Линеен модел на корелационна зависимост и Метод на най-малките квадрати

Една от задачите на статистиката е изследването на връзката между две променливи. Когато зависимостта не е ясно изразена (функционална), казваме че има корелационна зависимост. Характерът на зависимостта се оценява от регресионния анализ. Зависимостта може да бъде между 2 и повее величини, а от друга страна – линейна, квадратична, експоненциална и др.

Предполагаме, че е направено наблюдение върху съвкупност от n на брой единици по два признака. Независимата променлива(фактор) се означава с x, а зависимата(резултативна)- с y. Данните са от конкретно наблюдение, поради което те изразяват корелационна зависимост между изследваните признаци. Най-общата представа за зависимостта между две променливи се получава като се начертае диаграма на разсейването. Ако връзката е приблизително линейна съставяме уравнението $\hat{y} = b_0 + b_1.x$, което се нарича регресионно уравнение (линеен регресионен модел).

b_0 и $b_1 -$ параметри на модела или регресионни коефициенти.

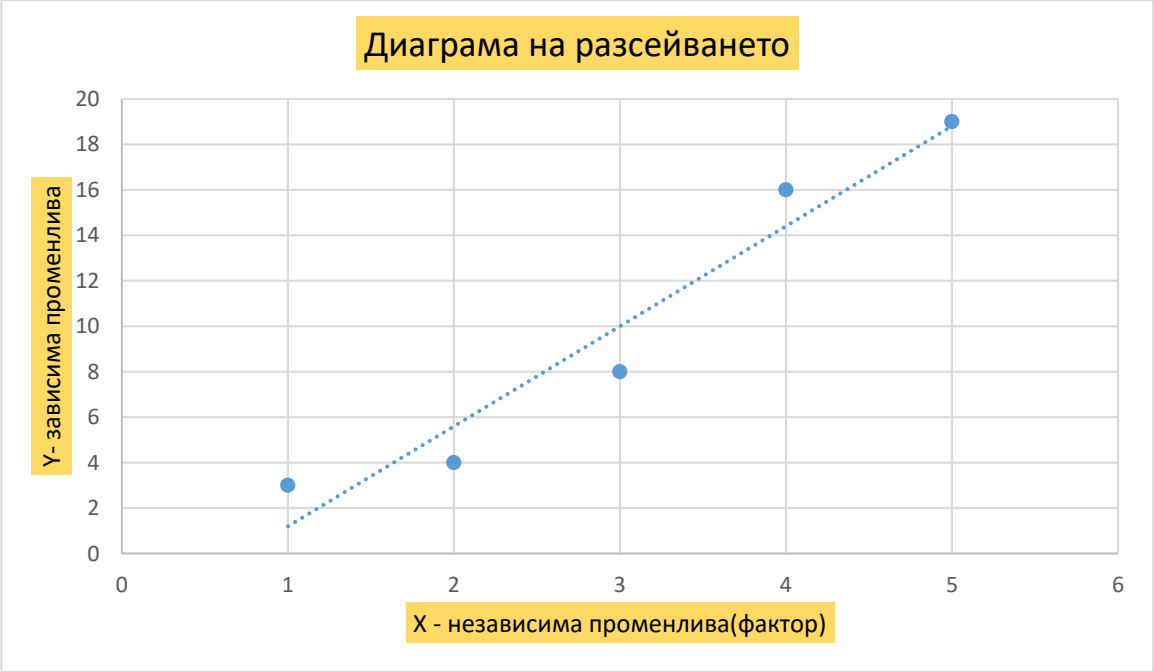
$b_0 -$ отрез, $b_1 -$ наклон

Търсим такова уравнение на правата $\hat{y} = b_0 + b_1.x$, че тя да минава най-близо до всички емпирични точки. За решаването на тази задача се използва метода на най-малките квадрати – МНМК.

