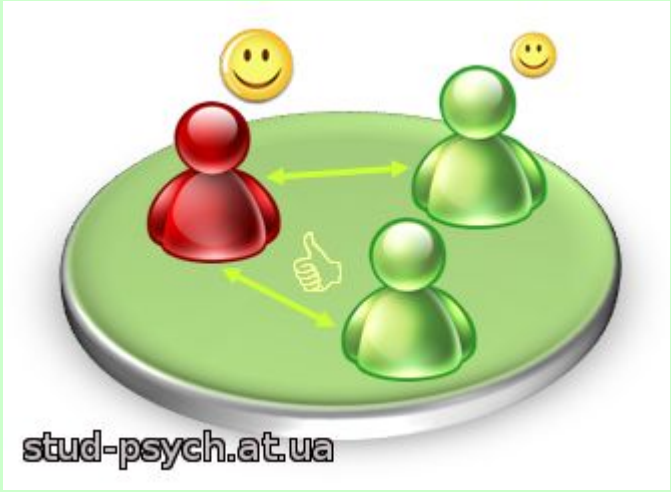
СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ КОРЕЛЯЦІЙНИХ ЗВ’ЯЗКІВ

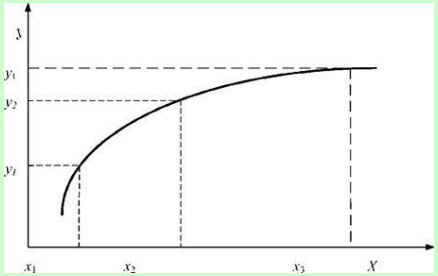
Соціально-економічні явища взаємозв'язані та взаємозумовлені і зв'язок (залежність) між ними носить причинно-наслідковий характер.

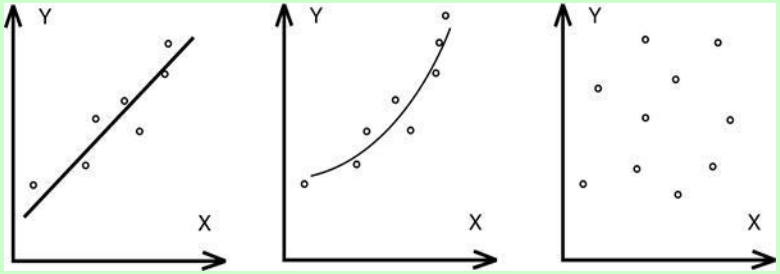
Фактор - причини і умови, що характеризують закономірності зв'язку. Ознаки, що є причинами та умовами зв'язку, називаються ***факторними (х),*** а ті, що змінюються під впливом факторних ознак, – ***результативними (у)***.



Види зв'язку між ознаками явищ

Функціональний зв'язок - між факторною та результативною ознаками кожному значенню ознаки *x* відповідає одне чітко визначене значення ознаки *y.*





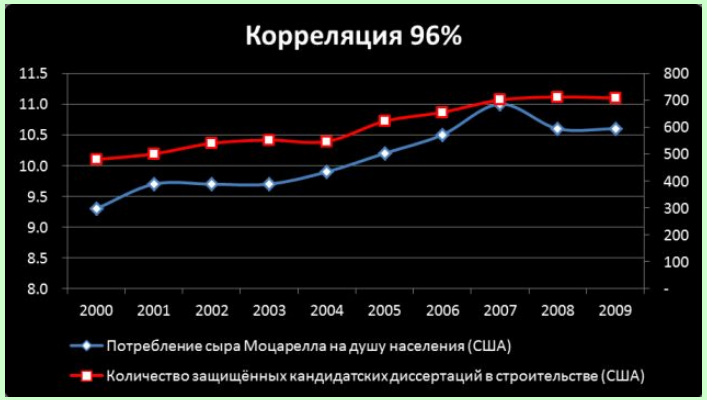
Стохастичний зв'язок - кожному окремому значенню факторної ознаки *x* відповідає певна множина значень результативної ознаки *y.* Такий зв'язок утворює *умовний розподіл* ознак, який варіює.

Зв'язки такого виду називають ще *статистичними, ймовірними.*

Теорія кореляції

Кореляція - термін із природознавства, означає співвідношення, відповідність між змінними у рівнянні регресії. Основоположниками цієї теорії є англійські вчені-біологи Ф. Гамільтон (1822 – 1911 рр.), К. Пірсон (1857 – 1936 рр.).

