Тема: Парная корреляция и регрессия

Вариант 3

- По данным хозяйств 1-25 изучить зависимость между Валовым доходом отрасли растениеводства, приходящимся на 100 га пашни (тыс. руб.) и Затратами труда в растениеводстве на 100 га пашни, тыс. чел.-час./га.

- ***Задание:***

- По данным своего варианта необходимо:
- 1. Вычислить описательные статистики. Проверить характер распределения при-знаков. При необходимости удалить аномальные наблюдения.
- 2. С помощью метода наименьших квадратов найти параметры a и b:
- • линейной функции;
- • степенной функции;
- • равносторонней гиперболы.

3. Дать экономическую интерпретацию каждому уравнению регрессии исчислив средний коэффициент эластичности , парный линейный коэффициент корреляции – r (для линейной модели), и индекс корреляции ρ (для нелинейных функций), коэффициент детерминации – D.
4. Оценить каждую модель через среднюю ошибку аппроксимации и F-критерий Фишера и сделать вывод, какая из моделей лучше описывает изучаемую зависимость.
5. Провести статистическую оценку надежности параметров парной корреляции (с помощью t-статистики Стьюдента и путем расчета доверительного интервала каждого из показателей).
6. Выполнить прогноз значения результативного признака при прогнозном значении факторного, составляющем 125% от его среднего уровня
7. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал

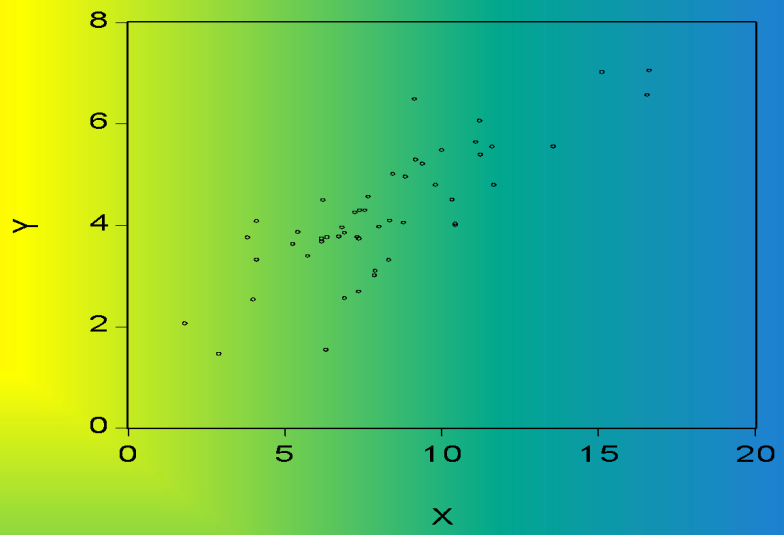
Построение уравнения регрессии

Постановка задачи

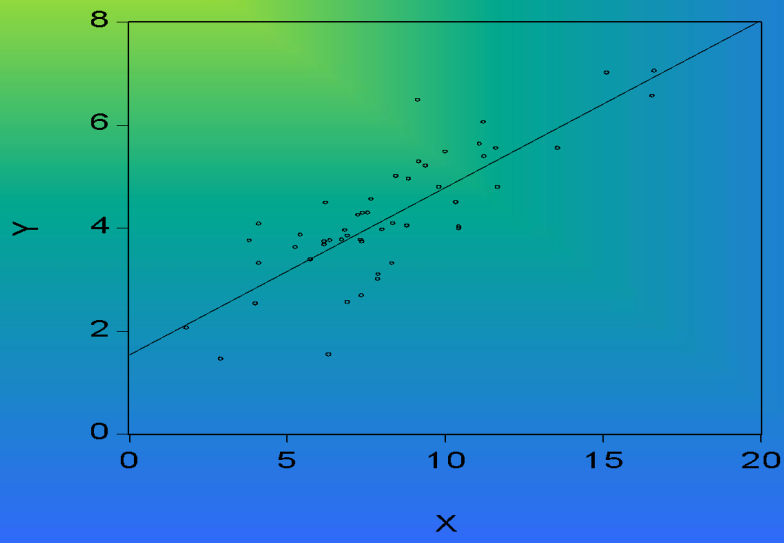
Данные наблюдений

	x	y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

Поле корреляции

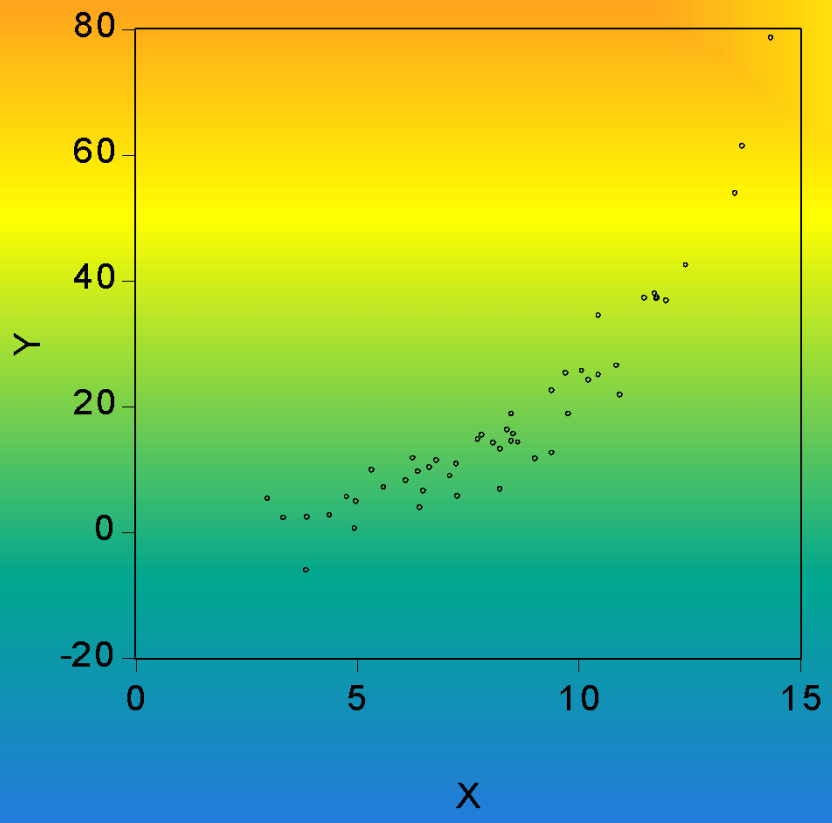


Зависимости $\hat{y} = f(x)$ соответствует некоторая кривая на плоскости. И по форме облака наблюдений можно определить вид регрессионной функции.



