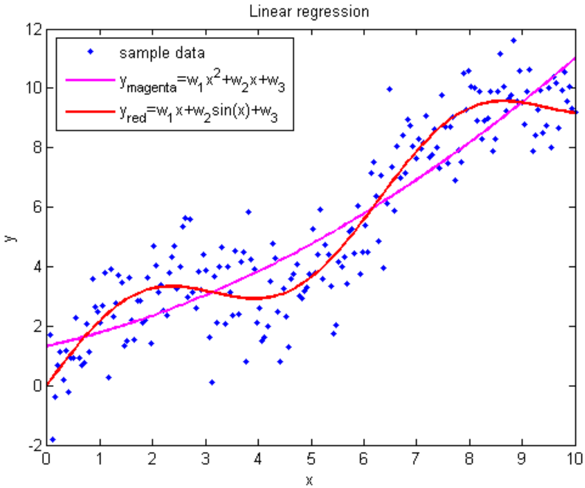
Регресійний аналіз даних

ГРАФОВІ ЙМОВІРНІСНІ МОДЕЛІ

Сумський державний університет

Регресія


Linear regression
sample data
$y_{magenta}=w_1x^2+w_2x+w_3$
$y_{red}=w_1x+w_2\sin(x)+w_3$
y
x
12
10
8
6
4
2
0
-2
0
1
2
3
4
5
6
7
8
9
10

Регресія

Кореляційний аналіз

При звичайній кореляції вивчається залежність між зміною двох ознак X та Y (степінь зв'язку варіації двох змінних)

-

Регресія

Кореляційний аналіз

При звичайній кореляції вивчається залежність між зміною двох ознак X та Y (степінь зв'язку варіації двох змінних)

-

Регресійний аналіз

Кількісно визначає як змінюється величина X відносно зміни Y на одиницю

-

Регресія

Оскільки змінних величин дві, то регресія може бути двосторонньою

Регресія

Оскільки змінних величин дві, то регресія може бути двосторонньою

визначення зміни ознаки X при зміні Y на одиницю

Регресія

Оскільки змінних величин дві, то регресія може бути двосторонньою

визначення зміни ознаки X при зміні Y на одиницю

визначення зміни ознаки Y при зміні X на одиницю

Регресія

Регресія виражається декількома способами

Регресія

Регресія виражається декількома способами

емпіричні лінії регресії

Регресія

Регресія виражається декількома способами

емпіричні лінії регресії

складання рівняння регресії і побудова теоретичної лінії регресії

Регресія

Регресія виражається декількома способами

- емпіричні лінії регресії
- складання рівняння регресії і побудова теоретичної лінії регресії
- обчислення коефіцієнта регресії

Емпіричні лінії регресії

Зазвичай користуються так званою кореляційною решіткою, де границі класів заміняють середніми значеннями

y \ x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	f_y	Средние по x для классов ряда y ($\bar{x}/y$)
y_6	a_1								$\bar{x}_6$
y_5	b_1	b_2	b_3						$\bar{x}_5$
y_4		c_2	c_3	c_4	c_5				$\bar{x}_4$
y_3		d_2	d_3	d_4	d_5				$\bar{x}_3$
y_2			e_3	e_4	e_5	e_6			$\bar{x}_2$
y_1				f_4	f_5	f_6	f_7		$\bar{x}_1$
f_x									
Средние по y для классов ряда x ($\bar{y}/x$)	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$	$\bar{y}_6$	$\bar{y}_7$		

Емпіричні лінії регресії

регресія x по y

У крайньому правому стовбчику записані середні значення ознаки x для класів ряду y

Емпіричні лінії регресії

регресія x по y

У крайньому правому стовбчику записані середні значення ознаки x для класів ряду y

регресія y по x

У нижньому горизонтальному рядку записані відповідні значення ознаки y для класів ряду x

Емпіричні лінії регресії

$y \backslash x$	225	275	325	375	425	475	525	575	f_y	$\overline{x}/y$
195								1	1	575
185					1	9	15	2	27	508
175			4	25	35	21	9	1	95	430
165		3	40	44	24	8			119	373
155	1	17	17	17	1				53	325
145	2	1	1						4	263
135	1								1	225
f_x	4	21	62	86	61	38	24	4	300	
$\overline{y}/x$	145	156	160	166	170	175	182	185		

Емпіричні лінії регресії

В класах x та y вказані середні значення

Значення цифр в останньому стовбчику і рядку отримуються шляхом обробки кожного горизонтального рядку та вертикального стовбчика, як окремого варіаційного ряду

Емпіричні лінії регресії

Наприклад другий знизу рядок таблиці

Із 4-х показників два мають вагу 225 г, один – 275 г і один – 325 г. Середнϵ значення по чотирьом даним буде 263 г.

