Задание

1. Повторить теоретический материал темы, законспектировать пример решения задачи, подготовить 5-7 контрольных вопросов.

2. Фотоотчет и сообщение присылать на электронную почту

С уважением, Хвастова Светлана Ивановна

!!! Если возникнут вопросы обращаться по телефону 0721389311 (ватсап). Электронная почта: xvsviv@rambler.ru

Лекция на тему: «Отыскание параметров выборочного уравнения прямой линии среднеквадратической регрессии по несгруппированным данным.»

Пусть изучается система количественных признаков (X, Y). *Допустим в результате* n *независимых опытов получены* n *пар чисел* $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, *которые могут быть представлены в виде таблицы:*

Таблица 1

x	x_1	x_2	$\ldots$	x_n
y	y_1	y_2	$\ldots$	y_n

Рассмотрим случай, когда есть основания предполагать наличие линейной корреляционной зависимости между величинами X *и* Y *(в генеральной совокупности их значений) т.е. когда линейные уравнения регрессии имеют вид:*

$$M(Y / X) = Ax + B \qquad (1)$$

(1) –уравнение линейной регрессии Y *на* X.

$$M(X / Y) = Cy + D \qquad (2)$$

(2) –уравнение линейной регрессии X *на* Y.

$M(Y / X)$ *–условное матожидание* Y *при конкретном* X;

$M(X / Y)$ *– условное матожидание* X *при конкретном* Y.

В этих случаях для описания корреляционных зависимостей между величинами X *и* Y *по результатам выборочных наблюдений вводят выборочные уравнения линейной регрессии* Y *на* X *и* X *на* Y:

$$\bar{y}_x = \rho_{yx} x + b \qquad (3)$$

(3) –уравнение линейной регрессии Y *на* X.

$$\bar{x}_y = \rho_{xy} y + d \qquad (4)$$

(4) –уравнение линейной регрессии X *на* Y,

где ρ_{yx}, ρ_{xy} *- выборочные коэффициенты линейной регрессии, имеющие смысл выборочных оценок коэффициентов* A *и* C *в формулах* (1), (2). *При этом* $\overline{y}_x$ *и* $\overline{x}_y$ *являются оценками условных математических ожиданий* $M(Y/X)$ *и* $M(X/Y)$, *а параметры* b *и* d *- оценками* B *и* D.

Пусть в результате n *независимых опытов получены* n *пар значений системы* (X,Y): $(x_1,y_1),(x_2,y_2),\ldots(x_n,y_n)$, *которые могут быть заданы (табл. 1). По этим статистическим данным найдем сначала параметры (коэффициенты) уравнения* (3) *регрессии* Y *на* X:

$$\overline{y}_x = \rho_{yx}x + b \tag{5}$$

Так как различные значения X *и соответствующие им значения* Y *наблюдались по одному разу, то группировать данные нет необходимости; также нет надобности использовать понятие условной средней (*$M(Y/X)$ *−условное матожидание* Y *при конкретном* X; $M(X/Y)$ *− условное матожидание* X *при конкретном* Y*), поэтому уравнение (3) можно записать:*

$$Y = \rho_{yx}x + b. \tag{6}$$

Подберем параметры ρ_{yx} *и* b *так, чтобы точки* $(x_1,y_1),(x_2,y_2),\ldots(x_n,y_n)$, *построенные по данным наблюдениям на плоскости* XOY *лежали как можно ближе к прямой* (5).

Разность $Y_i - y_i$ *– является отклонением ординаты* y_i, *вычисленной с помощью уравнения (5) при* $x = x_i$ *от наблюдаемой ординаты* y_i *(табличное значение), соответствующей значению* x_i. *Используем в дальнейшем метод наименьших квадратов, а именно подберем* ρ_{yx} *и* b *так, чтобы сумма квадратов отклонений была минимальной, т.е. составим функцию* $F(\rho,b)$, $(\rho_{yx} = \rho)$:

$$F(\rho,b) = \sum_{i=1}^{n}(Y_i - y_i)^2 = \sum_{i=1}^{n}(\rho x_i + b - y_i)^2.$$

Исследуем функцию $F(\rho,b)$ *на минимум, приравниваем к нулю ее частные производные 1-го порядка:*

$$\frac{\partial F}{\partial \rho} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0; \qquad \frac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0. \tag{7}$$

Получим систему двух линейных уравнений относительно ρ *и* b.

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^2\right)\rho + \left(\sum_{i=1}^{n} x_i\right)b = \sum_{i=1}^{n}(x_i y_i) \\ \left(\sum_{i=1}^{n} x_i\right)\rho + nb = \sum_{i=1}^{n} y_i. \end{cases} \tag{8}$$

Решив эту систему, получим:

$$\rho_{yx} = \rho = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}.$$

Применяя безындексную форму:

$$\rho_{yx} = \rho = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2},$$

$$b = \frac{\begin{vmatrix} \sum x^2 & \sum xy \\ \sum x & \sum y \end{vmatrix}}{\begin{vmatrix} \sum x^2 & \sum x \\ \sum x & n \end{vmatrix}} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}.$$

Аналогично можно найти выборочные уравнения прямой линии регрессии X *на* $Y: X = \rho_{xy} y + d$, *формулы для параметров* ρ_{xy} *и* d *имеют вид:*

$$\rho_{xy} = \rho = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2},$$

$$d = \frac{\sum y^2 \sum x - \sum y \sum xy}{n\sum y^2 - \left(\sum y\right)^2}.$$