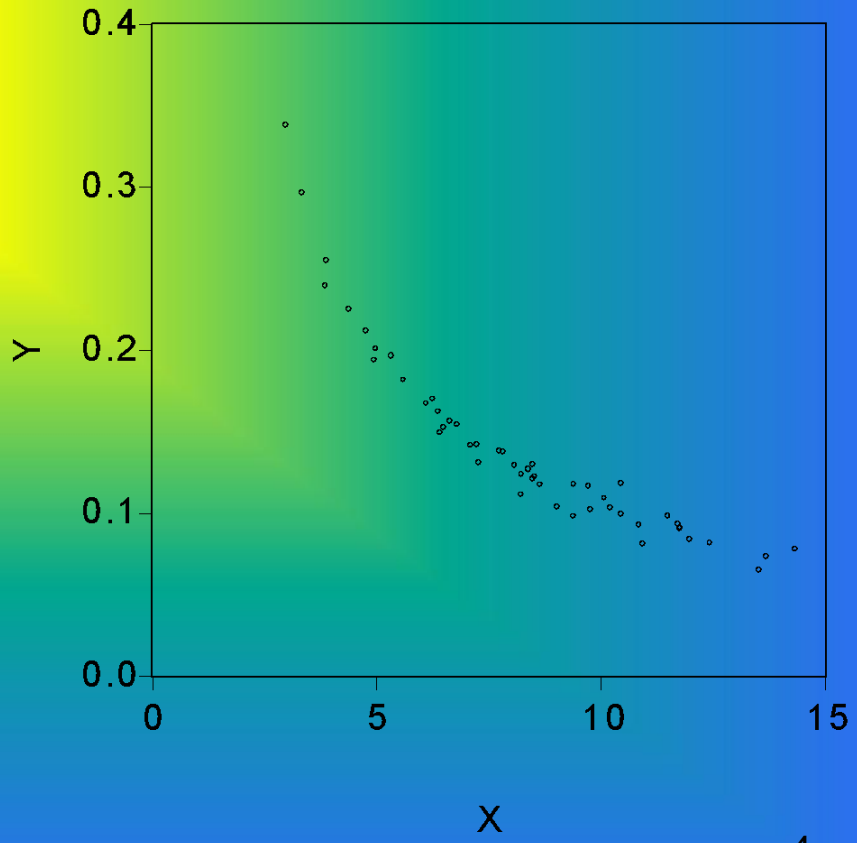
Степенная

$$Y = \alpha e^{\beta X}$$



Гиперболическая

$$Y = \alpha + \frac{\beta}{X}$$



2. Оценка параметров модели

Оценка параметров линейной парной регрессии – метод наименьших квадратов (МНК)

$$S = \sum (y_i - \hat{y}_i)^2 \to \min \quad \text{или} \quad \sum \varepsilon^2 \to \min$$

$$S = \sum (y_i - \hat{y}_i)^2 = \sum (y - a - bx)^2$$

$$S'_a = -2\sum y + 2na + 2b\sum x = 0$$

$$S'_b = -2\sum yx + 2a\sum x + 2b\sum x^2 = 0$$

Отсюда получаем систему уравнений:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx \end{cases}$$

Разделим оба уравнения на n:

$$\begin{cases} \frac{na}{n} + \frac{b\sum x}{n} = \frac{\sum y}{n}, \\ \frac{a\sum x}{n} + \frac{b\sum x^2}{n} = \frac{\sum yx}{n} \end{cases}$$

$$a = \frac{\sum y}{n} - \frac{b\sum x}{n} = \bar{y} - b\bar{x}$$

Подставляем во второе уравнение:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

Оценка параметров нелинейных моделей

Зависимость	Формула	Линеаризующее преобразование	Зависимость между параметрами
Гиперболическая	$y = a + \frac{b}{x}$	**y1=y** **X=1/x**	**a1=a** **b1=b**
Логарифмическая	$y = a + b \times \ln x$	**y1=y** **X=ln x**	**a1=a** **b1=b**
Экспоненциальная	$y = e^{a+bx}$	**Y=ln y** **x1=x**	**a1=a** **b1=b**
Степенная	$y = a \times x^{b}$	**Y=ln y (Y=lg y)** **X=ln x (X=lg x)**	**ln a=C (lg a=C)** **b1=b**
Показательная	$y = a \times b^{x}$	**Y=ln y (Y=lg y)** **x1=x**	**ln a=C (lg a=C)** **ln b=B (lg b=B)**

3. Проверка качества уравнения регрессии

H0: уравнение статистически не значимо

$yi = \hat{y}i + \varepsilon i$

$D(y) = D(\hat{y}) + D(\varepsilon)$

$$\frac{1}{n}\sum(y-\bar{y})^2 = \frac{1}{n}\sum(\hat{y}-\bar{y})^2 + \frac{1}{n}\sum(y-\hat{y})^2$$

полная (общая) сумма квадратов отклонений = **сумма квадратов отклонений, объясненная регрессией** + **(остаточная) сумма квадратов отклонений, не объясненная регрессией**

F-критерий Фишера:

$$F = \frac{\frac{D(\hat{y})}{k}}{\frac{D(\varepsilon)}{n-m-1}} \quad \textit{или} \quad \frac{R^2}{1-R^2} \times \frac{n-m-1}{m}$$

где m – число независимых переменных в уравнении регрессии (для парной регрессии m = 1);

n – число единиц совокупности.

Если **Fфакт > Fтабл**, то *H0* о случайной природе связи отклоняется и признается статистическая значимость и надежность уравнения.

