Розділ 4. ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

При вивченні тісноти зв'язку між різними ознаками економічного чи соціального об'єкта головною задачею є встановлення виду кореляційної залежності результативної ознаки (Y) від факторної (X), тобто виду функціональної залежності $\overline{Y}=f(X)$. В першу чергу це пов'язано з необхідністю прогнозування досліджуваних процесів. Математико-статистичний апарат, що дозволяє встановити вид кореляційної залежності називається **регресійним аналізом**, а функція, яка описує цю залежність, називається **рівнянням регресії**.

4.1. Встановлення виду кореляційної залежності

Регресійний аналіз проводиться за такими етапами:

1) Встановлення виду кореляційної залежності результативної ознаки Y від факторної ознаки X.

2) Побудова регресійної моделі.

3) Перевірка статистичної значущості побудованої моделі.

Перший етап регресійного аналізу є найважливішим, оскільки помилки у виборі виду залежності призводять до побудови регресійної моделі, що не відповідає емпіричним даним і не може використовуватися для прогнозування.

Вибіркові дані для вивчення кореляційного зв'язку між ознаками X та Y, зазвичай, мають вигляд пар їх значень: $(x_1;y_1)$, $(x_2;y_2)$, …, $(x_n;y_n)$, x_i – значення величини X, y_i – значення Y, n – кількість пар значень, $i=\overline{1,n}$. Якщо їх кількість достатньо велика, то для зручності розрахунків дані групуються (див. п. 3.2) і будується статистичний ряд, що містить значення X, відповідні середні значення Y та частоти (табл. 4.1).

Таблиця 4.1

x_i	x_1	x_2	…	x_k
$\overline{y}_{x_i}$	$\overline{y}_{x_1}$	$\overline{y}_{x_2}$	…	$\overline{y}_{x_k}$
n_i	n_1	n_2	…	n_k

Згруповані дані (табл. 4.1) зображуються графічно, що часто дозволяє визначити вид залежності Y від X.

Ламана лінія, що сполучає точки з координатами $(x_i; \overline{y}_{x_i})$, називається **емпіричною лінією регресії**.

Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками (рис. 4.1).

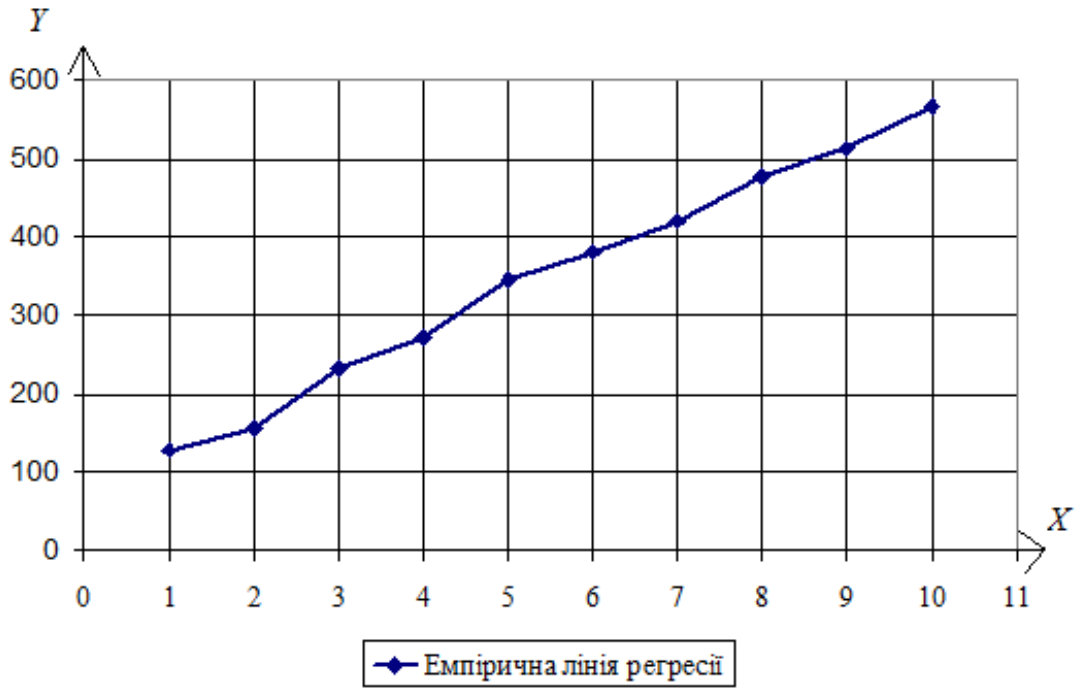

Рисунок 4.1. Гіпотетична лінійна залежність

В іншому випадку висувається гіпотеза про наявність нелінійного зв'язку (рис. 4.2).