Однако иногда уравнение регрессии (5) удобно записать в другой форме, вводя <u>*выборочный коэффициент корреляции*</u>*. Найдя* b *из уравнения второго системы (8) и подставляя в уравнение (5), получим:*

$$\left(\sum_{i=1}^{n} x_i\right)\rho + nb = \sum_{i=1}^{n} y_i : n$$

$$\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\rho_{yx} + b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right),$$

$$\bar{x}\rho_{yx} + b = \bar{y}, \quad b = \bar{y} - \bar{x}\rho_{yx},$$

$$Y = \rho_{yx} x + b = \rho_{yx} x + \bar{y} - \bar{x}\rho_{yx} = \rho_{yx}\left(x - \bar{x}\right) + \bar{y},$$

$$Y - \bar{y} = \rho_{yx}\left(x - \bar{x}\right). \qquad (9)$$

Если ввести соотношение $\rho_{yx} = \tilde{r}_{yx} \dfrac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$, *где*

$$\tilde{r}_{yx} = \frac{\frac{1}{n}\sum xy - \overline{x}\cdot\overline{y}}{\tilde{\sigma}_x\tilde{\sigma}_y} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\tilde{\sigma}_x\tilde{\sigma}_y}, \qquad (10)$$

(10) – *выборочный коэффициент корреляции.*

Действительно, $\rho_{yx} = \rho = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$,

разделим числитель и знаменатель на n^2:

$$\rho_{yx} = \frac{\frac{1}{n}\sum xy - \frac{\sum x}{n}\cdot\frac{\sum y}{n}}{\frac{1}{n}\sum x^2 - \left(\frac{\sum x}{n}\right)^2} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\tilde{\sigma}_x^2}, \text{ но}$$

$\overline{x}^2 - \left(\overline{x}\right)^2 = M\left[x^2\right] - m_x^2 = \tilde{D}_x = \tilde{\sigma}_x^2$, *отсюда*

$$\tilde{r}_{yx} = \frac{\rho_{yx}\cdot\tilde{\sigma}_x}{\tilde{\sigma}_y} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\tilde{\sigma}_x^2}\cdot\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\tilde{\sigma}_x\tilde{\sigma}_y}. \text{ Тогда}$$

$$Y - \overline{y} = \tilde{r}_{yx}\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}\left(x - \overline{x}\right), \qquad (11)$$

(11) – <u>*выборочное уравнение регрессии*</u> Y *на* X. *Аналогично находится выборочное уравнение линейной регрессии* X *на* Y.

$$X - \overline{x} = \tilde{r}_{xy}\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}\left(y - \overline{y}\right). \qquad (12)$$

$\tilde{\sigma}_x, \tilde{\sigma}_y$ –выборочные средние квадратические отклонения

Пример 1. *Методами корреляционного анализа исследовать зависимость между урожайностью пшеницы и картофеля на соседних участках на основании статистических данных. Построить выборочное уравнение линейной регрессии.*

Таблица 2

Годы:	1926	1927	1928	1929	1930	1931	1932	1933
Урожай пшеница (ц)	20,1	23,6	26,3	19,9	16,7	23,2	31,9	33,5
Урожай картофель(т)	7,2	7,1	7,4	6,1	6,0	7,3	9,4	9,2

Решение

Таблица 3

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	20,1	7,2	404,1	51,84	144,72
2	23,6	7,1	556,96	50,41	167,56
3	26,3	7,4	691,69	54,76	194,62
4	19,9	6,1	396,01	37,21	121,39
5	16,7	6,0	278,89	36,0	100,2
6	23,2	7,3	538,24	53,29	169,36
7	31,4	9,4	985,96	88,36	295,16
8	33,5	9,2	1122,25	84,64	308,2
	$\sum x = 194.7$	$\sum y = 59.7$	$\sum x^2 = 4974.01$	$\sum y = 456.51$	$\sum xy = 1501.21$

Выборочное уравнение линейной регрессии Y на X имеет вид:

$$Y = \rho_{yx} x + b,$$

$$\rho_{yx} = \rho = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{8 \cdot 1501.21 - 194.7 \cdot 59.7}{8 \cdot 4974.01 - \left(194.7\right)^2} = 0.205,$$

$$b = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2} = \frac{4974.01 \cdot 59.7 - 194.7 \cdot 1501.21}{8 \cdot 4974.01 - \left(194.7\right)^2} = 2.475.$$

Уравнение линейной регрессии Y на X имеет вид:

$$Y = 0.205x + 2.475.$$

Аналогично уравнение линейной регрессии X на Y имеет вид:

$$X = \rho_{xy} y + d,$$

$$\rho_{xy} = \rho = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2} = \frac{8 \cdot 1501.21 - 194.7 \cdot 59.7}{8 \cdot 456.51 - \left(59.7\right)^2} = 4.388,$$

$$d = \frac{\sum y^2 \sum x - \sum y \sum xy}{n\sum y^2 - \left(\sum y\right)^2} = \frac{456.51 \cdot 194.7 - 59.7 \cdot 1501.21}{8 \cdot 456.51 - \left(59.7\right)^2} = -8.407.$$

Таким образом, уравнение линейной регрессии X на Y имеет вид:

$$X = 4.388y - 8.407.$$