Тема: Регрессия

ЗАДАНИЕ. По некоторым территориям районов края известны значения средней суточного душевого дохода в у.е. (фактор X) и процент от общего дохода, расходуемого на покупку продовольственных товаров (фактор Y), табл. 1. Требуется для характеристики зависимости У от X рассчитать параметры линейной, степенной, показательной функции и выбрать оптимальную модель (провести оценку моделей через среднюю ошибку аппроксимации (А) и F- критерий Фишера.

Таблица 1

Район	фактор Y	фактор X
Пожарский (1)	68,8	45,1
Кавалеровский (2)	61,2	59,0
Дальнегорский (3)	59,9	57,2
Хасанский (4)	56,7	61,8
Лесозаводский (5)	55,0	58,8
Хорольский (6)	54,3	47,2
Анучинский (7)	49,3	55,2

РЕШЕНИЕ.

1а. Для расчета параметров a и b линейной регрессии $y = a + b \cdot x$ решаем систему нормальных уравнений относительно a и b:

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

По исходным данным рассчитываем $\sum a$, $\sum x$, $\sum yx$, $\sum x^2$, $\sum y^2$:

Таблица 2

	y	x	yx	x^2	y^2	y_x	$y - y_x$	A_i
1	68,8	45,1	3102,88	2034,01	4733,44	61,3	7,5	10.9
2	61,2	59,0	3610,80	3481,00	3745,44	56,5	4,7	7,7
3	59,9	57,2	3426,28	3271,84	3588,01	57,1	2,8	4,7
4	56,7	61,8	3504,06	3819,24	3214,89	55,5	1,2	2,1
5	55,0	58,8	3234,00	3457,44	3025,00	56,5	-1,5	2,7
6	54,3	47,2	2562,96	2227,84	2948,49	60,5	-6,2	11,4
7	49,3	55,2	2721,36	3047,04	2430,49	57,8	-8,5	17,2
Итого	405,2	384,3	22162,34	21338,41	23685,76	405,2	0,0	56,7
Ср. знач. (Итого/n)	57,89 $\overline{y}$	54,90 $\overline{x}$	3166,05 $\overline{yx}$	3048,34 $\overline{x^2}$	3383,68 $\overline{y^2}$	X	X	8,1
σ	5,74	5,86	X	X	X	X	X	X
σ^2	32,92	34,34	X	X	X	X	X	X

$$b = \frac{cov(x,y)}{\sigma_x^2} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{3166{,}05 - 57{,}89 \cdot 54{,}9}{34{,}34} \approx -0{,}35,$$

$$a = \bar{y} - b \cdot \bar{x} = 57{,}89 + 0{,}35 \cdot 54{,}9 \approx 76{,}88.$$

Уравнение регрессии: y = 76,88 – 0,35x. С увеличением среднедневной заработной платы на 1 руб. доля расходов на покупку продовольственных товаров снижается в среднем на 0,35 %-ных пункта.

Рассчитаем линейный коэффициент парной корреляции:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = -0{,}35 \cdot \frac{5{,}86}{5{,}74} = -0{,}357.$$

Связь умеренная, обратная.

Определим коэффициент детерминации:

$$r_{xy}^2 = (-0{,}35)^2 = 0{,}127.$$

Вариация результата на 12,7% объясняется вариацией фактора x. Подставляя в уравнение регрессии фактические значения x, определим теоретические (расчетные) значения $\bar{y}_x$. Найдем величину средней ошибки аппроксимации A:

$$A = \frac{1}{n}\sum A_i = \frac{1}{n}\sum \left|\frac{y - y}{y}\right| \cdot 100\% = \frac{56{,}7 \cdot 100\%}{7} = 8{,}1\%.$$

В среднем расчетные значения отклоняются от фактических на 8,1%. Рассчитаем F-критерий:

$$F_{факт} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2) = \frac{0{,}127}{0{,}873} \cdot 5 = 0{,}7.$$

$F_{табл} = 6{,}6 > F_{факт}$, при a = 0,05.

Полученное значение указывает на необходимость принять гипотезу H_0 о случайной природе выявленной зависимости и статистической незначимости параметров уравнения и показателя тесноты связи.

1б. Построению степенной модели $y = a \cdot x^b$ предшествует процедура линеаризации переменных. В примере линеаризация производится путем логарифмирования обеих частей уравнения:

$$lg\, y = lg\, a + b \cdot lg\, x,$$

$$Y = C + b \cdot X,$$

где Y = lg(y), X = lg(x), C = lg(a).