Між ознаками *x* та *y* існує кореляційна залежність, коли **середня величина однієї з них змінюється в залежності від значення іншої.**

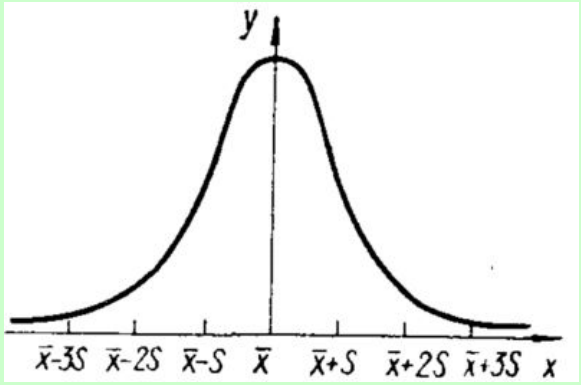


Умови використання теорії кореляції

а) наявність ***однорідності*** тих одиниць, які підлягають дослідженню (наприклад, відбір підприємств, які випускають однотипну продукцію, мають однаковий характер технології і тип обладнання тощо);

б) достатньо ***велика кількість спостережень,*** при яких погашається вплив випадковостей на результативну ознаку і має силу закон великих чисел;

в) ***нормальний характер*** *розподілу* результативної ознаки, на якому побудовані всі положення теорії кореляції.



Кореляційно-регресійний аналіз

КРА полягає у виборі виду рівняння регресії, обчисленні його параметрів та встановленні адекватності (відповідності) теоретичної залежності фактичним даним.

Якщо змінна *y* залежить від однієї змінної, то рівняння регресії є найпростішим і називається ***рівняння парної регресії*.**

Якщо *y* залежить від більш ніж однієї незалежної змінної, то така залежність має назву рівняння ***множинної або багатофакторної регресії***

Види рівнянь регресії

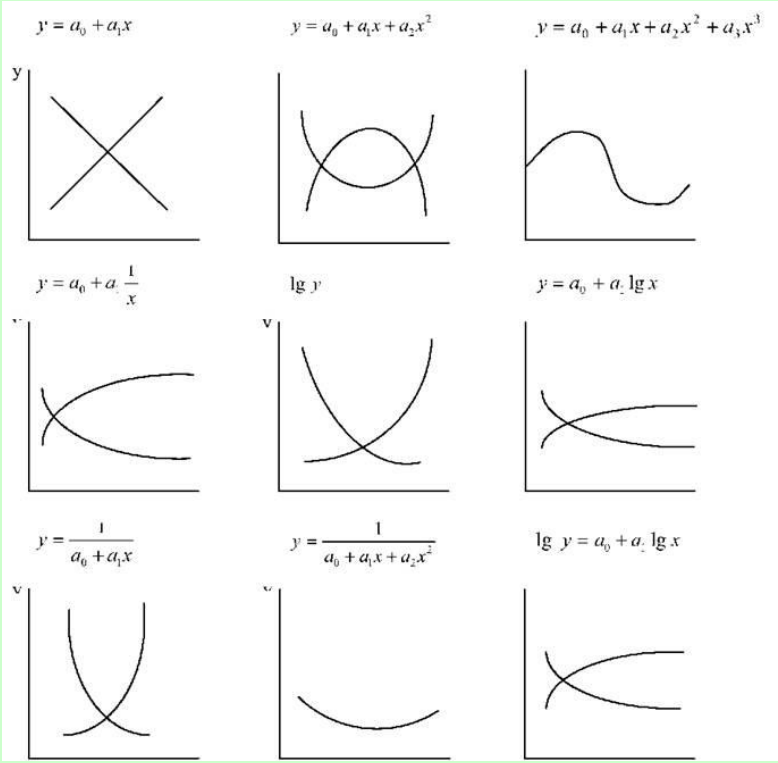
1. Лінійна $\hat{y} = b_0 + b_1 x$

2. Квадратична $\hat{y} = b_0 + b_1 x^2$

3. Гіперболічна $\hat{y} = b_0 + b_1 / x$

4. Степенева $\hat{y} = b_0 x^{b_1}$

5. Логарифмічна $\hat{y} = b_0 + b_1 \ln x$



*На практиці найчастіше використовується лінійний метод найменших квадратів, що використовується у випадку системи лінійних рівнянь**На практиці найчастіше використовується лінійний метод найменших квадратів, що використовується у випадку системи лінійних рівнянь.* *Зокрема важливим застосуванням у цьому випадку є оцінка параметрів у лінійній регресії, що широко застосовується в економічній статистиці.*

Метод найменших квадратів (МНК)

