ВЫЧИСЛЕНИЕ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ И ПОСТРОЕНИЕ ЭМПИРИЧЕСКОЙ И ТЕОРЕТИЧЕСКОЙ ЛИНИИ РЕГРЕССИИ

Цель работы: ознакомление с прямолинейной корреляцией; выработка умения и навыков вычисления выборочного коэффициента корреляции и составления уравнений теоретических линий регрессии.

Содержание работы: на основе опытных данных вычислить выборочный коэффициент корреляции; построить для него доверительный интервал с надежностью $\gamma = 0{,}95$; дать смысловую характеристику полученного результата; построить эмпирическую и теоретическую линии регрессии Y на X по предложенной ниже методике; вычислить корреляционное отношение.

МЕТОД КОРРЕЛЯЦИИ

С помощью метода корреляции в математической статистике изучается взаимосвязь явлений. Особенность изучения этой взаимосвязи состоит в том, что нельзя изолировать влияние посторонних факторов. Поэтому метод корреляции применяется для того, чтобы при сложном взаимодействии посторонних влияний выяснить, какова была бы зависимость между признаками, если бы посторонние факторы не изменялись.

В корреляции рассматриваются две задачи: 1) определение характера корреляционной связи между обследуемыми признаками; 2) определение тесноты этой связи. О характере связи между признаками X и Y можно судить по расположению точек в системе координат. Если эти точки располагаются около прямой, то предполагается, что между Y_x и X существует линейная зависимость. Уравнение $\overline{Y}_x = \varphi(x)$ называется уравнением линии регрессии Y на X.

Уравнение $\overline{X}_y = \psi(y)$ называется уравнением линии регрессии, X на Y.

Уравнения прямых регрессии

$$\overline{Y}_x - \overline{Y} = \rho_{y/x} \cdot \left(X - \overline{X}\right) \text{ и } \overline{X}_y - \overline{X} = \rho_{x/y} \cdot \left(Y - \overline{Y}\right)$$

составляются на основании выборочных данных, приведенных в корреляционной таблице.

$\overline{X}, \overline{Y}$ – средние значения соответствующих признаков;

$\rho_{y/x}, \rho_{x/y}$ – коэффициенты регрессии У на Х и Х на У – вычисляются по формулам

$$\rho_{y/x} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\delta_x^2}; \quad \rho_{x/y} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\delta_y^2}$$

где $\overline{XY}$ – среднее значение произведения X на $У$;

δ_x^2 и δ_y^2 – дисперсии признаков X и $У$.

В прямолинейной корреляции теснота связи между признаками характеризуется выборочным коэффициентом корреляции «r_b», который принимает значения в пределах от «-1» до «1».

Если значение коэффициента корреляции отрицательное, то это говорит об обратной связи между изучаемыми признаками; если оно положительное – о прямой связи. Если коэффициент корреляции равен 0, то связи между признаками нет.

Теснота связи	Величина r	
	Прямая связь	Обратная связь
Линейной связи нет	$0 \div 0{,}2$	$0 \div -0{,}2$
Слабая	$0{,}2 \div 0{,}5$	$-0{,}2 \div -0{,}5$
Средняя	$0{,}5 \div 0{,}75$	$-0{,}5 \div -0{,}75$
Сильная	$0{,}75 \div 0{,}95$	$-0{,}75 \div -0{,}95$
Функциональная	$0{,}95 \div 1$	$-0{,}95 \div -1$

Выборочный коэффициент корреляции

вычисляется по формуле (1):

$$r_b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\delta_x \cdot \delta_y}$$

где $\overline{XY}$ – среднее значение произведений X на Y,

$\overline{X}$, $\overline{Y}$ – средние значения соответствующих признаков,

δ_x, δ_y – средние квадратические отклонения, найденные для признака X и для признака Y.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Дана таблица значений температуры смазочного масла заднего моста автомобиля *У* в зависимости от температуры окружающего воздуха *Х*.

Таблица 1

У	4	8	12	16	12	12	12	12	16	4	12	12	12	4	8	8	4
Х	5	15	15	15	35	15	35	15	35	5	15	5	25	25	25	25	25
У	12	16	8	12	8	24	12	12	12	16	12	16	12	16	16	20	12
Х	25	25	25	25	25	65	35	35	35	45	35	45	35	15	35	45	35
У	16	12	20	16	16	20	16	20	16	20	16	20	20	20	24	20	
Х	45	35	45	55	55	45	55	45	55	45	55	55	55	55	55	55	

1. ВЫЧИСЛЕНИЕ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Данные таблицы 1 сведем в корреляционную таблицу.