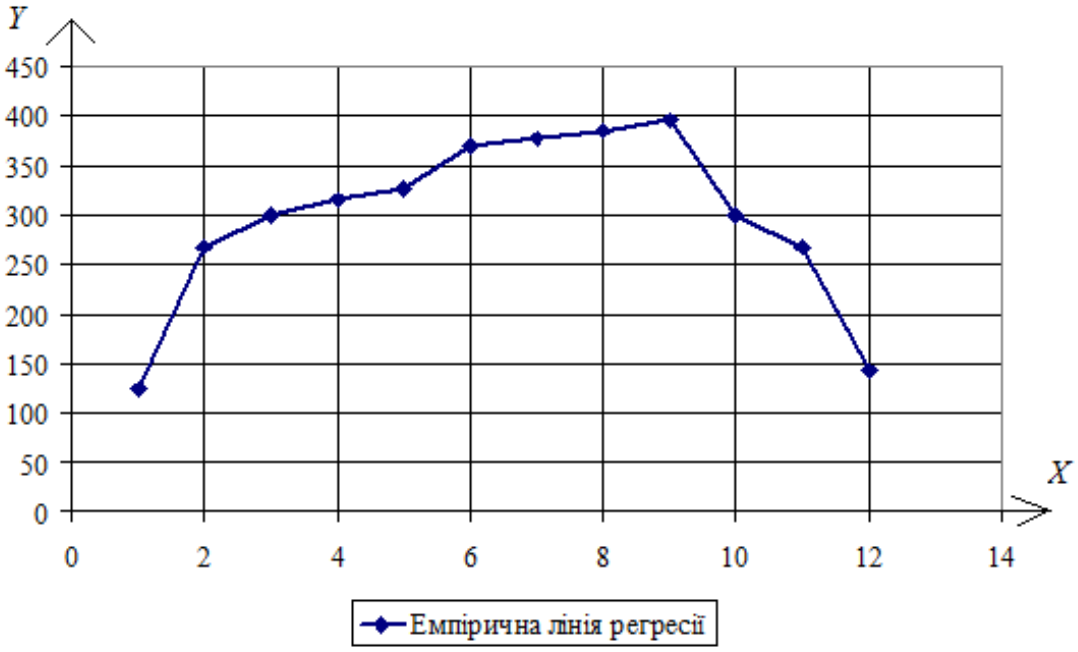

Рисунок 4.2. Гіпотетична нелінійна залежність

4.2. Лінійна регресія

Якщо висунуто гіпотезу про наявність лінійної залежності результативної ознаки (Y) від факторної (X), то рівняння регресії має вид:

$$\overline{y_x} = ax + b\,, \qquad (4.1)$$

де a, b − параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння (4.1). Параметри рівняння регресії можна знайти за **методом найменших квадратів**.

Ідея методу найменших квадратів

Нехай при вивченні залежності *Y* від *X* було отримано вибіркові дані: $x_1, x_2, ..., x_n$ – значення величини *X*, $y_1, y_2, ..., y_n$ – відповідні значення *Y*. За вибірковими даними було побудовано рівняння регресії $y = ax + b$. Якщо в рівняння підставити замість *x* значення $x_1, x_2, ..., x_n$, то будуть отримані теоретичні значення *Y*: $y_{1,теор}, y_{2,теор}, ..., y_{n,теор}$, які відрізняються від $y_1, y_2, ..., y_n$. Різниця значень $y_{i,теор} - y_i$ називається помилкою регресійної моделі і позначається e_i. Якщо параметри рівняння підбираються так, щоб сума квадратів помилок була мінімальною, то говорять, що вони отримані за методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії за методом найменших квадратів знаходяться з системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum\limits_{i=1}^{k} x_i^2 n_i + b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum\limits_{i=1}^{k} x_i n_i + b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}. \qquad (4.2)$$

Якщо вибіркові дані не згруповані, то система (4.1) значно спрощується:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i \end{cases}. \qquad (4.3)$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o, \qquad (4.4)$$

де $Q = \sum\limits_{i=1}^{k} \left(\overline{y_{x_i}} - \overline{y} \right)^2 n_i$ − загальна варіація, тобто сума квадратів відхилень емпіричних значень *Y* від середнього $\overline{y} = \dfrac{\sum\limits_{i=1}^{k} \overline{y_{x_i}} n_i}{n}$;

$Q_p = \sum\limits_{i=1}^{k} \left(y_{i,теор} - \overline{y} \right)^2 n_i$ − варіація регресії, тобто сума квадратів відхилень теоретичних значень *Y* від середнього, що обумовлена регресією;

$Q_o = \sum_{i=1}^{k} \left(y_{i,теор} - \overline{y}_{x_i} \right)^2 n_i$ – варіація залишків, тобто сума квадратів відхилень теоретичних значень Y від емпіричних.