Невідомі параметри a_j обираються таким чином, щоб сума квадратів відхилень емпіричних (фактичних) значень y_i від розрахункових була мінімальною:

$$S = \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2 \to \min$$

Необхідна умова екстремуму функції

$$\frac{\partial S}{\partial p_j} = 0$$

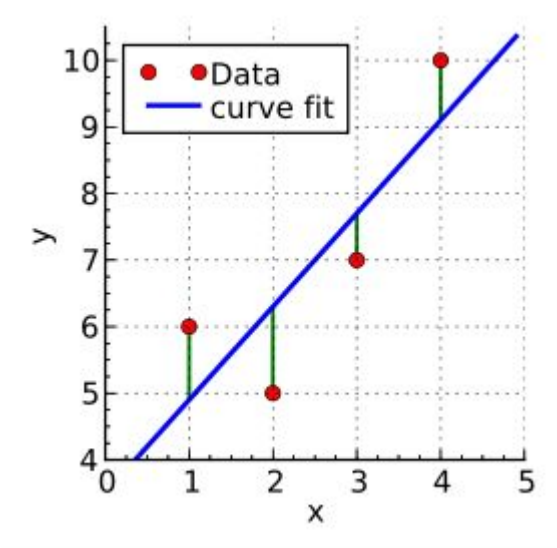
j=1..*p,* де *p* – число параметрів у системі

Цей метод застосовується для знаходження параметрів будь-якого регресійного рівняння з будь-яким числом незалежних змінних.

приклад

В результаті досліду, отримали чотири точки (x;y): (1;6), (2;5), (3;7), (4;10). Ми хочемо знайти лінію $y = b_0 + b_1 x$ яка найкраще підходить для цих точок. Інакше кажучи, ми хотіли б знайти числа b_0 і b_1, які приблизно розв'язують лінійну систему

$$6 = b_0 + 1b_1$$
$$5 = b_0 + 2b_1$$
$$7 = b_0 + 3b_1$$
$$10 = b_0 + 4b_1$$



Метод **найменших квадратів** розв'язання цієї проблеми полягає у спробі зробити якомога меншою суму квадратів *похибок* між правою і лівою сторонами цієї системи

Лінійна парна регресія

Сума квадратів для парної лінійної регресії матиме вигляд

$$S = \sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i \right)^2$$

Прирівнені до нуля її похідні дають систему нормальних рівнянь для визначення параметрів лінійної системи

$$\begin{cases} \dfrac{\partial S}{\partial b_0} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i \right) = 0; \\ \dfrac{\partial S}{\partial b_1} = 2\sum_{i=1}^{n} \left(b_0 + b_1 x_i - y_i \right) x_i = 0, \end{cases} \qquad \begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

розділивши обидві частини рівняння на *n*, отримаємо систему нормальних рівнянь:

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y}; \\ b_0 \bar{x} + b_1 \bar{x}^2 = \overline{xy}, \end{cases} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \quad \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}; \quad \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

Підставляючи значення з першого рівняння системи

$$b_0 = \bar{y} - b_1 \bar{x}$$

в рівняння регресії отримаємо $b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{K_{XY}}{s_x^2}$

$$K_{XY} = \overline{xy} - \bar{x} \cdot \bar{y} \qquad s_x^2 = \overline{x^2} - \bar{x}^2$$

де ***b*** $_1$ – вибірковий коефіцієнт регресії, *Kxy* – вибірковий кореляційний момент або вибіркова кореляція, s^2_x – вибіркова дисперсія змінної ***X***.

b $_1$ – вибірковий коефіцієнт регресії – показує, наскільки одиниць зміниться результуючий показник при зміні фактора на одиницю. , тобто швидкість змін.

Знак коефіцієнту регресії вказує на напрям змін.

Приклад

За статистичними даними витрат домогосподарств потрібно перевірити, чи є залежність між рівнем доходу населення та часткою витрат на харчування, та описати цю залежність.