Таблица 2

y \ x	5	15	25	35	45	55	65	m_y
4	2		2					4
8		1	4					5
12		4	3	10				17
16		2		2	3	6		13
20					5	4		9
24						1	1	2
m_x								

Процесс вычисления упростим, переходя к **условным вариантам** u_i и v_j

$$u_i = \frac{X_i - C_1}{h_1}; \quad v_j = \frac{Y_j - C_2}{h_2}$$

где C_1 – ложный нуль для X;

h_1 – шаг значений признака X;

C_2 – ложный нуль для Y;

h_2 – шаг значений признака Y.

В этом случае

$$r_b = \frac{\sum m_{uv} \cdot u \cdot v - n \cdot \bar{u} \cdot \bar{v}}{n \cdot \delta_u \cdot \delta_v} \quad (2)$$

где n – объем выборки;

m_{uv} – частота пары вариант u и v;

$$\bar{u} = \frac{\sum m_u \cdot u}{n}; \quad \bar{v} = \frac{\sum m_v \cdot v}{n};$$

$$\delta_u = \sqrt{\overline{u^2} - \left(\bar{u}\right)^2}; \quad \delta_v = \sqrt{\overline{v^2} - \left(\bar{v}\right)^2}$$

Из корреляционной таблицы выбираем наибольшую частоту

$$m_{xy}=10;\ C_1=35;\ C_2=12$$

Составляем корреляционную таблицу 3 в условных вариантах, где наибольшая частота $m_{xy}=10$ кодируется 0 в рядах u и v.

Все результаты приведем в этой таблице.

Таблица 3

v \ u	-3	-2	-1	0	1	2	3	m_v	$m_v v$	$m_v v^2$
-2	2 (6)		2 (2)					4	-8	16
-1		1 (2)	4 (1)					5	-5	5
0		4 (0)	3 (0)	10 (0)				17	0	0
1		2 (-2)		2 (0)	3 (1)	6 (2)		13	13	13
2					5 (2)	4 (4)		9	18	36
3						1 (6)	1 (9)	2	6	13
m_u	2	7	9	12	8	11	1	50	$\Sigma=24$	$\Sigma=88$
$m_u u$	-6	-14	-9	0	8	22	3	$\Sigma=4$		
$m_u u^2$	18	28	9	0	8	44	9	$\Sigma=16$		
$m_{uv} uv$	12	-2	8	0	13	34	9	$\Sigma=74$		

Подставляя результаты вычисления в формулы, получим:

$$\bar{u} = \frac{4}{50} = 0{,}08 \quad ; \quad \bar{v} = \frac{24}{50} = 0{,}48 \quad ;$$

$$\overline{u^2} = \frac{116}{50} = 2{,}32 \quad ; \quad \overline{v^2} = \frac{88}{50} = 1{,}76 \quad ;$$

$$\delta_u = \sqrt{2{,}32 - (0{,}08)^2} = 1{,}52 \; ; \delta_v = \sqrt{1{,}76 - (0{,}48)^2} = 1{,}24 \; ;$$

$$r_b = \frac{74 - 50 \cdot 0{,}08 \cdot 0{,}48}{50 \cdot 1{,}52 \cdot 1{,}24} = 0{,}76 \text{ или } 76\%$$

ВЫВОД. Коэффициент корреляции показывает тесную связь, существующую между температурой смазочного масла заднего моста и температурой окружающего воздуха.

2. ОПРЕДЕЛЕНИЕ НАДЕЖНОСТИ (ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА) КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Для оценки надежности полученного коэффициента корреляции определяют его погрешность по формуле:

$$\delta_{r_b} = \frac{1 - r_b^2}{\sqrt{n}}$$

В предположении, что (X,Y) имеет нормальное распределение или близкое к нему, можно считать коэффициент корреляции распределенным нормально с параметрами r_b и δ_{r_b}

Найдем доверительный интервал для r с надежностью $\gamma = 0{,}95$

$$r_b - t_\gamma \cdot \delta_{r_b} < r < r_b + t_\gamma \cdot \delta_{r_b}$$

t_γ найдем по таблице значений функции $\Phi(X)$.

$$\gamma = 2 \cdot \Phi(X)$$

$$\Phi(X) = \frac{\gamma}{2} = 0{,}475; \quad t_\gamma = 1{,}96;$$

$$\delta_{r_b} = \frac{1 - 0{,}76^2}{\sqrt{50}} = 0{,}059 \approx 0{,}06$$

Доверительный интервал для коэффициента корреляции запишется так:

$$0,76 - 1,96 \cdot 0,06 < r < 0,76 + 1,96 \cdot 0,06$$

$$0,76 - 0,12 < r < 0,76 + 0,12$$

$$0,64 < r < 0,88$$

ВЫВОД. Это означает, что при условиях данного опыта следует ожидать влияние температуры окружающего воздуха на температуру смазочного масла заднего моста не менее, чем на 64%.