Если **Fфакт < Fтабл**, то *H0* не отклоняется и признается статистическая незначимость уравнения регрессии.

t-критерий Стьюдента

H_0: a=0; b=0

Стандартные ошибки параметров регрессии и коэффициента корреляции:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S^2_{ост}}{\sum (x - \bar{x})^2}} = \frac{S_{ост}}{\sigma_x \sqrt{n}}$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2} \times \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S^2_{ост} \frac{\sum x^2}{n^2 \sigma^2_x}} = S_{ост} \frac{\sqrt{\sum x^2}}{n \sigma_x}$$

$$m_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}}$$

Оценка значимости параметров уравнения и коэффициента корреляции проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}$$

Если **tфакт > tтабл**, то H_0 отклоняется, т.е. a, b, r не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x.

Если **tфакт < tтабл**, то H_0 не отклоняется и признается случайная природа формирования a, b, r.

Доверительные интервалы – это пределы, в которых лежит точное значение определяемого показателя с заданной вероятностью.

Доверительные интервалы для параметров a и b уравнения линейной регрессии определяются соотношениями:

$$\gamma_a = a \pm t_{табл} \cdot m_a; \quad \gamma_{a_{\min}} = a - t_{табл} \cdot m_a \quad \gamma_{a_{\max}} = a + t_{табл} \cdot m_a$$

$$\gamma_b = b \pm t_{табл} \cdot m_b; \quad \gamma_{b_{\min}} = b - t_{табл} \cdot m_b \quad \gamma_{b_{\max}} = b + t_{табл} \cdot m_b$$

Точечный и интервальный прогноз по уравнению линейной регрессии

Точечный прогноз заключается в получении прогнозного значения *y*, которое определяется путем подстановки в уравнение регрессии соответствующего (прогнозного) значения *x*.

Интервальный прогноз заключается в построении доверительного интервала прогноза.

При построении доверительного интервала прогноза используется *стандартная ошибка прогноза*:

$$m_{\hat{y}_p} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Строится *доверительный интервал прогноза*:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm t_{табл} \cdot m_{\hat{y}_p}$$

Исходные данные

№	Валовой доход растениеводства, тыс. руб.;	Площадь пашни, га	Отработано за год в растениеводстве, тыс. чел.-час.;	Валовый доход растениеводства на 100 га пашни У	Затраты труда в раст-ве на 100 га пашни X
1	4709	21003	404	22,42	1,92
2	10585	6847	309	154,59	4,51
3	18740	19206	403	97,57	2,1
4	8938	4009	25	222,95	0,62
5	3543	3191	62	111,03	1,94
6	4001	3104	107	128,9	3,45
7	3756	3122	57	120,31	1,83
8	665	1306	16	50,92	1,23
9	3194	2838	79	112,54	2,78
10	3407	4852	31	70,22	0,64
11	1667	1790	30	93,13	1,68
12	1979	3053	78	64,82	2,55
13	2141	1987	47	107,75	2,37
14	3807	1803	74	211,15	4,1
15	2137	2790	302	76,59	10,82
16	18183	17489	559	103,97	3,2
17	5291	13813	801	38,3	5,8
18	5746	2883	98	199,31	3,4
19	3614	2601	89	138,95	3,42
20	8494	3412	144	248,94	4,22
21	11403	4277	405	266,61	9,47
22	2642	2497	70	105,81	2,8
23	4195	4759	154	88,15	3,24

