

Лабораторная работа №3

Корреляционная зависимость между двумя признаками, построение эмпирической и теоретической линии регрессии

Цель работы: построить эмпирическую линию линейной регрессии, вычислить выборочный коэффициент корреляции, составить теоретическое уравнение линии линейной регрессии и найти доверительный интервал для коэффициента корреляции с надежностью $\gamma = 0,95$

1. Найти условные средние \bar{y}_x записать их в корреляционные таблицы.
2. Построить эмпирическую линию регрессии Y на X . Для этого в системе координат XOY отметить точки (x_i, \bar{y}_x) и соединить их отрезками прямой.
3. Вычислить исправленный выборочный корреляционный момент по формуле

$$S_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^l \sum_{j=1}^m n_{ij} x_i y_j - n \cdot \bar{x} \cdot \bar{y} \right).$$

Найти выборочный коэффициент корреляции

$$r_e = \frac{S_{xy}}{s_x \cdot s_y}.$$

4. Если значение выборочного коэффициента корреляции существенно отличается от единицы, то следует проверить гипотезу H_0 : о равенстве нулю генерального коэффициента корреляции r при конкурирующей гипотезе H_1 : $r \neq 0$. Если нулевая гипотеза будет принята, то выборочный коэффициент незначим, следовательно, X и Y некоррелированы. В качестве критерия проверки нулевой гипотезы примем значение статистики $T = \frac{r_e \sqrt{n-2}}{\sqrt{1-r_e^2}}$. Величина T имеет распределение Стьюдента с $k = n-2$ степенями свободы. При $|T| < t_{kp}$ нет оснований опровергнуть гипотезу H_0 , то есть, X и Y некоррелированы. Если $|T| > t_{kp}$, то гипотеза H_0 отвергается. В этом случае следует принять, что коэффициент корреляции существенно отличен от нуля и признаки X и Y коррелированы. Критические точки распределения Стьюдента приведены в таблице (см. приложение 6 [1]).
5. Если выборочный коэффициент корреляции будет признан значимым, то записать уравнение линейной регрессии Y на X .

$$\hat{y}_x - \bar{y} = r_e \frac{s_y}{s_x} (x - \bar{x}).$$

6. Найти доверительный интервал для коэффициента корреляции надежностью $\gamma = 0,95$ по формуле:

$$r_e - t_\gamma \frac{1-r_e^2}{\sqrt{n}} < r < r_e + t_\gamma \frac{1-r_e^2}{\sqrt{n}},$$

$t_\gamma = t(\gamma, n)$ — коэффициент Стьюдента.

Найдем корреляционную зависимость между X и Y , используя вариационный ряд лабораторной работы №1.

1. Дополним корреляционную таблицу строчкой условных значений признака Y . Для этого вычислим значения признака Y при условии, что X принимает значения x_i :

$$\bar{y}_{x=55,5} = \frac{437 \cdot 4 + 471 \cdot 1}{5} = 443,8;$$

$$\bar{y}_{x=70,5} = \frac{539 \cdot 21 + 573 \cdot 12}{39} = 551,4;$$

$$\bar{y}_{x=60,5} = \frac{471 \cdot 8 + 505 \cdot 1}{9} = 474,8;$$

$$\bar{y}_{x=75,5} = \frac{573 \cdot 6 + 607 \cdot 2}{8} = 581,5;$$

$$\bar{y}_{x=65,5} = \frac{505 \cdot 16 + 539 \cdot 4}{20} = 511,8;$$

$$\bar{y}_{x=80,5} = \frac{607 \cdot 1 + 641 \cdot 2}{3} = 629,7.$$

Таблица 5

Y\X	55,5	60,5	65,5	70,5	75,5	80,5	n_{y_j}
437	4						4
471	1	8	16				9
505		1	4	21			17
539				12	6		25
573					2	1	18
607						2	3
641							2
n_{x_i}	5	9	20	33	8	3	78
\bar{y}_{x_i}	443,8	474,8	511,8	551,4	581,5	629,7	

2. В системе координат XOY отметим точки с координатами $(x_i; \bar{y}_{x_i})$ и соединим их отрезками прямой. Получим ломаную эмпирическую линию регрессии Y на X .

