Распределение 50 предприятий по размерам основных производственных фондов X (млн. руб.) и выпуску продукции Y (млн. руб.) дано в таблице.

X \ У	40-50	50-60	60-70	70-80	80-90	90-100	Итого
35-45	1	1	1				3
45-55		3	2				5
55-65		4	1	11			16
65-75				6	9		15
75-85				1	1	1	3
85-95					4	4	8
Итого	1	8	4	18	14	5	50

Необходимо:

1. Вычислить групповые средние $\overline{x}_i$ и $\overline{y}_j$, построить эмпирические линии регрессии.

2. Предполагая, что между переменными X и Y существует линейная корреляционная зависимость:

а) найти уравнения прямых регрессии, построить их графики на одном чертеже с эмпирическими линиями и дать экономическую интерпретацию полученных уравнений;

б) вычислить коэффициент корреляции; на уровне значимости $\alpha = 0,05$ оценить его значимость и сделать вывод о тесноте и направлении связи между переменными X и Y;

в) используя соответствующее уравнение регрессии, оценить средний выпуск продукции предприятия, основные производственные фонды (ОПФ), которого составляют 81 млн. руб.

Решение.

В корреляционную таблицу заносим данные заданные по условию задачи.

y / x			40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
	y_j												
	x_i	v_j / u_i											
35-45			1	1	1				3				
45-55				3	2				5				
55-65				4	1	11			16				
65-75						6	9		15				
75-85						1	1	1	3				
85-95							4	4	8				
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$			
$\overline{x}_y$									$\sum$				
$v_j n_y$													
$v_j^2 n_y$													
$\sum v_j u_i n_{ij}$													

От интервальных рядов перейдем к дискретным рядам x_i и y_j, используя формулы

$$x_i = \frac{x_{i-1} + x_{i+1}}{2} \quad \text{и} \quad y_j = \frac{y_{j-1} + y_{j+1}}{2}.$$

Получаем

$$x_1 = \frac{35+45}{2} = \frac{80}{2} = 40 \qquad y_1 = \frac{40+50}{2} = \frac{90}{2} = 45$$

$$x_2 = \frac{45+55}{2} = \frac{100}{2} = 50 \qquad y_2 = \frac{50+60}{2} = \frac{110}{2} = 55$$

$$x_3 = \frac{55+65}{2} = \frac{120}{2} = 60 \qquad y_3 = \frac{60+70}{2} = \frac{130}{2} = 65$$

$$x_4 = \frac{65+75}{2} = \frac{140}{2} = 70 \qquad y_4 = \frac{70+80}{2} = \frac{150}{2} = 75$$

$$x_5 = \frac{75+85}{2} = \frac{160}{2} = 80 \qquad y_5 = \frac{80+90}{2} = \frac{170}{2} = 85$$

$$x_6 = \frac{85+95}{2} = \frac{180}{2} = 90 \qquad y_6 = \frac{90+100}{2} = \frac{190}{2} = 95$$

Вычисленные серединные значения интервальных рядов заносим в корреляционную таблицу.

x	x_i	v_j / u_i	y: 40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
		y_j	45	55	65	75	85	95					
		v_j											
35-45	40		1	1	1				3				
45-55	50			3	2				5				
55-65	60			4	1	11			16				
65-75	70					6	9		15				
75-85	80					1	1	1	3				
85-95	90						4	4	8				
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$			
$\overline{x}_y$									$\sum$				
$v_j n_y$													
$v_j^2 n_y$													
$\sum v_j u_i n_{ij}$													

http://exponenta.ucoz.ru/

Для упрощения расчетов перейдем к новым условным вариантам u_i и v_j:

$$u_i = \frac{x_i - C_1}{k_1} \quad \text{и} \quad v_j = \frac{y_j - C_2}{k_2},$$

где C_1 и C_2 – середины серединных интервалов, k_1 и k_2 – ширина интервалов по x, и по Y соответственно.

При четном числе интервалов пользуемся правилом: рекомендуется взять середину интервала, имеющего большую частоту.

