Регрессионный анализ

Регрессия ***(с лат. «движение назад»)*** – функция, позволяющая по величине одного признака *X* находить среднее (ожидаемое) значение другого признака У, связанного с *X* корреляционно.

Регрессионный анализ — вид анализа, позволяющий выявить количественную зависимость результативного признака от одного или нескольких факторных признаков

Задача регрессионного анализа заключается в установлении формы зависимости, в определении уравнения регрессии, использование уравнения для оценки неизвестных значений зависимой переменной

Классификация регрессии:

1. *по количеству факторов:*
   - ***однофакторная*** (парная);
   - ***многофакторная*** (множественная).
2. *по аналитической форме* (для парной):
   - ***линейная***: $\tilde{y}_x = a_0 + a_1 x$;
   - ***параболическая***: $\tilde{y}_x = a_0 + a_1 x + a_2 x^2$;
   - ***гиперболическая***: $\tilde{y}_x = a_0 + a_1 / x$;
   - ***показательная***: $\tilde{y}_x = a_0 \cdot a_1^x$;
   - ***степенная***: $\tilde{y}_x = a_0 \cdot x^{a1}$;
   - ***логарифмическая***: $\tilde{y}_x = a_0 + a_1 \lg x$ и т.д.

Интерпретация параметров уравнения регрессии:

- параметр a_0 показывает усредненное влияние на результативный признак неучтенных факторов (невыделенных для исследования);
- параметр a_1 (а в уравнении параболы и a_2) – коэффициент регрессии показывает, насколько изменяется в среднем значение результативного признака при изменении факторного на единицу его собственного измерения; коэффициент регрессии применяется также для определения коэффициента эластичности:

$$Э_x = a_1 \cdot \frac{\bar{x}}{\bar{y}},$$

где $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$, который показывает, на сколько % изменится величина результативного признака при изменении факторного признака на 1%.

Определение. ***Метод наименьших квадратов (МНК)*** – это метод оценки параметров уравнения регрессии a_0, a_1, a_2,…, в основе которого лежит предположение о независимости наблюдений исследуемой совокупности. Параметры модели (a_0, a_1, a_2, …) подбираются так, чтобы сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по уравнению регрессии, была минимальной, а именно:

$$S = \sum (y - y_x)^2 \to \min.$$

Замечание. Система нормальных уравнений для нахождения параметров линейной парной регрессии $\tilde{Y}_x = a_0 + a_1 x$ методом наименьших квадратов имеет следующий вид:

$$\begin{cases} na_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases},$$

где n – объем исследуемой совокупности (число единиц наблюдения).

- построение ***корреляционной таблицы*** (Для выявления связи между признаками по достаточно большому числу наблюдений используется корреляционная таблица. В корреляционной таблице можно отобразить только парную связь. Для составления корреляционной таблицы данные необходимо предварительно сгруппировать по обоим признакам (x и y), затем построить таблицу, по строкам в которой отложить группы результативного признака, а по столбцам – группы факторного. Корреляционная таблица дает общее представление о направлении связи. Если оба признака (x и y) располагаются в возрастающем порядке, а частоты (n_{xy}) сосредоточены по диагонали из левого верхнего угла в правый нижний, то можно судить о прямой связи между признаками; в противном случае – об обратной);
- ***дисперсионный анализ*** (На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет p уровней $F_1, F_2, \ldots, F_p$ на изучаемую величину X. Основная идея метода состоит в сравнении «факторной дисперсии», порождаемой воздействием фактора, и «остаточной дисперсии», обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на X).