Для расчетов используем данные табл. 3.

Таблица 3

	Y	X	YX	Y^2	X^2	y_x	$y-y_x$	$(y-y_x)^2$	A_i
1	1,8376	1,6542	3,0398	3,3768	2,7364	61,0	7,8	60,8	11,3
2	1,7868	1,7709	3,1642	3,1927	3,1361	56,3	4,9	24,0	8,0
3	1,7774	1,7574	3,1236	3,1592	3,0885	56,8	3,1	9,6	5,2
4	1,7536	1,7910	3,1407	3,0751	3,2077	55,5	1,2	1,4	2,1
5	1,7404	1,7694	3,0795	3,0290	3,1308	56,3	-1,3	1,7	2,4
6	1,7348	1,6739	2,9039	3,0095	2,8019	60,2	-5,9	34,8	10,9
7	1,6928	1,7419	2,9487	2,8656	3,0342	57,4	-8,1	65,6	16,4
Итого	12,3234	12,1587	21,4003	21,7078	21,1355	403,5	1,7	197,9	56,3
Среднее значение	1,7605	1,7370	3,0572	3,1011	3,0194	X	X	28,27	8,0
σ	0,0425	0,0484	X	X	X	X	X	X	X
σ^2	0,0018	0,0023	X	X	X	X	X	X	X

Рассчитаем C и b:

$$b = \frac{\overline{Y \cdot X} - \overline{Y} \cdot \overline{X}}{\sigma_X^2} = \frac{3{,}0572 - 1{,}7605 \cdot 1{,}7370}{0{,}0484^2} \approx -0{,}298,$$

$$C = \overline{Y} - b \cdot \overline{X} = 1{,}7605 + 0{,}298 \cdot 1{,}7370 \approx 2{,}278.$$

Получим линейное уравнение: $Y_x = 2{,}278 - 0{,}298 \cdot X$.

Выполнив его потенцирование, получим:

$$y_x = 10^{2{,}278} \cdot x^{-0{,}298} = 189{,}7 \cdot x^{-0{,}298}.$$

Подставляя в данное уравнение фактические значения x, получаем теоретические значения результата y_x. По ним рассчитаем показатели: тесноты связи - индекс корреляции ρ_{xy} и среднюю ошибку аппроксимации $\overline{A_i}$:

$$\rho_{xy} = \sqrt{1 - \frac{\sum\left(y - y_x\right)^2}{\sum\left(y - \overline{y}\right)^2}} = \sqrt{1 - \frac{28{,}27}{32{,}92}} = 0{,}3758,$$

$\overline{A} = 8{,}0\%$.

Характеристики степенной модели указывают, что она несколько лучше линейной функции описывает взаимосвязь.

1в. Построению уравнения показательной кривой $y = a \cdot b^x$ предшествует процедура линеаризации переменных при логарифмировании обеих частей уравнения:

$lg\, y = lg\, a + x \cdot lg\, b$,

$Y = C + B \cdot x$.

Для расчетов используем данные табл. 4.

Таблица 4

	Y	x	Yx	Y^2	x^2	y_x	$y - y_x$	$(y - y_x)^2$	A_i
1	1,8376	45,1	82,8758	3,3768	2034,01	60,7	8,1	65,61	11,8
2	1,7868	59,0	105,4212	3,1927	3481,00	56,4	4,8	23,04	7,8
3	1,7774	57,2	101,6673	3,1592	3271,84	56,9	3,0	9,00	5,0
4	1,7536	61,8	108,3725	3,0751	3819,24	55,5	1,2	1,44	2,1
5	1,7404	58,8	102,3355	3,0290	3457,44	56,4	-1,4	1,96	2,5
6	1,7348	47,2	81,8826	3,0095	2227,84	60,0	-5,7	32,49	10,5
7	1,6928	55,2	93,4426	2,8656	3047,04	57,5	-8,2	67,24	16,6
Итого	12,3234	384,3	675,9974	21,7078	21338,41	403,4	-1,8	200,78	56,3
Ср. зн.	1,7605	54,9	96,5711	3,1011	3048,34	X	X	28,68	8,0
	0,0425	5,86	X	X	X	X	X	X	X
	0,0018	34,339	X	X	X	X	X	X	X

Значения параметров регрессии A и B составили:

$$B = \frac{\overline{Y \cdot x} - \overline{Y} \cdot \overline{x}}{\sigma_x^2} = \frac{96{,}5711 - 1{,}7605 \cdot 54{,}9}{5{,}86^2} \approx -0{,}0023,$$

$A = \overline{Y} - B \cdot \overline{x} = 1{,}7605 + 0{,}0023 \cdot 54{,}9 = 1{,}887$.