Анализируя интервальные ряды по x и по Y приходим к выводу: $k_1 = 10$, $C_1 = 60$ (т.к. $n_x = 16 > 15$),

$k_2 = 10$, $C_2 = 75$ ($n_y = 18 > 4$).

$$u_1 = \frac{40-60}{10} = -\frac{20}{10} = -2 \qquad v_1 = \frac{45-75}{10} = -\frac{30}{10} = -3$$

$$u_2 = \frac{50-60}{10} = -\frac{10}{10} = -1 \qquad v_2 = \frac{55-75}{10} = -\frac{20}{10} = -2$$

$$u_3 = \frac{60-60}{10} = 0 \qquad v_3 = \frac{65-75}{10} = -\frac{10}{10} = -1$$

$$u_4 = \frac{70-60}{10} = \frac{10}{10} = 1 \qquad v_4 = \frac{75-75}{10} = 0$$

$$u_5 = \frac{80-60}{10} = \frac{20}{10} = 2 \qquad v_5 = \frac{85-75}{10} = \frac{10}{10} = 1$$

$$u_6 = \frac{90-60}{10} = \frac{30}{10} = 3 \qquad v_6 = \frac{95-75}{10} = \frac{20}{10} = 2$$

Вычисленные условные варианты заносим в корреляционную таблицу.

y			40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
		y_j	45	55	65	75	85	95					
x	x_i	v_j / u_i	−3	−2	−1	0	1	2					
35-45	40	−2	1	1	1				3				
45-55	50	−1		3	2				5				
55-65	60	0		4	1	11			16				
65-75	70	1				6	9		15				
75-85	80	2				1	1	1	3				
85-95	90	3					4	4	8				
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$			
$\overline{x}_y$									$\sum$				
$v_j n_y$													
$v_j^2 n_y$													
$\sum v_j u_i n_{ij}$													

Найдем вспомогательные величины

$u_i n_x$, $u_i^2 \cdot n_x$, $v_j n_y$, $v_j^2 \cdot n_y$.

И найдем соответствующие суммы

$\sum u_i n_x$, $\sum u_i^2 n_x$, $\sum v_j n_y$, $\sum v_j^2 \cdot n_y$

Результат вычислений занесем в соответствующие строки и столбцы корреляционной таблицы.

Для вычисления сумм вида $\sum v_j u_i n_{ij}$ и $\sum u_i v_j n_{ij}$ найдем произведения $u_i v_j$ и запишем их под диагональю в соответствующей клетке корреляционной таблицы.

x	x_i	u_i \ v_j	40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
		y_j	45	55	65	75	85	95					
		v_j	−3	−2	−1	0	1	2					
35-45	40	−2	1 / 6	1 / 4	1 / 2				3		−6	12	
45-55	50	−1		3 / 2	2 / 1				5		−5	5	
55-65	60	0		4 / 0	1 / 0	11 / 0			16		0	0	
65-75	70	1				6 / 0	9 / 1		15		15	15	
75-85	80	2				1 / 0	1 / 2	1 / 4	3		6	12	
85-95	90	3					4 / 3	4 / 6	8		24	72	
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$	34	116	
$\overline{x}_y$									$\sum$				
$v_j n_y$			−3	−16	−4	0	14	10	1				
$v_j^2 n_y$			9	32	4	0	14	20	79				
$\sum v_j u_i n_{ij}$													

Далее, по каждой строке и по каждому столбцу находим произведения $u_i v_j n_{ij}$, и полученные значения складываем.

Например, $\sum u_1 v_j n_{1j} = 1 \cdot 6 + 1 \cdot 4 + 1 \cdot 2 = 12$.

Аналогично находим остальные суммы.

Если все расчеты выполнены верно, то сумма значений столбца $\sum u_i v_j n_{ij}$ должна совпасть с суммой значений строки $\sum v_j u_i n_{ij}$.

x	x_i	y / y_j / $u_i \backslash v_j$	40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
		y_j	45	55	65	75	85	95					
		v_j	−3	−2	−1	0	1	2					
35-45	40	−2	1 / 6	1 / 4	1 / 2				3		−6	12	12
45-55	50	−1		3 / 2	2 / 1				5		−5	5	8
55-65	60	0		4 / 0	1 / 0	11 / 0			16		0	0	0
65-75	70	1				6 / 0	9 / 1		15		15	15	9
75-85	80	2				1 / 0	1 / 2	1 / 4	3		6	12	6
85-95	90	3					4 / 3	4 / 6	8		24	72	36
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$	34	116	
$\overline{x}_y$									$\sum$				
$v_j n_y$			−3	−16	−4	0	14	10	1				
$v_j^2 n_y$			9	32	4	0	14	20	79				
$\sum v_j u_i n_{ij}$			6	10	4	0	23	28					71