Емпіричні лінії регресії

Наприклад другий знизу рядок таблиці

Із 4-х показників два мають вагу 225 г, один – 275 г і один – 325 г. Середнϵ значення по чотирьом даним буде 263 г.

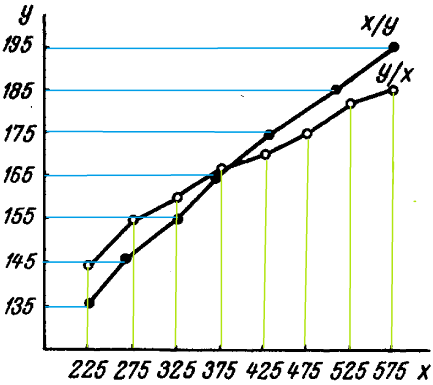
Наприклад сьомий вертикальний стовбчик основної частини таблиці

Із 24-х показників дев’ять мають розмір 175 см, а 15 – 185 см. Середнϵ значення обчислюють по формулі.

$$\frac{175 \cdot 9 + 185 \cdot 15}{24} = 181$$

Емпіричні лінії регресії

На снові показників $\bar{x}/y$ та $\bar{y}/x$ на одному графіку будують лінії регресії



Емпіричні лінії регресії

$y \backslash x$	225	275	325	375	425	475	525	575	f_y	$\overline{x}/y$
195								1	1	575
185					1	9	15	2	27	508
175			4	25	35	21	9	1	95	430
165		3	40	44	24	8			119	373
155	1	17	17	17	1				53	325
145	2	1	1						4	263
135	1								1	225
f_x	4	21	62	86	61	38	24	4	300	
$\overline{y}/x$	145	156	160	166	170	175	182	185		

Емпіричні лінії регресії

Позначають лінії регресії зазвичай x/y та y/x замість $\bar{x}/y$ та $\bar{y}/x$, оскільки у деяких випадках це можуть бути не середні значення класів відповідних ознак, а конкретні значення x та y

Емпіричні лінії регресії

Позначають лінії регресії зазвичай x/y та y/x замість $\bar{x}/y$ та $\bar{y}/x$, оскільки у деяких випадках це можуть бути не середні значення класів відповідних ознак, а конкретні значення x та y

Методом регресії можна користуватися, коли дані зводяться до декількох одиничних спостережень x та y

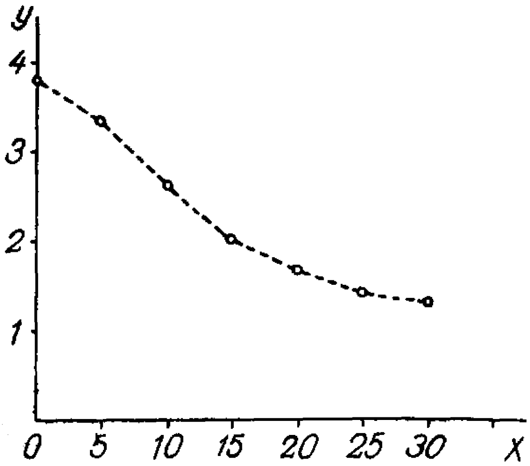
Емпіричні лінії регресії

Позначають лінії регресії зазвичай x/y та y/x замість $\bar{x}/y$ та $\bar{y}/x$, оскільки у деяких випадках це можуть бути не середні значення класів відповідних ознак, а конкретні значення x та y

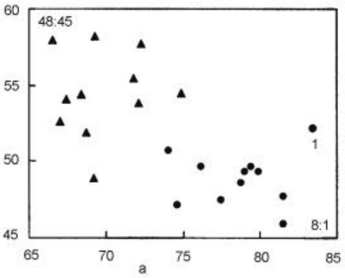
Методом регресії можна користуватися, коли дані зводяться до декількох одиничних спостережень x та y