Пример

Условие. Имеются следующие выборочные данные по предприятиям одной из отраслей промышленности в отчётном году:

№	Выпуск продукции, тыс.ед. (*У*)	Затраты на производство, млн.руб. (*Х*)
1	160	18,24
2	140	17,08
3	105	13,44
4	150	17,85
5	158	18,17
6	170	19,21
7	152	17,936
8	178	19,58
9	180	19,44
10	164	18,86
11	151	17,818
12	142	17,04
13	120	15
14	100	13
15	176	19,36
16	148	17,612
17	110	13,97
18	146	17,666
19	155	17,98
20	169	19,266
21	156	17,94
22	135	16,335
23	122	15,25
24	130	15,86
25	200	21
26	125	15,25
27	152	17,784
28	173	19,03
29	115	14,49
30	190	19,95

Необходимо:

- найти линейное уравнение регрессии, оценив его коэффициенты методом МНК;
- дать интерпретацию коэффициентов найденного уравнения;
- на одной координатной плоскости построить точечный график, соответствующий исходным данным, и найденную в прямую;
- вычислить коэффициент эластичности;
- сделать соответствующие выводы.

Линейное уравнение регрессии: $\tilde{y}_x = a_0 + a_1 x$.

Система нормальных уравнений:

$$\begin{cases} na_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases}$$

	X	y	X^2	Y^2	XY
	18,24	160	332,6976	25600	2918,4
	17,08	140	291,7264	19600	2391,2
	13,44	105	180,6336	11025	1411,2
	17,85	150	318,6225	22500	2677,5
	18,17	158	330,1489	24964	2870,86
	19,21	170	369,0241	28900	3265,7
	17,936	152	321,7001	23104	2726,272
	19,58	178	383,3764	31684	3485,24
	19,44	180	377,9136	32400	3499,2
	18,86	164	355,6996	26896	3093,04
	17,818	151	317,4811	22801	2690,518
	17,04	142	290,3616	20164	2419,68
	15	120	225	14400	1800
	13	100	169	10000	1300
	19,36	176	374,8096	30976	3407,36
	17,612	148	310,1825	21904	2606,576
	13,97	110	195,1609	12100	1536,7
	17,666	146	312,0876	21316	2579,236
	17,98	155	323,2804	24025	2786,9
	19,266	169	371,1788	28561	3255,954
	17,94	156	321,8436	24336	2798,64
	16,335	135	266,8322	18225	2205,225
	15,25	122	232,5625	14884	1860,5
	15,86	130	251,5396	16900	2061,8
	21	200	441	40000	4200
	15,25	125	232,5625	15625	1906,25
	17,784	152	316,2707	23104	2703,168
	19,03	173	362,1409	29929	3292,19
	14,49	115	209,9601	13225	1666,35
	19,95	190	398,0025	36100	3790,5
Σ	521,407	4472	9182,8	685248	79206,16

Система нормальных уравнений (с учётом нужных сумм):

$$\begin{cases} 30a_0 + 521{,}407a_1 = 4472 \\ 521{,}407a_0 + 9182{,}8a_1 = 79206{,}16 \end{cases}$$

Решим систему методом Крамера:

$$\Delta = \begin{vmatrix} 30 & 521{,}407 \\ 521{,}407 & 9182{,}8 \end{vmatrix} = 3618{,}74035,\ \Delta a_0 = \begin{vmatrix} 4472 & 3521{,}407 \\ 79206{,}16 & 9182{,}8 \end{vmatrix} = -233164{,}67,$$

$$\Delta a_1 = \begin{vmatrix} 4472 & 521{,}407 \\ 79206{,}16 & 9182{,}8 \end{vmatrix} = 44452{,}696.$$

Тогда $a_0 = \frac{\Delta a_0}{\Delta} \approx -64{,}4$ и $a_1 = \frac{\Delta a_1}{\Delta} \approx 12{,}3$.

Уравнение регрессии: $\tilde{y}_x = -64{,}4 + 12{,}3x$.