Групповые средние найдем по формулам:

$$\overline{y}_x = \frac{\sum y_j n_{ij}}{n_x} \quad \text{и} \quad \overline{x}_y = \frac{\sum x_i \cdot n_{ij}}{n_y}$$

Получаем,

$$\overline{y}_{x=40} = \frac{45 \cdot 1 + 55 \cdot 1 + 65 \cdot 1}{3} = \frac{165}{3} = 55$$

$$\overline{y}_{x=50} = \frac{55 \cdot 3 + 65 \cdot 2}{5} = \frac{295}{5} = 59$$

$$\overline{y}_{x=60} = \frac{55 \cdot 4 + 65 \cdot 1 + 75 \cdot 11}{16} = \frac{1110}{16} = 69,375$$

$$\overline{y}_{x=70} = \frac{75 \cdot 6 + 85 \cdot 9}{15} = \frac{1215}{15} = 81$$

$$\overline{y}_{x=80} = \frac{75 \cdot 1 + 85 \cdot 1 + 95 \cdot 1}{3} = \frac{255}{3} = 85$$

$$\overline{y}_{x=90} = \frac{85 \cdot 4 + 95 \cdot 4}{8} = \frac{720}{8} = 90$$

$$\overline{x}_{y=45} = \frac{40 \cdot 1}{1} = 40$$

$$\overline{x}_{y=55} = \frac{40 \cdot 1 + 50 \cdot 3 + 60 \cdot 4}{8} = \frac{430}{8} = 53,75$$

$$\overline{x}_{y=65} = \frac{40 \cdot 1 + 50 \cdot 2 + 60 \cdot 1}{4} = \frac{200}{4} = 50$$

$$\overline{x}_{y=75} = \frac{60 \cdot 11 + 70 \cdot 6 + 80 \cdot 1}{18} = \frac{1160}{18} = 64,444$$

$$\overline{x}_{y=85} = \frac{70 \cdot 9 + 80 \cdot 1 + 90 \cdot 4}{14} = \frac{1070}{14} = 76,428$$

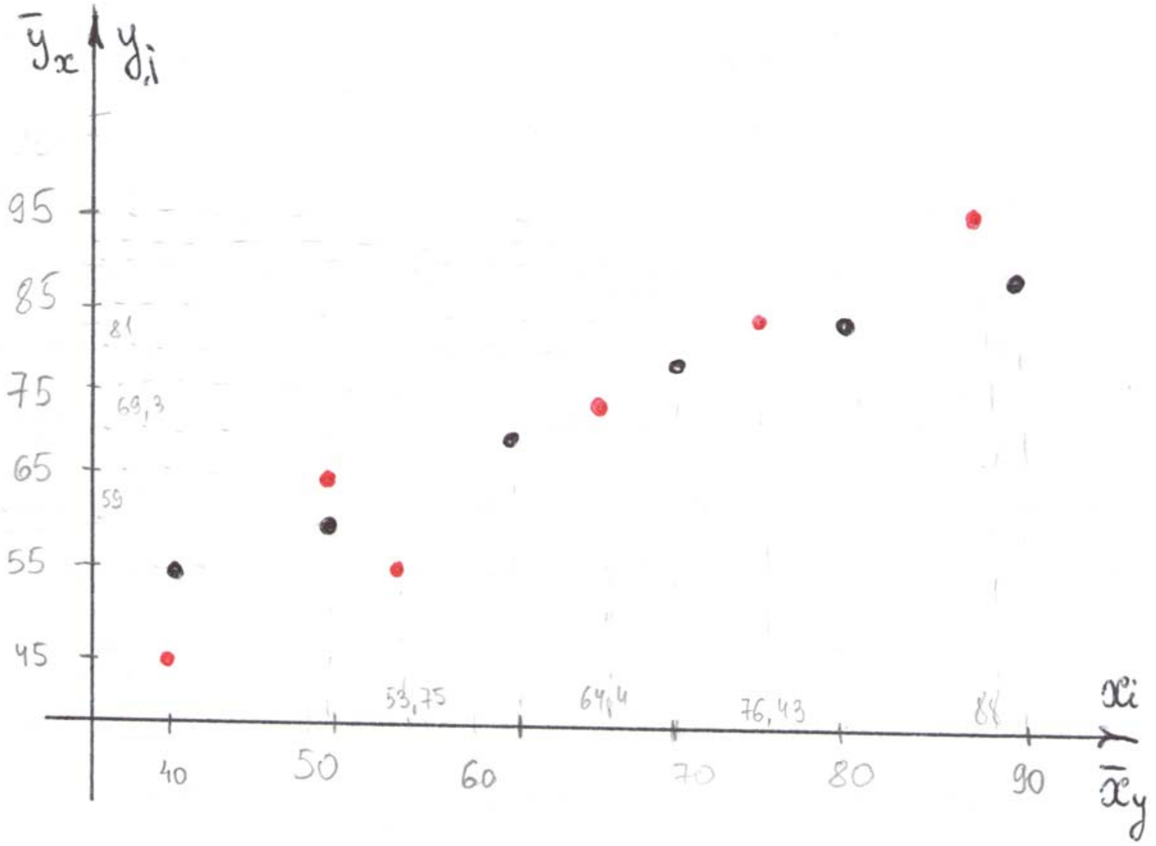
$$\overline{x}_{y=95} = \frac{80 \cdot 1 + 90 \cdot 4}{5} = \frac{440}{5} = 88$$