Тоді на кореляційне поле можна нанести пари даних x та y, і за їх розташуванням зробити висновок щодо зв'язку

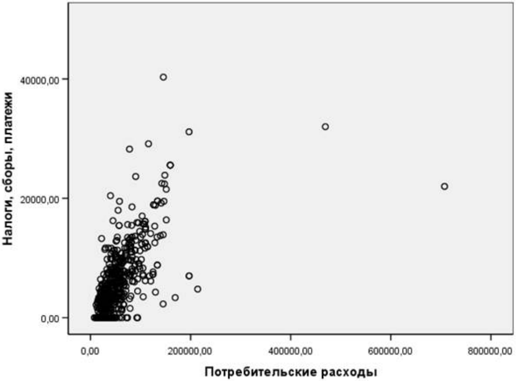
Емпіричні лінії регресії



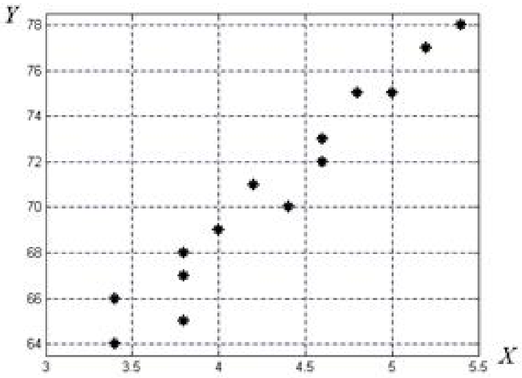
Емпіричні лінії регресії



Емпіричні лінії регресії



Емпіричні лінії регресії



Метод ковзаючої середньої

Один вигляд емпіричної лінії регресії може підказати форму зв'язку (лінійний, параболічний та ін.)

Метод ковзаючої середньої

Один вигляд емпіричної лінії регресії може підказати форму зв'язку (лінійний, параболічний та ін.)

Щоб початково проаналізувати зв'язок можна виключити випадкові коливання емпіричної лінії регресії за допомогою метода ковзаючої середньої

Метод ковзаючої середньої

Один вигляд емпіричної лінії регресії може підказати форму зв'язку (лінійний, параболічний та ін.)

Щоб початково проаналізувати зв'язок можна виключити випадкові коливання емпіричної лінії регресії за допомогою метода ковзаючої середньої

Отримані значення ознаки y, що відповідають фіксованим значенням x, замінюють новими, отриманими шляхом додавання трьох, або п'яти поруч розташованих даних

Метод ковзаючої середньої

$$y_1 = \frac{y_1 + y_2 + y_3}{3}$$

$$y_1 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$$

Метод ковзаючої середньої

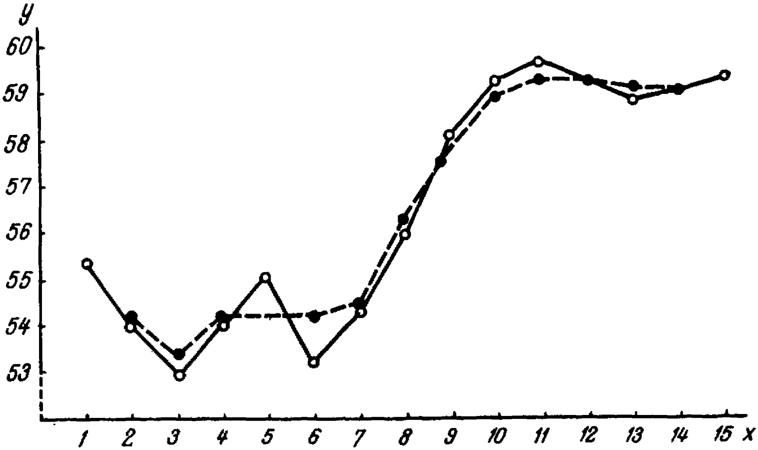
$$y_2 = \frac{y_2 + y_3 + y_4}{3} \quad \text{или} \quad y_2 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5};$$

$$y_3 = \frac{y_3 + y_4 + y_5}{3} \quad \text{или} \quad y_3 = \frac{y_3 + y_4 + y_5 + y_6 + y_7}{5} \quad \text{и т. д.}$$