У випадку незгрупованих даних загальна варіація, варіації регресії і залишків знаходяться за формулами: $Q = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$; $Q_p = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - \overline{y} \right)^2$; $Q_o = \sum_{i=1}^{n} \left(y_{i,\text{теор}} - y_i \right)^2$; а середнє значення за формулою $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Для перевірки статистичної значущості рівняння регресії розраховується F-статистика за формулою:

$$F = \frac{Q_p (n-l)}{Q_o (l-1)}, \tag{4.5}$$

де n – кількість спостережень, l – кількість груп у кореляційній таблиці або кількість параметрів моделі у випадку незгрупованих даних. Розраховане значення F-статистики порівнюється з критичним значенням $F_{кр}$ розподілу Фішера, яке можна знайти за статистичними таблицями або за допомогою вбудованої функції Excel $FРАСПОБР(\alpha, k_1, k_2)$, де $k_1 = l - 1$; $k_2 = n - l$ – степені свободи, α – рівень значущості.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2, що показує частину варіації значень результативної ознаки Y, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}. \tag{4.6}$$

Значення коефіцієнта детермінації знаходяться в інтервалі $[0;1]$, тобто $0 \le R^2 \le 1$. Чим ближче R^2 до 1, тим краще отримане рівняння регресії пояснює поведінку результативної ознаки. Наприклад, якщо $R^2 = 0{,}98$, то 98% варіації результативної ознаки Y пояснюється рівнянням регресії.

Приклад 4.1. Побудувати регресійну модель, що описує залежність сумарних виробничих затрат Y (тис. грн.) від об'ємів виробництва X (тис. од.). Відповідні статистичні дані задано у табл. 4.2.

Таблиця 4.2

X	41	44	52	57	59	64	68	70	73	75
Y	670	657	713	736	778	812	833	876	911	932

Розв'язок. В табл. 4.2 задано вибіркові дані: значення x_i, $i = \overline{1,n}$ величини X та відповідні значення y_i, $i = \overline{1,n}$; кількість пар – $n = 10$ невелика, тому для проведення регресійного аналізу їх можна не групувати.

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.3).

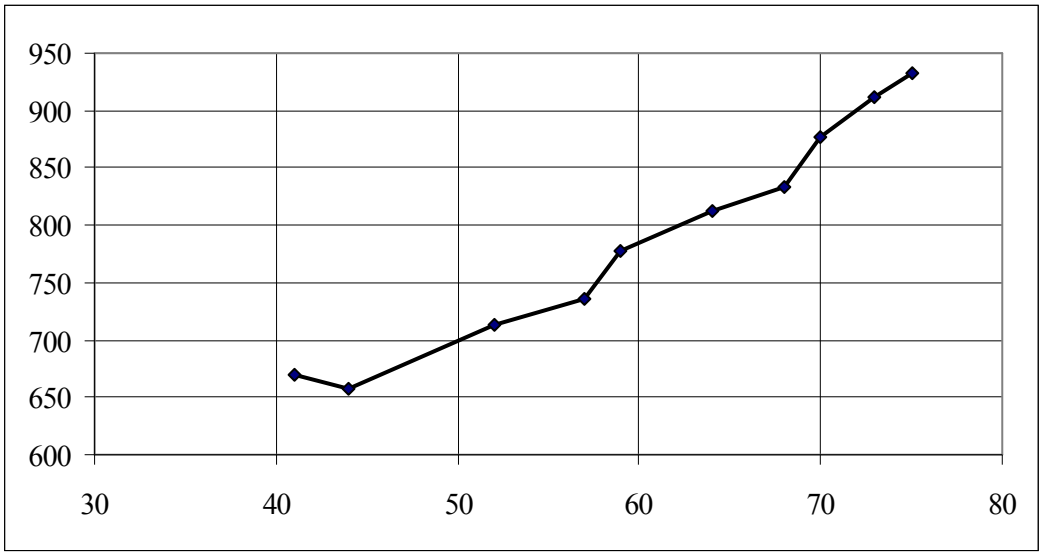

Рисунок 4.3. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до прямої лінії, то висуваємо гіпотезу про лінійну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $y = ax + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему (4.3) для не згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.3).