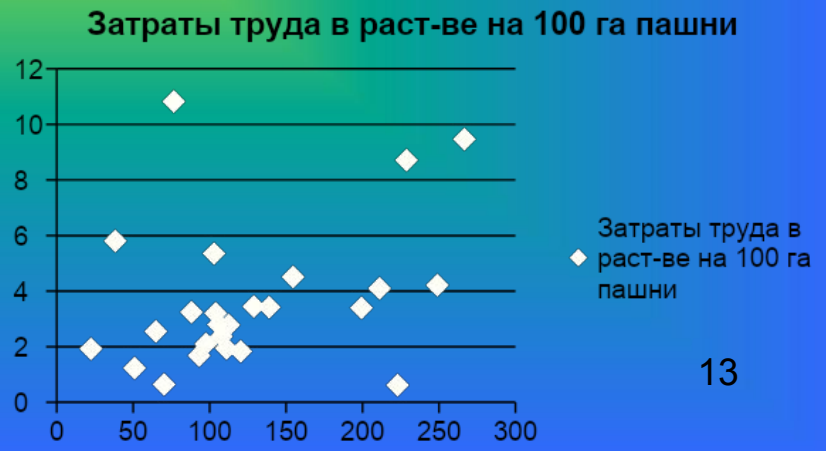
$$\sigma_x = 2,61$$

$$\sigma_y = 66,87$$

$$v_x = \frac{2,61}{3,67} = 0,71 \qquad v_y = \frac{66,87}{126,66} = 0,53$$

Поскольку коэффициенты вариации по каждому из признаков превышают значение 0,35, то можно сделать вывод о неоднородности совокупности.

Поле корреляции



- Исключим из совокупности не типичные явления, т.е. следующие хозяйства: 1, 2, 4, 8, 14, 15, 17, 18, 20, 21, 24, 25.

$$\sigma_x = 0,82$$

$$\sigma_y = 21,07$$

$$v_x = \frac{0,82}{2,46} = 0,33$$

$$v_y = \frac{21,07}{103,32} = 0,20$$

Поскольку коэффициенты вариации по каждому из признаков не превышают значения 0.35, то может сделать вывод об однородности изучаемой совокупности.

№	Валовый доход растениеводства на 100 га пашни У	Затраты труда в раст-ве на 100 га пашни Х
3	97,57	2,1
5	111,03	1,94
6	128,9	3,45
7	120,31	1,83
9	112,54	2,78
10	70,22	0,64
11	93,13	1,68
12	64,82	2,55
13	107,75	2,37
16	103,97	3,2
19	138,95	3,42
22	105,81	2,8
23	88,15	3,24

Валовый доход растениеводства на 100 га пашни Y		*Затраты труда в раст-ве на 100 га пашни X*	
Среднее	103,3192308	Среднее	2,461538462
Стандартная ошибка	5,844173568	Стандартная ошибка	0,226590804
Медиана	105,81	Медиана	2,55
Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонение	21,07146746	Стандартное отклонение	0,816984763
Дисперсия выборки	444,006741	Дисперсия выборки	0,667464103
Эксцесс	-0,013177306	Эксцесс	0,442390022
Асимметричность	-0,314102043	Асимметричность	-0,753212066
Интервал	74,13	Интервал	2,81
Минимум	64,82	Минимум	0,64
Максимум	138,95	Максимум	3,45
Сумма	1343,15	Сумма	32
Счет	13	Счет	13

Исследуя полученные показатели описательной статистики, мы наблюдаем: По факторному признаку наблюдается незначительная левосторонняя асимметрия и незначительный плосковершинный эксцесс. По результативному признаку наблюдается незначительная левосторонняя асимметрия и незначительный островершинный эксцесс. Так как значения не превышают критические, то распределение совокупности можно считать близким к нормальному.

$$b = \frac{262,3 - 2,46 * 103,32}{0,82^2} = 12,13$$

$$a = 103,32 - 12,13 * 2,46 = 73,48$$

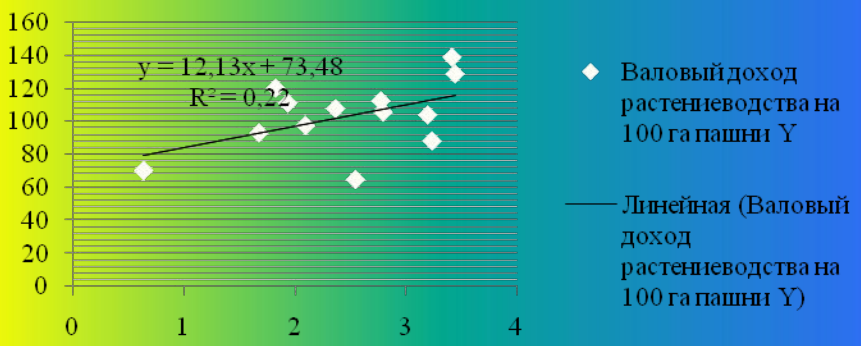
$$\overline{Э} = 12,13 * \frac{2,46}{103,32} = 0,29$$