Вычисленные групповые средние $\overline{x}_y$ и $\overline{y}_x$ заносим в корреляционную таблицу.

x \ y			40-50	50-60	60-70	70-80	80-90	90-100	n_x	$\overline{y}_x$	$u_i n_x$	$u_i^2 n_x$	$\sum u_i v_j n_{ij}$
		y_j	45	55	65	75	85	95					
	x_i	u_i \ v_j	−3	−2	−1	0	1	2					
35-45	40	−2	1 / 6	1 / 4	1 / 2				3	55	−6	12	12
45-55	50	−1		3 / 2	2 / 1				5	59	−5	5	8
55-65	60	0		4 / 0	1 / 0	11 / 0			16	69,375	0	0	0
65-75	70	1				6 / 0	9 / 1		15	81	15	15	9
75-85	80	2				1 / 0	1 / 2	1 / 4	3	85	6	12	6
85-95	90	3					4 / 3	4 / 6	8	90	24	72	36
n_y			1	8	4	18	14	5	50 $n_x = n_y$	$\sum$	34	116	
$\overline{x}_y$			40	53,75	50	64,444	76,428	88	$\sum$				
$v_j n_y$			−3	−16	−4	0	14	10	1				
$v_j^2 n_y$			9	32	4	0	14	20	79				
$\sum v_j u_i n_{ij}$			6	10	4	0	23	28					$\boxed{71}$

На координатной плоскости $xO\bar{y}_x$ и $\bar{x}_yOy$

Отмечаем точки с координатами $(x_i; \bar{y}_x)$ и $(\bar{x}_y; y_j)$



$(x_i; \bar{y}_x)$ – эмпирическая точечная линия y на x

$(\bar{x}_y; y_j)$ – эмпирическая точечная линия x на y

Пусть между X и Y существует линейная корреляционная зависимость. Для нахождения уравнений прямых регрессии посчитаем:

- средние арифметические $\overline{x}$, $\overline{y}$;
- дисперсии S_x^2, S_y^2;
- коэффициент ковариации μ.

Получаем,

$$\overline{x} = k_1 \overline{u} + C_1 = k_1 \cdot \frac{\sum u_i n_x}{n} + C_1 = 10 \cdot \frac{34}{50} + 60 = 0{,}68 + 60 =$$

$= 66{,}8$ (млн. руб.) — средняя величина ОПФ.

$$\overline{y} = k_2 \overline{v} + C_2 = k_2 \cdot \frac{\sum v_j n_y}{n} + C_2 = 10 \cdot \frac{1}{50} + 75 = 0{,}2 + 75 =$$

$= 75{,}2$ (млн. руб.) — средняя стоимость выпуска продукции.

$$S_x^2 = k_1^2 \left(\overline{u^2} - \overline{u}^2 \right) = k_1^2 \cdot \left(\frac{\sum u_i^2 \cdot n_x}{n} - \left(\frac{\sum u_i \cdot n_x}{n} \right)^2 \right) =$$

$$= 10^2 \cdot \left(\frac{116}{50} - \left(\frac{34}{50} \right)^2 \right) = 100 \cdot \left(2{,}32 - 0{,}68^2 \right) =$$

$= 185{,}76$ — дисперсия стоимости ОПФ

$$S_y^2 = k_2^2\left(\overline{v^2} - \overline{v}^2\right) = k_2^2 \cdot \left(\frac{\sum v_j^2 \cdot n_y}{n} - \left(\frac{\sum v_j n_y}{n}\right)^2\right) =$$

$$= 10^2 \cdot \left(\frac{79}{50} - \left(\frac{1}{50}\right)^2\right) = 100 \cdot \left(1,58 - 0,02^2\right) =$$