3. ПОСТРОЕНИЕ ЭМПИРИЧЕСКОЙ И ТЕОРЕТИЧЕСКОЙ ЛИНИЙ РЕГРЕССИИ *У* НА *Х*

Для построения эмпирической линии регрессии составим таблицу 4.

Таблица 4

X	5	15	25	35	45	55	65
$\overline{Y_x}$	4	12,6	8,44	12,6	18,5	18,2	24

$\overline{Y_x}$ – условная средняя значений признака Y при условий, что X принимает определенное значение, для вычисления $\overline{Y_x}$ воспользуемся таблицей 2.

$$\overline{Y}_{x_1}=\frac{2\cdot 4}{2};\quad \overline{Y}_{x_2}=\frac{1\cdot 8+4\cdot 12+2\cdot 16}{7}\approx\frac{88}{7}=12{,}6;$$

$$\overline{Y}_{x_3}=\frac{2\cdot 4+4\cdot 8+3\cdot 12}{9}=8{,}44;\quad \overline{Y}_{x_4}=\frac{120+32}{12}=12{,}6;$$

$$\overline{Y}_{x_5}=\frac{148}{8}=18{,}5.\quad \overline{Y}_{x_6}=\frac{200}{11}\approx 18{,}2$$

Принимая пары чисел $(X, \overline{Y_x})$ за координаты точек, строим их в системе координат и соединяем отрезками прямой. Полученная ломаная линия и будет **эмпирической линией регрессии**.

Уравнение теоретической прямой линии регрессии Y на X имеет вид:

$$\overline{Y}_x - \overline{Y} = r_b \cdot \frac{\delta_y}{\delta_x} \cdot \left(X - \overline{X}\right)$$

$$r_b \cdot \frac{\delta_y}{\delta_x} = \rho_{y/x}$$

где $\overline{Y}$ – выборочная средняя признака Y;

$\overline{X}$ – выборочная средняя признака X.

$$\overline{Y} = \bar{v} \cdot h_2 + C_2 = 0{,}48 \cdot 4 + 12 = 13{,}92$$

$$\overline{X} = \bar{u} \cdot h_1 + C_1 = 35{,}8$$

$$\delta_x = h_1 \cdot \delta_u = 1{,}52 \cdot 10 = 15{,}2$$

$$\delta_y = h_2 \cdot \delta_v = 1{,}24 \cdot 4 = 4{,}96$$

$$r_b = 0{,}76$$

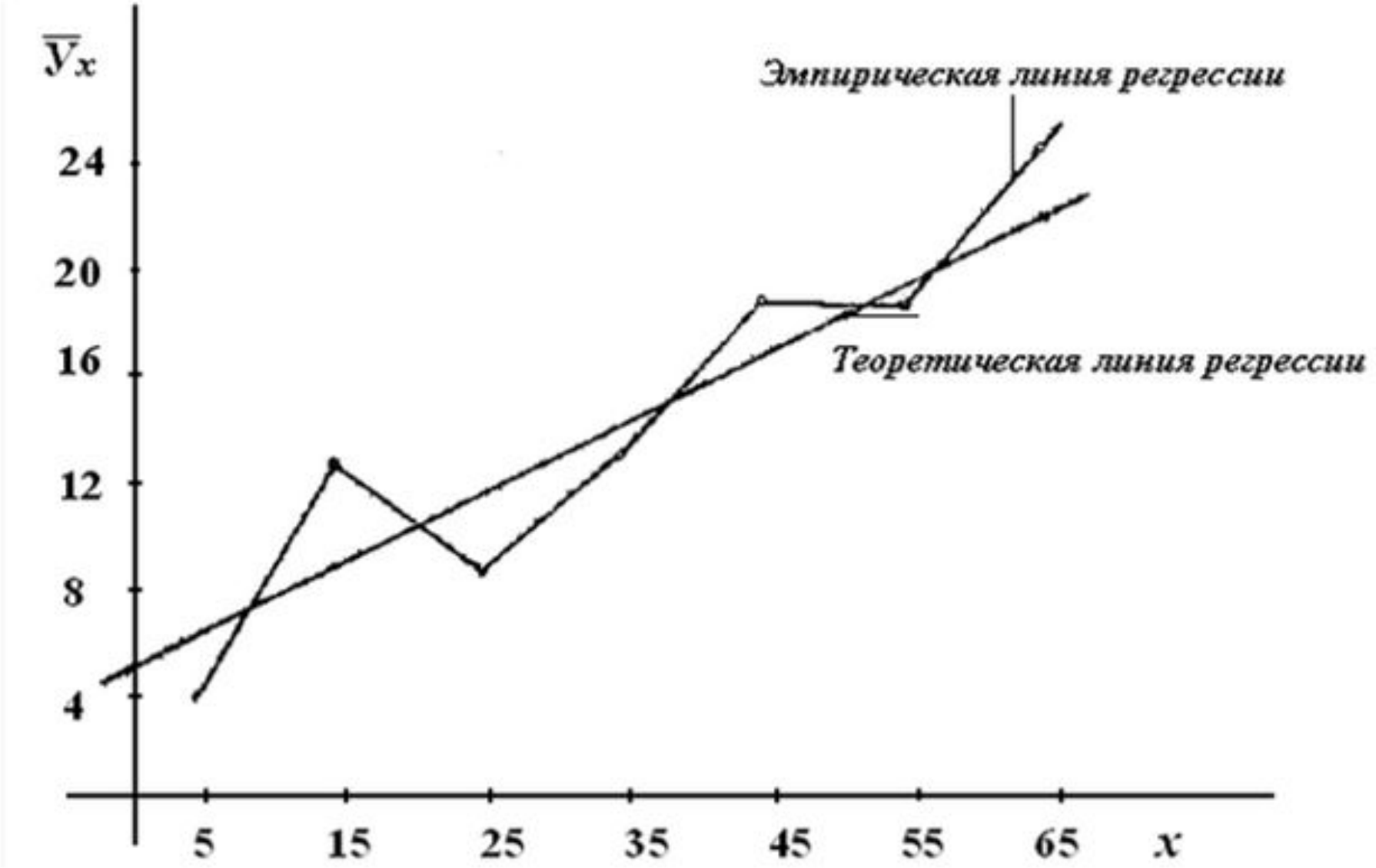
Уравнение прямой регрессии Y на X запишется так:

$$\overline{Y}_x - 13{,}92 = 0{,}76 \cdot \frac{4{,}96}{15{,}2} \cdot (X - 35{,}8)$$

или окончательно

$$\overline{Y}_x = a \cdot x + b; \quad \overline{Y}_x = 0{,}25 \cdot x + 4{,}97$$

Построим обе линии регрессии



при $x = 0$; $\overline{y_x} = 4{,}97$ при $x = 65$; $\overline{y_x} = 21{,}22$;

КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ

Для оценки тесноты связи линейной и нелинейной корреляционной зависимости между признаками Y и X служит корреляционное отношение

$$\eta_{y/x} = \frac{\delta_{\bar{y}_x}}{\delta_y},$$

где δ_y – среднее квадратическое отклонение признака Y;

$\delta_{\bar{y}_x}$ – среднее квадратическое отклонение условных средних относительно $\bar{Y}_x$ общей средней $\bar{Y}$, определяемое из равенства

$$\delta_{\bar{y}_x} = \sqrt{\frac{\sum m_x \cdot (\bar{Y}_x - Y)}{n}}$$

Корреляционное отношение обладает следующими свойствами:

1. Корреляционное отношение всегда находится между нулем и единицей
$$0 \le \eta_{y/x} \le 1$$
2. Если $\eta_{y/x} = 0$, то признак Y с признаком X корреляционной зависимостью не связан.
3. Если $\eta_{y/x} = 1$, то между признаками Y и X существует функциональная зависимость $Y=f(X)$.
4. Чем ближе корреляционное отношение к единице, тем сильнее корреляционная зависимость между признаками Y и X: чем ближе $\eta_{y/x}$ к нулю, тем эта зависимость слабее.
5. Выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции:
$$\eta_{y/x} \ge |r_b|$$
6. Если $\eta_{y/x}$ равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.

Для вычисления корреляционного отношения составим расчетную таблицу 5

№ n/n	m_x	$\overline{y}_x$	$\overline{y}_x-\overline{y}$	$(\overline{y}_x-\overline{y})^2$	$m_x(\overline{y}_x-\overline{y})^2$
1	2	4	-9,92	98,4064	196,8128
2	7	12,6	-1,32	1,7424	12,19568
3	9	8,44	-5,48	30,0304	270,2736
4	12	12,6	-1,32	1,7424	20,9088
5	8	18,5	4,58	20,9764	167,8112
6	11	18,2	4,28	18,3184	201,5024
7	1	24,0	10,08	101,6064	101,6064
Σ					971,1109

В таблице $\overline{y}=13,92;\ n=50;\ \delta_y=4,96.$

Вычислим

$$\delta_{\overline{y}_x}=\sqrt{\frac{971,1109}{50}}\approx 4,41,$$

тогда

$$\eta_{y/x}=\frac{4,41}{4,96}\approx 0,89$$

ФОРМА ОТЧЕТА

1. Условие задачи (табл.2).
2. Вычисление коэффициента корреляции методом условных вариант (табл.3).
3. Вывод о связи между обследуемыми признаками.
4. Определение надежности коэффициента корреляции.
5. Вычисление условных средних $\overline{Y}_x$ (табл.4).
6. Построение эмпирической линии регрессии (рис.5).
7. Построение теоретической линии регрессии «*У*» на «*Х*».
8. Вычисление корреляционного отношения (табл.5).