$$r = b \cdot \frac{\sigma_x}{\sigma_y}; \qquad r = 12,13 * \frac{0,82}{21,07} = 0,47$$

$$D = r^2 * 100\% = 0,47^2 * 100\% = 22,09\%$$

$$\overline{A} = \frac{1}{n} \cdot \sum \left| \frac{Y - \widetilde{Y}}{Y} \right| \cdot 100\% = \frac{1}{n} \cdot \Sigma A_i \quad \frac{1}{13} * 187,23 = 14,4$$

$$t_b = \frac{12,13}{6,73} = 1,8 \qquad t_a = \frac{73,48}{16,66} = 4,41 \qquad t_r = \frac{0,47}{0,26} = 1,81$$



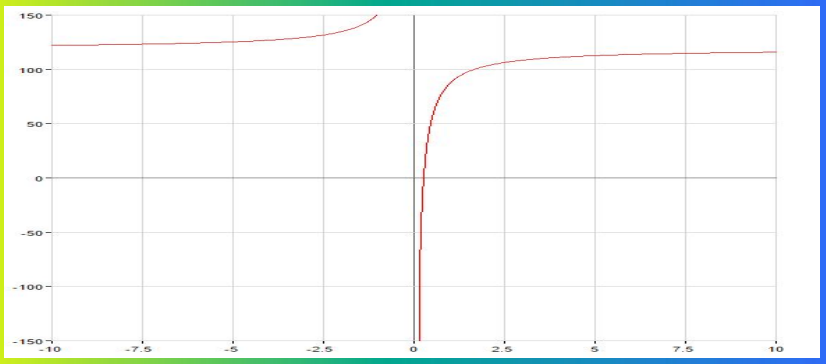
36,81<*a*<110,15 -2,68<*b*<26,94 -0,1 <*r*<1,04

Уравнение парной линейной регрессии является статистически незначимым

• Равносторонняя гипербола.

№ хозяйства	Z	Y	Z_i^2	Y_i^2	$Z_i \cdot Y_i$
1	0,292398	138,95	0,085496	19307,1	40,62865
2	0,289855	128,9	0,084016	16615,21	37,36232
3	0,546448	120,31	0,298606	14474,5	65,74317
4	0,359712	112,54	0,129393	12665,25	40,48201
5	0,515464	111,03	0,265703	12327,66	57,23196
6	0,421941	107,75	0,178034	11610,06	45,46414
7	0,357143	105,81	0,127551	11195,76	37,78929
8	0,3125	103,97	0,097656	10809,76	32,49063
9	0,47619	97,57	0,226757	9519,905	46,4619
10	0,595238	93,13	0,354308	8673,197	55,43452
11	0,308642	88,15	0,09526	7770,423	27,20679
12	1,5625	70,22	2,441406	4930,848	109,7188
13	0,392157	64,82	0,153787	4201,632	25,41961
	6,430188	1343,15	4,537974	144101,3	621,4337
сред.знач.	0,494			11084,72	47,8026

$$y = 118{,}32 - \frac{30{,}32}{x}$$



$$\overline{Э} = \frac{30{,}32}{118{,}32 * 2{,}46 - 30{,}32} = 0{,}12$$

$$\rho = \sqrt{1 - \frac{3972{,}777}{5328{,}081}} = 0{,}5$$

$$D = 0{,}5^2 * 100\% = 25$$

Индекс корреляции показывает, что связь между среднегодовым заработком 1 работника сельскохозяйственного предприятия и валовой продукцией на 100 га сельскохозяйственных угодий сильная.