Интерпретация коэффициентов регрессии:

- коэффициент регрессии $a_1 \approx 12{,}3$, означает, что в среднем по изучаемой совокупности отклонение затрат на производство от средней величины на 1 млн. руб. приводило к отклонению с тем же знаком среднего выпуска продукции на 12,3 тыс. ед. При нестрогой интерпретации говорят: «С увеличением затрат на производство на 1 млн. руб. в среднем выпуск продукции возрастает на 12,3 тыс.ед.»;
- свободный член уравнения регрессии $a_0 \approx -64{,}4$ показывает усреднённое влияние неучтённых в модели факторов; отрицательная величина свободного члена уравнения означает, что область существования признака *У* не включает нулевого значения признака *Х* и близких к нулю значений. Можно рассчитать минимально возможную величину фактора *Х*, при которой обеспечивается наименьшее значение признака *У* (разумеется, положительное):

$$X_{min} = a_0 / a_1 = 64{,}4 / 12{,}3 = 5{,}2 \text{ млн. руб.}$$

Это наименьшая сумма затрат на производство, при которой предприятие способно выпускать продукцию.

Графическое изображение (рис. 8).

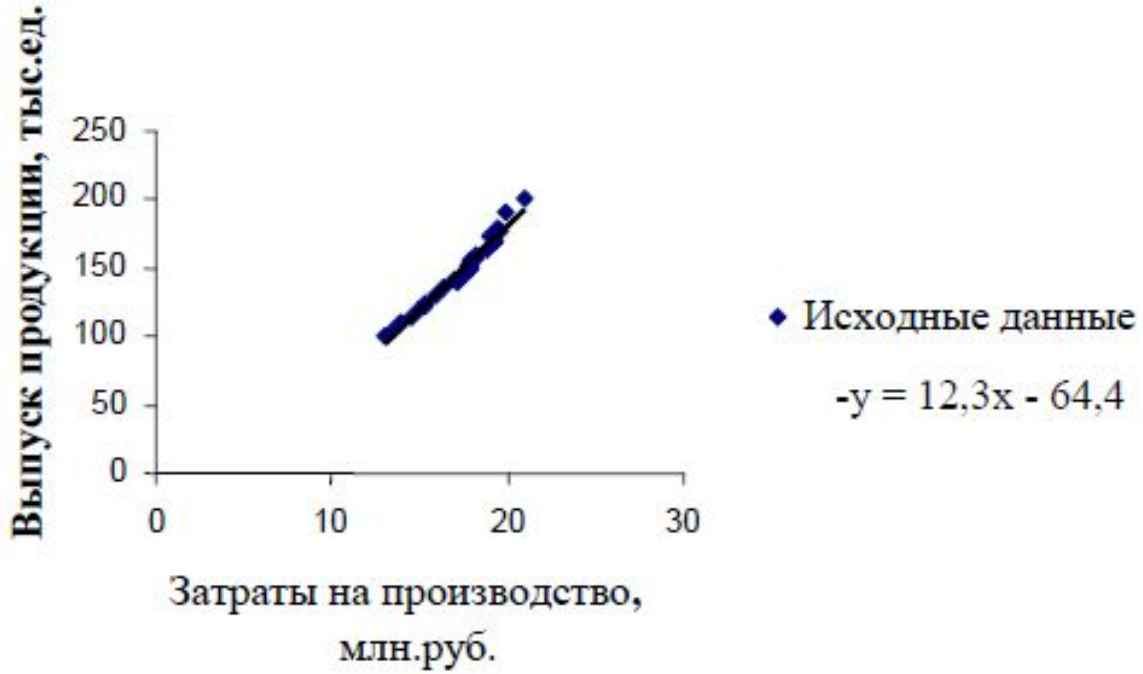

Рис. 8: Поле корреляции и уравнение регрессии.

Коэффициент эластичности:

Т.к. $\bar{x} = \frac{521{,}407}{20} \approx 26{,}07$, $\bar{y} = \frac{4472}{20} = 223{,}6$, то

$$Э_x = 12{,}3 \cdot \frac{26{,}07}{223{,}6} \approx 1{,}43.$$

При росте затрат на производство на 1%, выпуск продукции предприятия возрастет на 1,43%.

Вывод:

- с увеличением затрат на производство на 1 млн. руб. в среднем выпуск продукции возрастает на 12,3 тыс.ед.;
- при росте затрат на производство на 1%, выпуск продукции предприятия возрастет на 1,43%;
- наименьшая сумма затрат на производство, при которой предприятие способно выпускать продукцию составляет 5,2 млн.руб.