Метод ковзаючої середньої

x_i	y_i	y'	y''	x_i	y_i	y'	y''
1	55,4	—	—	9	58,0	57,7	57,4
2	54,0	54,1	—	10	59,2	58,9	58,4
3	53,0	53,4	54,3	11	59,6	59,3	59,0
4	54,1	54,1	53,9	12	59,2	59,2	59,2
5	55,1	54,1	53,9	13	58,8	59,0	59,2
6	53,2	54,2	54,5	14	59,0	59,0	—
7	54,3	54,5	55,3	15	59,3	—	—
8	56,0	56,3	56,1				

Метод ковзаючої середньої



Рівняння регресії

Емпірична лінія, хоч і відображає характер зв'язку, але зазвичай представляє собою ламану криву.

Рівняння регресії

Емпірична лінія, хоч і відображає характер зв'язку, але зазвичай представляє собою ламану криву.

Щоб точно визначити ознаку y, що відповідає фіксованому значенню x, треба записати рівняння регресії.

Рівняння регресії

Емпірична лінія, хоч і відображає характер зв'язку, але зазвичай представляє собою ламану криву.

Щоб точно визначити ознаку y, що відповідає фіксованому значенню x, треба записати рівняння регресії.

Прямолінійна регресія

у загальному випадку можна представити співвідношенням

$$y_i - \bar{y} = b(x_i - \bar{x})$$

$$y_i = \bar{y} + b(x_i - \bar{x})$$

$$y = a + bx$$

Рівняння регресії

Для того, щоб знайти коефіцієнти a і b необхідно вирішити алгебраїчну систему рівнянь

Рівняння регресії

Для того, щоб знайти коефіцієнти a і b необхідно вирішити алгебраїчну систему рівнянь

$$na + (\sum x_i)b = \sum y_i$$
$$(\sum x_i)a + (\sum x_i^2)b = \sum x_i y_i$$

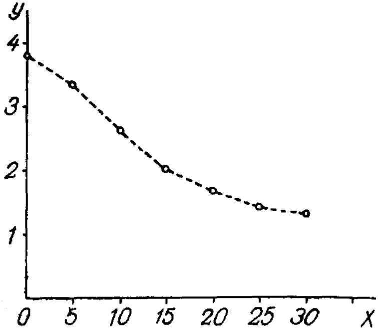
Рівняння регресії

Для того, щоб знайти коефіцієнти a і b необхідно вирішити алгебраїчну систему рівнянь

$$na + (\sum x_i)b = \sum y_i$$
$$(\sum x_i)a + (\sum x_i^2)b = \sum x_i y_i$$

Ці рівняння базуються на методі найменших квадратів, тобто обчислюються такі параметри системи, для яких сума квадратів відхилень значень y від теоретичних є найменшою.

Рівняння регресії



x_i	y_i
0	3,8
5	3,4
10	2,6
15	2,0
20	1,7
25	1,4
30	1,3

Рівняння регресії

x_i	y_i	x_i^2	$x_i y_i$
0	3,8	0	0
5	3,4	25	17
10	2,6	100	26
15	2,0	225	30
20	1,7	400	34
25	1,4	625	35
30	1,3	900	39
$\Sigma x_i = 105$ $\bar{x} = 15$	$\Sigma y_i = 16{,}2$ $\bar{y} = 2{,}3$	$\Sigma x_i^2 = 2275$ $n = 7$	$\Sigma x_i y_i = 181$

Рівняння регресії

$$1.\quad 7a + 105b = 16{,}2;$$

$$2.\quad 105a + 2275b = 181.$$

$$
\begin{array}{r}
-\begin{array}{r} 105a + 2275b = 181 \\ 105a + 1575b = 243 \end{array} \\
\hline
700b = -62
\end{array}
$$

$$b = -0{,}089\ (\approx -0{,}09).$$

Рівняння регресії

$$7a = 16{,}2 + 9{,}35 = 25{,}55;$$

$$a = 3{,}65.$$

$$y = 3{,}65 - 0{,}09\,x.$$

Коефіцієнт регресії

Коефіцієнт регресії R при прямолінійному зв'язку співпадає з коефіцієнтом b

Коефіцієнт регресії

Коефіцієнт регресії R при прямолінійному зв'язку співпадає з коефіцієнтом b

Загальне визначення

$$R_{x/y} = r\frac{\sigma_x}{\sigma_y}$$

$$R_{y/x} = r\frac{\sigma_y}{\sigma_x}$$

Коефіцієнт регресії

Коефіцієнт регресії R при прямолінійному зв'язку співпадає з коефіцієнтом b

Загальне визначення

$$R_{x/y} = r\frac{\sigma_x}{\sigma_y}$$

$$R_{y/x} = r\frac{\sigma_y}{\sigma_x}$$

r – коефіцієнт кореляції.