Получено линейное уравнение: $Y_x = 1{,}887 - 0{,}0023 \cdot x$. Произведем потенцирование полученного уравнения и запишем его в обычной форме:

$y_x = 10^{1{,}887} \cdot 10^{-0{,}0023} = 77{,}1 \cdot 0{,}9947^x$.

Тесноту связи оценим через индекс корреляции ρ_{xy}:

$$\rho_{xy} = \sqrt{1 - \frac{\sum\left(y - \hat{y}\right)^2}{\sum\left(y - \overline{y}\right)^2}} = \sqrt{1 - \frac{28{,}68}{32{,}92}} = 0{,}3589,$$

$\overline{A} = 8{,}0\%$.

Связь умеренная.

$\overline{A} = 8{,}0\%$, что говорит о повышенной ошибке аппроксимации, но в допустимых пределах. Показательная функция чуть хуже, чем степенная, описывает изучаемую зависимость.

1г. Уравнение равносторонней гиперболы $y = a + b \cdot \frac{1}{x}$ линеаризуется при замене: $z = \frac{1}{x}$. Тогда $y = a + b \cdot z$. Для расчетов используем данные табл. 5.

Таблица 5

	y	z	yz	z^2	y^2	y_x	$y - y_x$	$(y - y_x)^2$	A_i
1	68,8	0,0222	1,5255	0,000492	4733,44	61,8	7,0	49,00	10,2
2	61,2	0,0169	1,0373	0,000287	3745,44	56,3	4,9	24,01	8,0
3	59,9	0,0175	1,0472	0,000306	3588,01	5,9	3,0	9,00	5,0
4	56,7	0,0162	0,9175	0,000262	3214,89	55,5	1,2	1,44	2,1
5	55	0,0170	0,9354	0,000289	3025,00	56,4	-1,4	1,96	2,5
6	54,3	0,0212	1,1504	0,000449	2948,49	60,8	-6,5	42,25	12,0
7	49,3	0,0181	0,8931	0,000328	2430,49	57,5	-8,2	67,24	16,6
Итого	405,2	0,1291	7,5064	0,002413	23685,76	405,2	0,0	194,90	56,5
Среднее значение	57,9	0,0184	1,0723	0,000345	3383,68	X	X	27,84	8,1
	5,74	0,002145	X	X	X	X	X	X	X
	32,9476	0,000005	X	X	X	X	X	X	X

Значения параметров регрессии a и b составили:

$$b = \frac{\overline{y \cdot z} - \overline{y} \cdot \overline{z}}{\sigma_x^2} = \frac{1{,}0723 - 57{,}9 \cdot 0{,}0184}{0{,}002145^2} \approx 1051{,}4,$$

$$a = \overline{y} - b \cdot \overline{z} = 57{,}89 - 1051{,}4 \cdot 0{,}0184 = 38{,}5.$$

Получено уравнение: $y_x = 38{,}5 + 1051{,}4\frac{1}{x}$.

Индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sum\left(y - y_x\right)^2}{\sum\left(y - \overline{y}\right)^2}} = \sqrt{1 - \frac{27{,}84}{32{,}92}} = 0{,}3944,$$

$\overline{A} = 8{,}1\%$.

По уравнению равносторонней гиперболы получена наибольшая оценка тесноты связи: p_{xy} = 0,3944 (по сравнению с линейной, степенной и показательной регрессиями). $\overline{A}$ остается на допустимом уровне: 8,1%.

$$F_{факт} = \frac{\rho_{xy}^2}{1-\rho_{xy}^2} \cdot \frac{m-n-1}{n} = \frac{0{,}1555}{0{,}8445} \cdot 5 = 0{,}92$$

,

где $F_{табл} = 6{,}6 > F_{факт}$, при a = 0,05.

Следовательно, принимается гипотеза H_0 о статистически незначимых параметрах этого уравнения. Этот результат можно объяснить сравнительно невысокой теснотой выявленной зависимости и небольшим числом наблюдений.