$= 157,96$ – дисперсия стоимости выпуска продукции.

$$\mu = k_1 \cdot k_2 \cdot \left(\overline{uv} - \overline{u} \cdot \overline{v}\right) = k_1 \cdot k_2 \cdot \left(\frac{\sum u_i v_j n_{ij}}{n} - \frac{\sum u_i n_x}{n} \cdot \frac{\sum v_j n_y}{n}\right) =$$

$$= 10 \cdot 10 \cdot \left(\frac{71}{50} - \frac{34}{50} \cdot \frac{1}{50}\right) = 100 \cdot \left(1,42 - 0,68 \cdot 0,02\right) =$$

$= 140,64$

Уравнение регрессии y по x будем искать в виде:

$$\overline{y}_x - \overline{y} = b_{yx} \cdot (x - \overline{x}),$$

где $b_{yx} = \frac{\mu}{S_x^2}$ – коэффициент регрессии y по x.

Найдем значение b_{yx}, получаем

$$b_{yx} = \frac{\mu}{S_x^2} = \frac{140,64}{185,76} = 0,7571$$

Геометрически b_{yx} означает тангенс угла наклона прямой регрессии y по x к оси абсцисс, т.е.

$$b_{yx} = tg\,\alpha = 0,7571 \Rightarrow \alpha = arctg\,0,7571 \approx 37^\circ$$

С экономической точки зрения $b_{yx} = 0,7571$ показывает, что с увеличением x – стоимости ОПФ на 1 млн. руб., стоимость выпуска продукции в среднем увеличится на 0,7571 млн. руб.

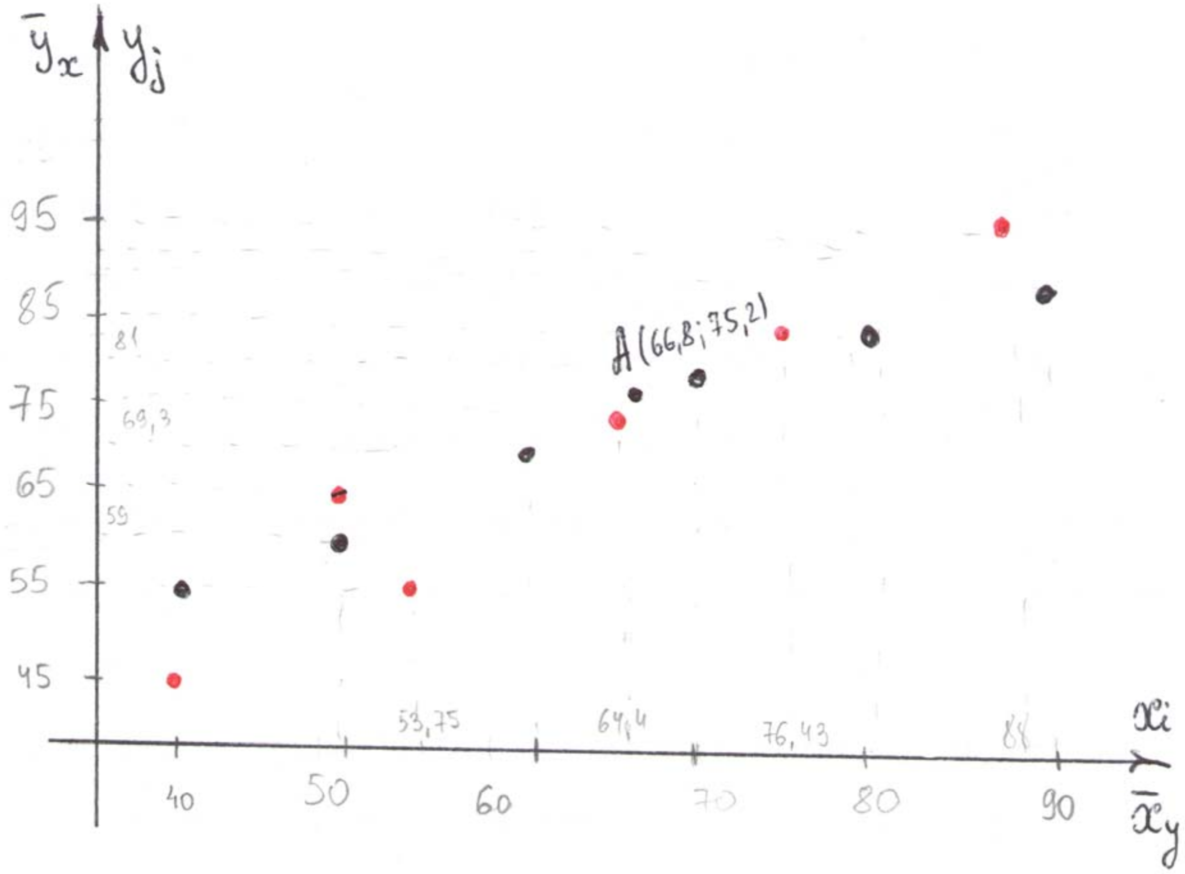
Подставим значения $\bar{x}$, $\bar{y}$ и b_{yx} в уравнение регрессии y по x, получаем