Коефіцієнт регресії

Загальне визначення

$$R_{x/y} = r\sqrt{\frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2}}$$

$$R_{y/x} = r\sqrt{\frac{\sum(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2}}$$

Коефіцієнт регресії

Коефіцієнт регресії $R(= b)$ виражений у відхиленнях від середнього арифметичного

$$b_{y/x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

$$b_{x/y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2};$$

Коефіцієнт регресії

Коефіцієнт регресії $R(=b)$ виражений у конкретних значеннях ознак

$$b_{y/x} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}},$$

$$b_{x/y} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}.$$

Коефіцієнт регресії

x_i	y_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}) \times (y_i - \overline{y})$
0	3,8	—15	225	1,5	2,25	—22,5
5	3,4	—10	100	1,1	1,21	—11,0
10	2,6	—5	25	0,3	0,09	—1,5
15	2,0	0	0	—0,3	0,0	0
20	1,7	5	25	—0,6	0,36	—3,0
25	1,4	10	100	—0,9	0,81	—9,0
30	1,3	15	225	—1,0	1,00	—15,0
$\Sigma x_i = 105$ $\overline{x} = 15$	$\Sigma y_i = 16,2$ $\overline{y} = 2,3$		$\Sigma = 700$		$\Sigma = 5,81$	$\Sigma = -62,0$

Коефіцієнт регресії

$$b_{y/x} = \frac{-62}{700} = -0{,}089 \approx -0{,}09;$$

$$b_{x/y} = \frac{-62}{5{,}81} = -10{,}67.$$

Коефіцієнт регресії

$$\bar{y} = a + b\bar{x}.$$

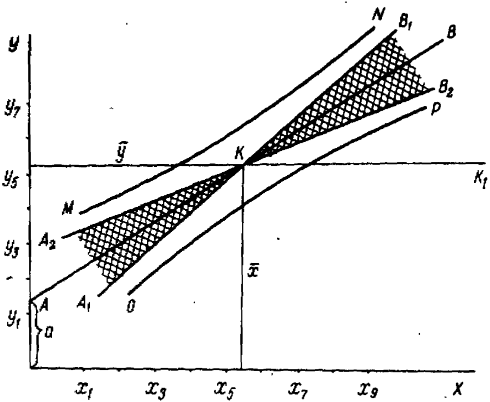
$$a = \bar{y} - b\bar{x}.$$

$$a = 2{,}3 - (-0{,}09) \cdot 15 = 2{,}3 + 1{,}35 = 3{,}65$$

$$y = 3{,}65 - 0{,}09\,x.$$

Коефіцієнт регресії

Достовірність лінії регресії та коефіцієнта регресії



Коефіцієнт регресії

Оновою для визначення можливого відхилення лінії регресії є сума квадратів відхилень фактичних значень від теоретичних

$$\sigma_{y \cdot x}^{2} = \frac{\Sigma (y_i - \hat{y}_i)^2}{n-2}.$$

$$\sigma_{y \cdot x} = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{n-2}}.$$

Коефіцієнт регресії

Похибка коефіцієнта регресії

$$s_b = \frac{\sigma_{y \cdot x}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Коефіцієнт регресії

Коваріація

$$\mathrm{cov}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Коефіцієнт регресії

Параболічна залежність

$$y = a + bx + cx^2.$$

1. $na + (\Sigma x_i) b + (\Sigma x_i^2) c = \Sigma y_i;$
2. $(\Sigma x_i) a + (\Sigma x_i^2) b + (\Sigma x_i^3) c = \Sigma x_i y_i;$
3. $(\Sigma x_i^2) a + (\Sigma x_i^3) b + (\Sigma x_i^4) c = \Sigma x_i^2 y_i.$

Коефіцієнт регресії

Показникова залежність

$$W = A \cdot B^x$$

$$\log W = \log A + (\log B)x$$

Коефіцієнт регресії

Показникова залежність