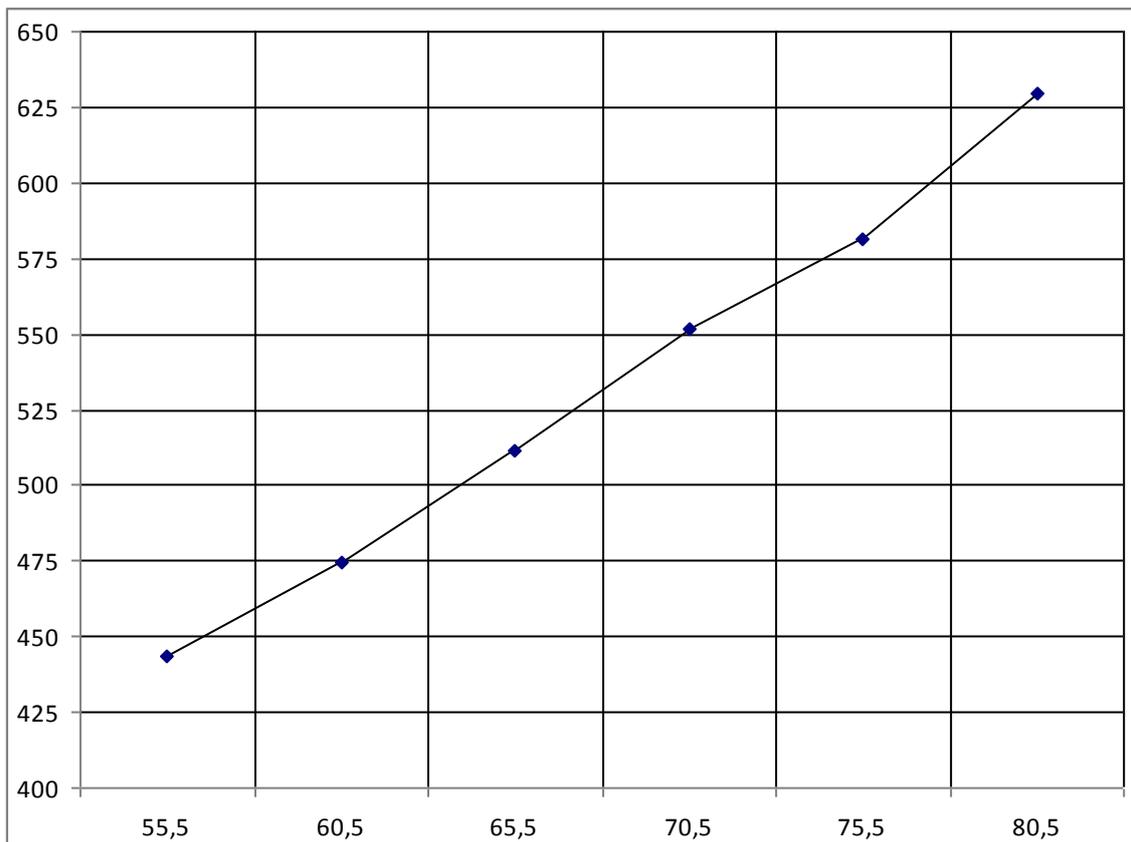


Рис. 3 Эмпирическая линия регрессии Y на X

3. Для вычисления значения суммы в формуле выборочного корреляционного момента S_{xy} составим расчетную таблицу 6.

Таблица 6

x_i	y_i	n_{ij}	$x_i \cdot y_j \cdot n_{ij}$
55,5	437	4	97014,0
55,5	471	1	26140,5
60,5	471	8	227964,0
60,5	505	1	30552,5
65,5	505	16	529240,0
65,5	539	4	141218,0
70,5	539	21	797989,5
70,5	573	12	484758,0
75,5	573	6	259569,0
75,5	607	2	91657,0
80,5	607	1	48863,5
80,5	641	2	103201,0
Σ			2838167

$$S_{xy} = \frac{1}{78-1} \cdot (2838167 - 78 \cdot 68 \cdot 531,656) = 237,19$$

$$r_6 = \frac{237,19}{5,73 \cdot 44,71} = 0,93$$

4. Так как полученное значение выборочного коэффициента корреляции $r_g = 0,93$ близко к единице, то X и Y следует признать коррелированными.
5. Запишем уравнение теоретической прямой линии регрессии Y на X :

$$\hat{y}_x - 531,656 = 0,93 \cdot \frac{44,71}{5,73} (x - 68) \quad \text{или}$$

$$\hat{y}_x = 7,26x + 38,21.$$

Найдем значения $\hat{y}(x_i)$ и занесем их в таблицу 7 .

Таблица 7

x_i	55,5	60,5	65,5	70,5	75,5	80,5
\bar{y}_{x_i}	443,8	474,8	511,8	551,4	581,5	629,7
$\hat{y}(x_i)$	441,1	477,4	513,7	550,0	586,3	622,6

Построим теоретическую прямую регрессии Y на X рис.4 .