Пример:

Направени са наблюдения върху два признака върху един и същ обект, като X е независимата променлива, а У е зависимата. Данните са показани в таблицата:

X	1	2	3	4	5
Y	3	4	8	16	19



Построяваме диаграма на разсейването – избираме точкова диаграма и линия на тенденцията. Можем да предположим, че точките се групират около права , т. е зависимостта е линейна.

1. Нека регресионното уравнение е $\hat{y} = b_0 + b_1.x$
2. Параметрите b_0 и b_1 определяме по МНМК. Прилагането на този метод се свежда до решаване на система, наречена система нормални уравнения.

$\begin{vmatrix} S_y = nb_0 + b_1.S_x \\ S_{yx} = b_0 S_x + b_1.S_{x^2} \end{vmatrix}$ – **n** е обемът на извадката.

$$S_x = x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

$$S_y = y_1 + y_2 + \cdots + y_n = \sum_{i=1}^{n} y_i$$

$$S_{x^2} = x_1^2 + x_2^2 + \cdots + x_n^2 = \sum_{i=1}^{n} x_i^2$$

$$S_{xy} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i$$

За удобство при извършване на изчисленията използваме следната таблица:

i	x_i	y_i	x_i^2	$x_i y_i$
1	1	3	1	3
2	2	4	4	8
3	3	8	9	24
4	4	16	16	64
5	5	19	25	95
$n = 5$	$S_x = 15$	$S_y = 50$	$S_{x^2} = 55$	$S_{xy} = 194$

Системата $\begin{vmatrix} S_y = nb_0 + b_1.S_x \\ S_{yx} = b_0 S_x + b_1.S_{x^2} \end{vmatrix}$ приема вида $\begin{vmatrix} 50 = 5b_0 + 15b_1 \\ 194 = 15b_0 + 55b_1 \end{vmatrix}$

Решенията са $b_0 = -3{,}2$ и $b_1 = 4{,}4$
Регресионната права е $\hat{y} = -3{,}2 + 4{,}4.x$
Стойностите $\hat{y}$ се наричат предсказана(прогнозна) стойност. Например, ако x приеме стойност 6, то за y се очаква да приеме стойност $\hat{y} = 23{,}2$

Задача:

Трябва да се разбере как цената на наемите на апартаменти във вашия град зависи от площта на жилището. Събират се данни за 10 апартамента, като за всеки се записва:

- Площ в квадратни метри (x)
- Месечен наем в лева ($\hat{y}$)

Трябва да се изгради **линеен регресионен модел** от вида:

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

където:

- Y е прогнозният наем
- X е площта
- b_0 и b_1 – параметри на модела или регресионни коефициенти.

1. Да се построи линейната регресия по дадените данни.
2. Да се използва модела, за да се предвиди наема за апартамент с площ 85 кв.м.

3. Да се прецени дали площта има значимо влияние върху цената на наема (може да се използва коефициентът на детерминация).

Апартамент	Площ в кв.м - x	Наем в лв. -$\hat{y}$
1	40	500
2	55	650
3	60	700
4	70	800
5	80	890
6	90	970
7	100	1050
8	110	1150
9	120	1230
10	130	1300

Решение:

1. Да се начертае диаграма: площта на x-оста, наемът на y-оста.
2. Да се използва метода на най-малките квадрати, за да се намерят коефициентите b_0 и b_1 от уравнението на правата.
3. Да се изрази линейната функция: $\hat{y} = b_0 + b_1.x$
4. Да се направи прогноза: колко ще струва наем на апартамент с площ 85 кв.м?
5. Да се изчисли коефициента на детерминация R^2, за да се прецени колко добре моделът обяснява зависимостта между площ и наем.

Уравнение на модела:

$$Y = 8{,}88X + 165$$

Коефициент на детерминация:-R^2

R^2=0,998
Това означава, че има много силна зависимост.
Прогноза за апартамент с площ 85 кв.м:
Y=919,56 лв