$$\bar{y}_x - 75,2 = 0,7571 \cdot (x - 66,8)$$

$$\bar{y}_x = 0,7571x - 0,7571 \cdot 66,8 + 75,2$$

$\bar{y}_x = 0,757x + 24,626$ – уравнение прямой регрессии y по x, показывает зависимость средней стоимости продукции от стоимости ОПФ.

Возвращаемся к графику с эмпирическими точечными линиями регрессии и отмечаем точку $A(\bar{x}; \bar{y})$, т.е. $A(66,8; 75,2)$.



$(x_i; \bar{y}_x)$ - эмпирическая точечная линия у на х

$(\bar{x}_y; y_j)$ - эмпирическая точечная линия х на у

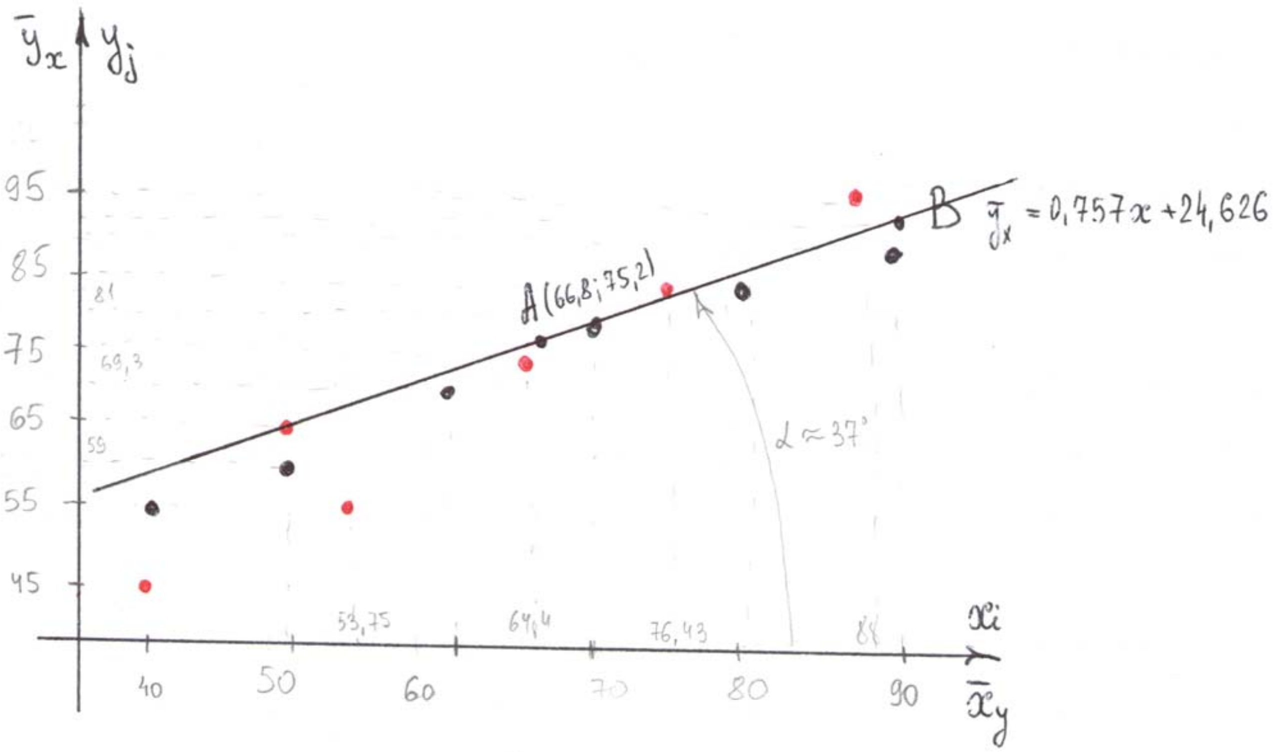
Для построения графика прямой регрессии у по х найдем координаты дополнительной точки В используя полученное уравнение регрессии:

$$\bar{y}_{x=90} = 0,757 \cdot 90 + 24,626 = 92,756$$

Получаем В (90; 92,76)

Через точки A и B проводим линию прямой регрессии y по x.