Область	Витрати на споживання, у.о. *Фактор X*	Частка витрат на харчування,% *Результат У*
Харківська	9394.5	34.3
Дніпровська	10329.6	30.3
Чернівецька	9055.9	33.4
Івано-Франківська	8541.4	40.8
Одеська	8070.2	39.1
Львівська	8805.4	36.6
Київська	12904.9	30.7
Сумська	6633.8	41.1
Волинська	6397.3	41.9

Визначення параметрів моделі за допомогою методу найменших квадратів

Складаємо проміжні розрахунки і визначаємо рівняння

n	Область	x	y	x^2	yx
1	Харківська	9394.5	34.3	88256630.25	322231.4
2	Дніпровська	10329.6	30.3	106700636.16	312986.9
3	Чернівецька	9055.9	33.4	82009324.81	302467.1
4	Івано-Франківська	8541.4	40.8	72955513.96	348489.1
5	Одеська	8070.2	39.1	65128128.04	315544.8
6	Львівська	8805.4	36.6	77535069.16	322277.6
7	Київська	12904.9	30.7	166536444.01	396180.4
8	Сумська	6633.8	41.1	44007302.44	272649.2
9	Волинська	6397.3	41.9	40925447.29	268046.9
	всього	80133	328.2	744054496.12	2860873

b0 54.31945

b1 -0.002005

Y=54.319-0.002x

Момент *Кху* характеризує розсіювання величин та зв’язок між ними.

Для характеристики зв’язку між величинами застосовується відношення моменту *Кху* до добутку середніх квадратичних відхилень *Sx* і Sy величин *x* та *y*.

Це відношення називається **коефіцієнтом кореляції**.

$$r = \frac{b_1 \cdot s_x}{s_y} \qquad r = \frac{\overline{xy} - x \cdot y}{s_x \cdot s_y} \qquad r = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sqrt{\sum (y - \bar{y})^2 \sum (x - \bar{x})^2}}$$

r	-0.87084909

Властивості коефіцієнта кореляції

Коефіцієнт кореляції приймає значення на відрізку [-1;1]. Чим ближче $|r|$ до 1, тим тіснішим є кореляційний зв'язок.

При $|r| = 1$, кореляційний зв'язок становиться функціональним. При цьому всі значення, що спостерігаються, лежать на одній лінії.

При $|r| = 0$, кореляційний зв'язок відсутній і лінія регресії паралельна осі x.

При $r>0$ ($b1>0$) кореляційний зв'язок називають *прямим*.

При $r<0$ ($b1<0$) кореляційний зв'язок називають *оберненим*.

Оцінка адекватності регресійної моделі. **Коефіцієнт детермінації**.

Коефіцієнт детермінації показує, яка частка коливань результативної ознаки ***y*** зумовлена коливанням факторної ознаки *x*.

$$R^2 = \frac{\sum (Y - \bar{y})^2}{\sum (y - \bar{y})^2}. \qquad r = \sqrt{R^2}.$$

Де Y - оціночне значення пояснювальної змінної

y – фактичне значення

R^2 0.75837814

Коефіцієнт детермінації завжди позитивний і перебуває в межах від нуля до одиниці.

*Наприклад, R^2=0,758. Це означає, що на 75,8% зміна У залежить від зміни **X**, а (1-R^2) = 0,242, тобто на 24,2% - від інших факторів.*

Властивості коефіцієнта детермінації

Коефіцієнт детермінації приймає значення на відрізку **[0;1]**, тобто $0 \le R^2 \le 1$. Чим ближче R^2 до одиниці, тим краще регресія апроксимує емпіричні дані.

Якщо $R^2=1$, між змінними ***x*** та ***y*** існує лінійна функціональна залежність.

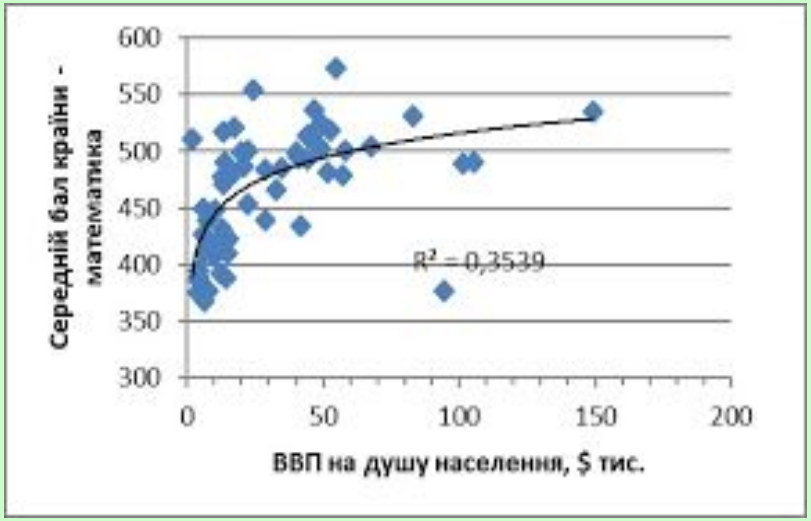
Якщо $R^2=0$, то варіація залежної змінної повністю обумовлена впливом випадкових та неврахованих у моделі змінних.

