости уравнение регрессии является уравнением прямой линии. Таких уравнений два:

$$y = a_1 + b_{yx} \cdot x, \quad (1)$$

$$x = a_2 + b_{x/y} \cdot y. \quad (2)$$

Если уравнение (1) называть прямым, то уравнение (2) будет ему обратным, и наоборот.

Параметры $a_1, a_2, b_{x/y}$ определяются на основании статистических данных признаков X и Y по формулам:

$$b_1 = b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad (3)$$

$$b_2 = b_{x/y} = r \cdot \frac{\sigma_y}{\sigma_x} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (y_i - \bar{y})^2}. \quad (4)$$

Коэффициенты регрессии имеют размерность, равную отношению размерностей изучаемых признаков X и Y, и тот же знак, что и коэффициенты корреляции.

$$a_2 = \bar{y} - b_{yx} \cdot \bar{x}, \quad (5)$$

$$a_2 = \bar{x} - b_{yx} \cdot \bar{y}. \quad (6)$$

Чтобы вычислить a_1 и a_2, надо просто в уравнения (1) и (2) подставить средние значения коррелируемых признаков. Для оценки качества уравнения регрессии вычисляются остаточные средние квадратические отклонения по формулам:

$$\sigma_{y/x} = \sigma_y \cdot \sqrt{1 - r^2}, \quad (7)$$

$$\sigma_{x/y} = \sigma_x \cdot \sqrt{1 - r^2}. \quad (8)$$

Эти оценки абсолютны и, следовательно, не могут быть сравнимы друг с другом. Поэтому вводят оценки относительной погрешности уравнений регрессии, которые определяются в процентах по формулам:

$$\sigma'_{y/x} = \frac{\sigma_{y/x}}{\bar{y}} \cdot 100\%, \quad (9)$$

$$\sigma'_{x/y} = \frac{\sigma_{x/y}}{\bar{x}} \cdot 100\%, \quad (10)$$

Значение этой оценки, если r = ± 1,00, и, если r = 0,00, максимально. Остаточное среднее квадратическое отклонение характеризует колеблемость *y* относительно линии регрессии по *x,* и наоборот в обратном случае.

Пример

Найти уравнения регрессии для веса (X) и роста (Y) группы студентов, если их значения таковы:

x_i, кг ~ 60, 65, 71, 73, 75, 80, 72.
y_i, см ~ 170, 168, 180, 182, 189, 190, 178.

Решение:

1. Занесем результаты тестирования в рабочую таблицу:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
60	-11	121	170	-10	100	110
65	- 6	36	168	-12	144	72
71	0	0	180	0	0	0
73	2	4	182	2	4	4
75	4	16	189	9	81	36
80	9	81	190	10	100	90
72	1	1	178	-2	4	2
$\bar{x} \approx 71$		$\sum = 259$	$\bar{y} \approx 180$		$\sum = 433$	$\sum = 314$

2. Рассчитаем нормированный коэффициент корреляции по формуле:

$$r^p_{x,y} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}};$$

$$r^p_{x,y} = \frac{314}{\sqrt{259 \cdot 433}} = \frac{314}{\sqrt{112147}} \approx \frac{134}{334,8} \approx 0,93.$$

3. Подставим полученные данные в уравнения регрессии:

$$y = \bar{y} + \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \cdot (x - \bar{x})$$

$$x = \bar{x} + \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \cdot (y - \bar{y})$$

Тогда уравнение регрессии примет вид:

$$y = 180 + \frac{314}{259} \cdot (x - 71) = 180 + 1{,}21 \cdot (x - 71) \approx 180 + 1{,}21x - 85{,}9 \approx 1{,}21x + 94{,}1$$

$$x = 71 + \frac{314}{433} \cdot (y - 180) = 71 + 0{,}72 \cdot (y - 180) \approx 71 + 0{,}72y - 129{,}6 \approx 0{,}72y - 58{,}6$$

Т.е. $\mathbf{y = 1{,}21 \cdot x + 94{,}1}$ (1)

$\mathbf{x = 0{,}72 \cdot y - 58{,}6}$ (2)

4. В конечные значения уравнений (1) и (2) подставим произвольные значения показателей *x* и *y* (например, 1-го исследуемого).

Тогда:

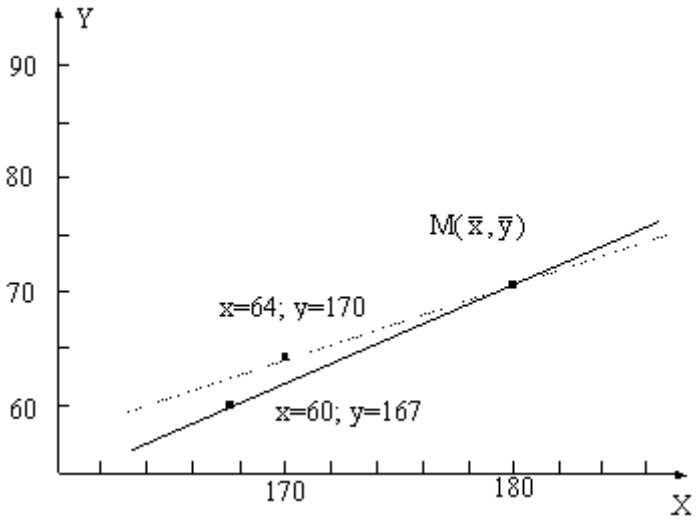
1) при x = 60 y= 1,21· 60 + 94,1 = 166,7 ≈ 167 ;
2) при y=170 x= 0,72·170 - 58,6 = 63,8 ≈ 64 .

5. Разобранную в данном примере корреляционную зависимость можно представить графически в виде, приведенном на рисунке 13, учитывая следующие особенности данного представления:

1. две линии уравнения регрессии на графике пересекаются в точке М с координатами средних значений показателей *x* и *y;*

2. чем ближе коэффициент корреляции по своему значению к |1|, тем меньше угол между линиями на графике. При r = ± 1 линии уравнения регрессии либо совпадают, либо расположены параллельно, так как корреляционная взаимосвязь между признаками в этом случае переходит в функциональную;

3. чем ближе значение коэффициента корреляции к нулю, тем больше угол между линиями на графике. При r = 0 линии уравнения регрессии на графике расположены перпендикулярно, т.е. взаимосвязь между



показателями отсутствует.