Отмечаем угол α между прямой регрессии и осью абсцисс.



$(x_i; \bar{y}_x)$ – эмпирическая точечная линия y на x

$(\bar{x}_y; y_j)$ – эмпирическая точечная линия x на y

Аналогично находим уравнение регрессии x по y в виде:

$$\bar{x}_y - \bar{x} = b_{xy} \cdot (y - \bar{y}),$$

где $b_{xy} = \dfrac{\mu}{S_y^2}$ - коэффициент регрессии x по y.

Найдем значение b_{xy}, получаем

$$b_{xy} = \frac{\mu}{S_y^2} = \frac{140,64}{157,96} = 0,8904$$

Геометрически коэффициент регрессии b_{xy} означает тангенс угла наклона прямой регрессии x по y к оси ординат, т.е.

$$b_{xy} = tg\,\beta = 0,8904 \quad \Rightarrow \quad \beta = arctg\,0,8904 \approx 42^\circ$$

Экономическая интерпретация выборочного коэффициента регрессии b_{xy}: чтобы увеличить стоимость выпускаемой продукции на 1 млн. руб. необходимо увеличить ОПФ на 0,8904 млн. руб. в среднем.

Подставим значения $\bar{x}$, $\bar{y}$ и b_{xy} в уравнение регрессии x по y, получаем

$$\bar{x}_y - 66{,}8 = 0{,}8904\,(y - 75{,}2)$$

$$\bar{x}_y = 0{,}8904\,y - 0{,}8904 \cdot 75{,}2 + 66{,}8$$

$\bar{x}_y = 0{,}890\,y - 0{,}158$ – уравнение прямой регрессии x по y, показывает зависимость средней стоимости ОПФ от стоимости выпущенной продукции.

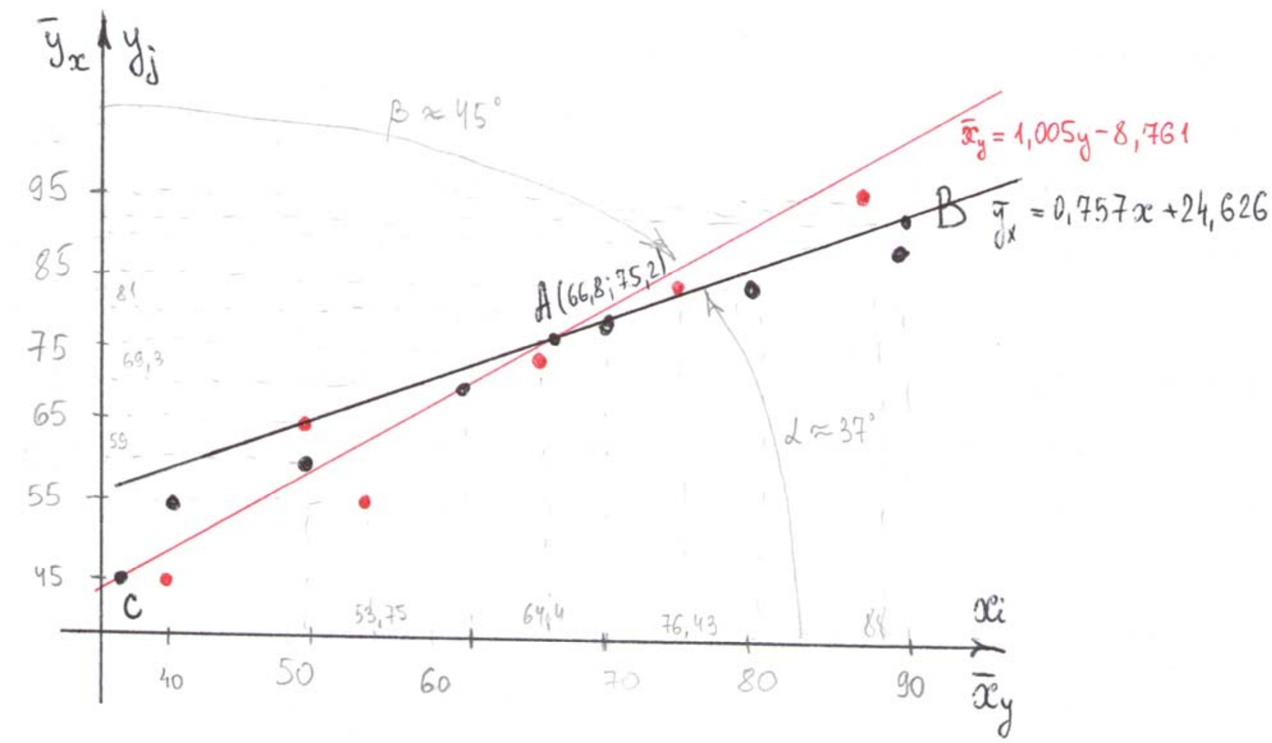
Для построения графика прямой регрессии x по y найдем координаты дополнительной точки C используя полученное уравнение регрессии:

$$\bar{x}_{y=45} = 0{,}890 \cdot 45 - 0{,}158 = 39{,}892$$

Получаем C (39,89 ; 45).

Через точки A и C проводим линию прямой регрессии x по y.

Отмечаем угол β между прямой регрессии и осью ординат.



$(x_i; \bar{y}_x)$ – эмпирическая точечная линия y на x

$(\bar{x}_y; y_j)$ – эмпирическая точечная линия x на y

http://exponenta.ucoz.ru/

Вычислим коэффициент корреляции по формуле

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}.$$

Правило определения знака перед корнем: знак плюс выбирают если коэффициенты регрессии одновременно положительны, знак минус – коэффициенты регрессии b_{yx} и b_{xy} одновременно отрицательны.

Получаем, $r = +\sqrt{0{,}7571 \cdot 0{,}8904} = 0{,}82$

Вычисленный коэффициент корреляции является выборочным т.к. получен на основе выборочных данных. Оценим значимость коэффициента корреляции с помощью t-критерия Стьюдента на уровне значимости $\alpha = 0{,}05$. Для этого выдвинем нулевую гипотезу H_0.

H_0: линейная корреляционная зависимость в генеральной совокупности отсутствует.

Чтобы проверить гипотезу вычислим значение статистики $t_{эмп}$ по формуле:

$$t_{эмп} = \sqrt{\frac{r^2}{1-r^2}(n-2)}.$$

Получаем, $t_{эмп} = \sqrt{\frac{0{,}82^2}{1-0{,}82^2}\cdot(50-2)} = \sqrt{98{,}5201} = 9{,}92$

Сравним значение $t_{эмп}$ с табличным значением $t_{кр} = t_{\gamma, k}$, где $\gamma = 1 - \alpha$ и $k = n - 2$ - число степеней свободы.

По таблице значений $t_{\gamma, k}$ - критерия Стьюдента при $\gamma = 1 - 0{,}05 = 0{,}95$ и $k = 50 - 2 = 48$ получаем $t_{\gamma, k} = t_{0{,}95, 48} = 2{,}02$.

Поскольку $t_{эмп} > t_{кр}$ $(9{,}92 > 2{,}02)$, то нулевая гипотеза отвергается, следовательно выборочный коэффициент корреляции значим на 5% уровне.

Для определения тесноты связи между переменными X и Y воспользуемся таблицей оценок коэффициента корреляции.

Поскольку $r = 0{,}82 \in [0{,}8; 0{,}9]$, то связь между переменными тесная.

Т.к. $r = 0{,}82 > 0$, то связь между переменными прямая, т.е. с увеличением стоимости ОПФ стоимость выпущенной продукции в среднем увеличивается.

Таблицы значений $t_{\gamma, k}$ – критерия Стьюдента и оценок коэффициента корреляции находится на сайте http://exponenta.ucoz.ru/II_kyrs/statisticheskie_tablicy.pdf

Чтобы оценить средний выпуск продукции предприятия, ОПФ которого составляют 81 млн.руб. воспользуемся уравнением регрессии y по x, т.е.

$$\bar{y}_x = 0,757x + 24,626.$$

Получаем $\bar{y}_{x=81} = 0,757 \cdot 81 + 24,626 = 85,943$ (млн. руб.)

Т.о., средний выпуск продукции предприятия, ОПФ которого составляют 81 млн.руб. в среднем составляет 85,943 млн. руб.