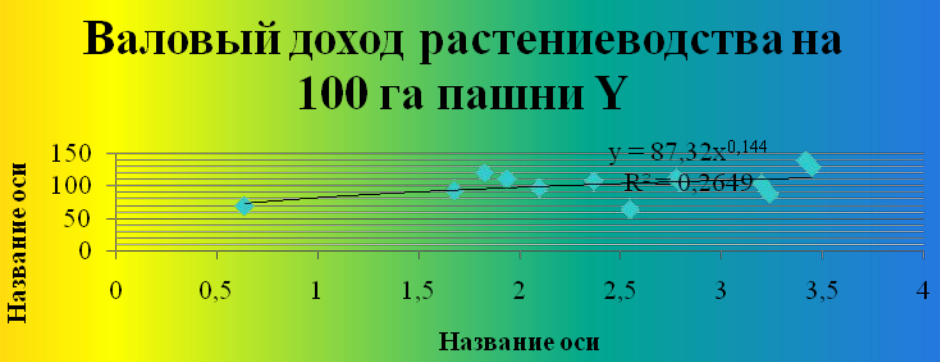
Средняя ошибка аппроксимации равна 13,47%, т.е. в среднем расчетные значения валового дохода на 100 га пашни , отличаются от фактических на 13,47%, что не входит в допустимый предел.

$$t_b = \frac{-30{,}32}{16{,}13} = -1{,}88 \qquad t_a = \frac{118{,}32}{9{,}44} = 12{,}53 \qquad t_r = \frac{0{,}5}{0{,}26} = 1{,}92$$

H_0 о значимости коэффициентов корреляции и регрессии подтверждается

№хозяйства	L	P	L_i^2	P_i^2	$L_i \cdot P_i$
1	1,23	4,93	1,51	24,35	6,07
2	1,24	4,86	1,53	23,61	6,02
3	0,6	4,79	0,37	22,94	2,89
4	1,02	4,72	1,05	22,31	4,83
5	0,66	4,71	0,44	22,18	3,12
6	0,86	4,68	0,74	21,9	4,04
7	1,03	4,66	1,06	21,73	4,8
8	1,16	4,64	1,35	21,57	5,4
9	0,74	4,58	0,55	20,98	3,4
10	0,52	4,53	0,27	20,56	2,35
11	1,18	4,48	1,38	20,06	5,27
12	-0,45	4,25	0,2	18,08	-1,9
13	0,94	4,17	0,88	17,4	3,91
$\sum$	10,74	60,02	11,33	277,67	50,19
сред. знач.	0,83	4,62	0,87	21,36	3,86

$y=87,32+ x^{0,144}$



$$\rho = \sqrt{1 - \frac{4490,331}{5328,081}} = 0,4$$

$$D = 0,4^2 * 100\% = 16\%$$

$$t_b = \frac{0,144}{0,1} = 1,44 \qquad t_a = \frac{4,5}{0,12} = 37,5 \qquad t_r = \frac{0,4}{0,28} = 1,43$$

$4,24<a<4,76$ $-0,08<b<0,36$ $-0,22<r<1,02$

Уравнение парной нелинейной гиперболической регрессии является статистически незначимым.

Интервальный прогноз

Ввиду того, что все три уравнения регрессии являются статистически незначимыми и ненадежными, рассчитать прогнозируемое значение ни по одному из рассмотренных уравнений не имеет смысла, поскольку данный прогноз не даст достоверного результата.

Тем не менее, для закрепления методики расчета прогнозов, выполним расчет прогнозного значения результата по линейной модели.

По условию задачи прогнозное значение фактора составляет 125% от x ср .

x= 3,69*1,25=4,61

И прогнозное значение при этом составит: y=73,48+12,13*4,61=129,4

- Найдем ошибку прогно

$$m_{ур} = \sigma_{ост} * \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} \quad m_{ур} = 19{,}05 * \sqrt{1 + \frac{1}{13} + \frac{(4{,}61 - 3{,}69)^2}{8{,}0096}} = 20{,}78$$

- Далее строиться доверительный интервал прогноза при уровне значимости

$$\gamma_{ур} = y_p \pm \Delta_{ур}$$

- Предельная ошибка прогноза, которая в 95% случаев не будет превышена, составит:

$$\Delta_{ур} = t_{табл} * m_{ур} = 2{,}201 * 20{,}78 = 45{,}74$$

- Доверительный интервал прогноза:
- (83.66;175,14)