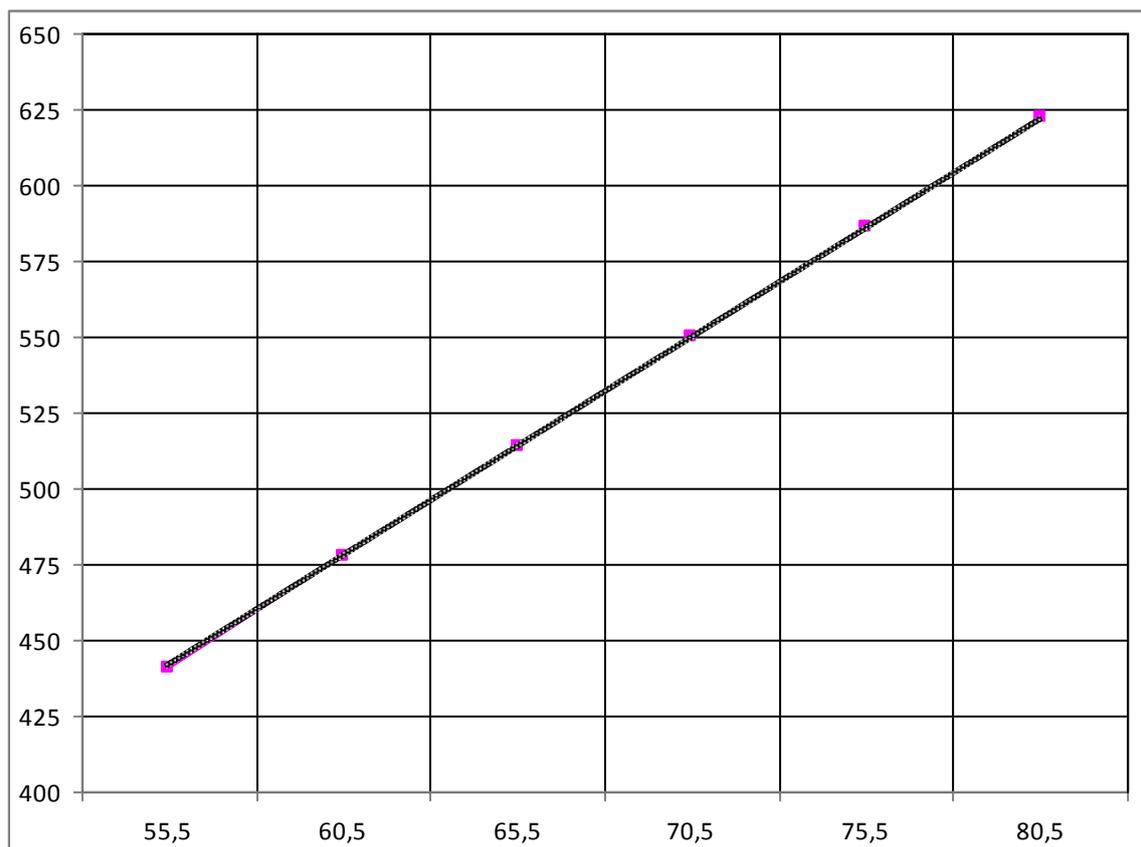


Рис.4 Теоретическая прямая регрессии Y на X

5. Найдем доверительный интервал для коэффициента корреляции r . Значение $t_\gamma = t(\gamma, n)$ найдем по таблице значений коэффициента Стьюдента (см. приложение 3 [1]). Для $\gamma = 0,95$ и $n = 78$ $t_\gamma = 1,991$.

$$0,93 - 1,991 \cdot \frac{1 - 0,93^2}{\sqrt{78}} < r < 0,93 + 1,991 \cdot \frac{1 - 0,93^2}{\sqrt{78}},$$

$$0,90 < r < 0,96.$$

- Вывод:** 1. Т.к. значение выборочного коэффициента корреляции близко к единице ($r_s = 0,93$), корреляционная зависимость величин X и Y тесная, прямая (т.е. с увеличением X в среднем растет и Y, $r > 0$) и близка к линейной.
2. Сравнивая \bar{y}_{x_i} и $\hat{y}(x_i)$, видим, что теоретическая прямая регрессии хорошо согласуется с данными выборки.

Литература

1. Гмурман В.Е., Руководство к решению задач по теории вероятностей и математической статистике. — М.: Высшая школа, 2005.
2. Гмурман В.Е., Теория вероятностей и математическая статистика. — М.: Высшая школа, 2005.
3. Колде Я.К., Практикум по теории вероятностей и математической статистике. — М. Высшая школа, 1991.