На практиці для оцінки ступеня апроксимації рівнянням регресії вихідних даних використовують наступні емпіричні правила:

1). $R^2>0,95$ - висока точність апроксимації.

2). $0,8<R^2<0,95$ - задовільна апроксимація.

3). $R^2<0,6$ - незадовільна апроксимація.



Оцінка значимості залежності

Оцінка значимості моделі проводиться за допомогою критерію Фішера

$$F_p = \frac{r^2}{1-r^2} * \frac{(n-m-1)}{m}$$

Ft 21.9708886

Де n – число спостережень

m – кількість факторів в моделі (в парній регресії =1)

Fp має бути більше за критичне значенням Ft, що є фіксованим табличним значенням для різних рівнів значимості α (найчастіше =0,05) і двох степенях свободи k1=m, k2=n-m-1

$k_2 \ / k_1$	1	2	3	4	5	6	8	12	24	∞
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54

Середня помилка апроксимації

Для оцінки якості моделі розраховують середню помилку апроксимації (*A*), яка показує, на скільки відсотків в середньому відрізняються фактичні значення результативного показника ***y*** від розрахункових значень ***У.***

$$A = \frac{1}{n}\sum \frac{|y - Y|}{y} 100$$

Модель регресії вважається достатньо точною, якщо А не перевищує 10%.

A 4.7365

n	Область	x	y	Y	y-Y	|y-Y|*100/y
1	Харківська	9394.5	34.3	35.48	-1.18	3.45
2	Дніпровська	10329.6	30.3	33.61	-3.31	10.92
3	Чернівецька	9055.9	33.4	36.16	-2.76	8.27
4	Івано-франківська	8541.4	40.8	37.19	3.61	8.84
5	Одеська	8070.2	39.1	38.14	0.96	2.46
6	Львівська	8805.4	36.6	36.66	-0.06	0.17
7	Київська	12904.9	30.7	28.44	2.26	7.35
8	Сумська	6633.8	41.1	41.02	0.08	0.20
9	Волинська	6397.3	41.9	41.49	0.41	0.97
	всього	80133	328.2	328.20	0.00	42.63

Прогнозування

Однією з задач економічного моделювання є прогнозування значень результуючого показника при певних значеннях фактору.

Доцільно представляти значення результату у вигляді *довірчого інтервалу.*

Довірчий інтервал визначається з заданою ймовірністю (значимістю) α з урахуванням величини граничної помилки $\Delta_{пр}$

$$\Delta_{\ddot{i}\,\eth} = \mu t_t \qquad \mu = \sqrt{\frac{(y-Y)^2}{n-m-1}}\left(1+\frac{1}{n}+\frac{(x^{\ddot{i}\,\eth}-\overline{x})^2}{\sum(x-\overline{x})^2}\right)$$

α найчастіше приймається 0,05. Це означає, що ймовірність того, що прогнозне значення результату буде знаходитись у межах довірчого інтервалу складає (1- α) 95%.

Визначіть з ймовірністю 95% інтервал можливих значень частки витрат на харчування, якщо витрати на споживання 14500 у.о.

n	Область	x	y	Y	y-Y	|y-Y|*100/y	(y-Y)^2	x-xcp	(x-xcp)^2
1	Харківська	9394.5	34.3	35.48	-1.18	3.45	1.40	490.83	240917.36
2	Дніпровська	10329.6	30.3	33.61	-3.31	10.92	10.94	1425.93	2033285.87
3	Чернівецька	9055.9	33.4	36.16	-2.76	8.27	7.63	152.23	23174.99
4	Івано-франківська	8541.4	40.8	37.19	3.61	8.84	13.01	-362.27	131237.14
5	Одеська	8070.2	39.1	38.14	0.96	2.46	0.93	-833.47	694666.68
6	Львівська	8805.4	36.6	36.66	-0.06	0.17	0.00	-98.27	9656.34
7	Київська	12904.9	30.7	28.44	2.26	7.35	5.09	4001.23	16009868.19
8	Сумська	6633.8	41.1	41.02	0.08	0.20	0.01	-2269.87	5152294.68
9	Волинська	6397.3	41.9	41.49	0.41	0.97	0.17	-2506.37	6281873.87
	всього	80133	328.2	328.20	0.00	42.63	39.17	0.00	30576975.12
	Xcp	8903.6667							
	Xпр	14500		25,245					

m 2.49359

t 2.36

Δ 5.884873

Yпр= 25.25 ±5.88