Таблиця 4.3

Розрахункова таблиця											Суми
x_i	41	44	52	57	59	64	68	70	73	75	603
y_i	670	657	713	736	778	812	833	876	911	932	7918
x_i^2	1681	1936	2704	3249	3481	4096	4624	4900	5329	5625	37625
$x_i y_i$	27470	28908	37076	41952	45902	51968	56644	61320	66503	69900	487643

Отже, складемо систему для знаходження параметрів рівняння регресії та розв'яжемо її за правилом Крамера:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \begin{cases} 37625a + 603b = 487643 \\ 603a + 10b = 7918 \end{cases}.$$

Знайдемо визначник основної матриці системи, яка складена із коефіцієнтів перед невідомими: $\Delta = \begin{vmatrix} 37625 & 603 \\ 603 & 10 \end{vmatrix} = 37625 \cdot 10 - 603^2 = 12641$.

Знайдемо допоміжні визначники, що отримуються із попереднього заміною відповідного стовпця коефіцієнтів на стовпець вільних членів:

$$\Delta a = \begin{vmatrix} 487643 & 603 \\ 7918 & 10 \end{vmatrix} = 487643 \cdot 10 - 603 \cdot 7918 = 101876;$$

$$\Delta b = \begin{vmatrix} 37626 & 487643 \\ 603 & 7918 \end{vmatrix} = 37626 \cdot 7918 - 48743 \cdot 603 = 3866021.$$

Знайдемо невідомі за формулами Крамера:

$$a = \frac{\Delta a}{\Delta} = \frac{101876}{12641} \approx 8,06; \quad b = \frac{\Delta b}{\Delta} = \frac{3866021}{12641} \approx 305,83.$$

Отже, шукане рівняння регресії має вигляд $y = 8,06x - 305,83$.

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для чого знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.4).

Передусім знайдемо $\bar{y}$: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \approx 791,8$.

Таблиця 4.4

x_i	y_i	$y_{i,\text{теор}}$	$\left(y_i - \bar{y}\right)^2$	$\left(y_{i,\text{теор}} - \bar{y}\right)^2$	$\left(y_{i,\text{теор}} - y_i\right)^2$
41	670	636,26	14835,24	24193,32	1138,52
44	657	660,44	18171,04	17256,64	11,80
52	713	724,91	6209,44	4474,42	141,82
57	736	765,20	3113,64	707,31	852,92
59	778	781,32	190,44	109,77	11,04
64	812	821,62	408,04	889,17	92,52
68	833	853,86	1697,44	3850,90	434,96
70	876	869,97	7089,64	6111,17	36,31
73	911	894,15	14208,64	10475,83	283,87
75	932	910,27	19656,04	14035,10	472,20
		Суми	85579,6	82103,63	3475,974

Отже, $Q = 85579,6$; $Q_p = 82103,63$; $Q_o = 3475,974$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $85579,6 = 82103,63 + 3475,974$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 10$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p\left(n-l\right)}{Q_o\left(l-1\right)} = \frac{82103\left(10-2\right)}{3475,974\left(2-1\right)} \approx 188,96.$$

Знайдемо $F_{\text{кр}}$: $F_{\text{кр}} = FРАСПОБР\left(0,001; 2-1; 10-2\right) \approx 25,41$. Розраховане значення *F*-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{82103,63}{85579,6} \approx 0,96$. Значення коефіцієнта детермінації свідчить, що 96% варіації результативної ознаки Y пояснюються рівнянням регресії.

Висновок: Сумарні виробничі затрати Y (тис. грн.) лінійно залежать від об'єму виробництва X (тис. од.). Залежність описується рівнянням $y = 8,06x - 305,83$, яке є статистично значущим на рівні значущості 0,001 та описує 96% вибіркових даних.

4.3. Нелінійна регресія

Якщо висунуто гіпотезу про наявність нелінійної залежності результативної ознаки (Y) від факторної (X), то регресійний аналіз проводиться за тими ж етапами, як і у випадку лінійної залежності. Вид рівнянь регресії і системи для знаходження їх параметрів для нелінійних залежностей, що найчастіше зустрічаються, надано у табл. 4.5.

Таблиця 4.5

Рівняння параболічної регресії:

$$\overline{y_x} = ax^2 + bx + c.$$

Система для знаходження параметрів:

для згрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{k} x_i^4 n_i + b\sum_{i=1}^{k} x_i^3 n_i + c\sum_{i=1}^{k} x_i^2 n_i = \sum_{i=1}^{k} x_i^2 n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^3 n_i + b\sum_{i=1}^{k} x_i^2 n_i + c\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i + c\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}$$

для незгрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$$

Рівняння гіперболічної регресії:

$$\overline{y_x} = \frac{a}{x} + b.$$

Система для знаходження параметрів:

для згрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{k} \frac{1}{x_i^2} n_i + b\sum_{i=1}^{k} \frac{1}{x_i} n_i = \sum_{i=1}^{k} \frac{1}{x_i} \overline{y_{x_i}} n_i \\ a\sum_{i=1}^{k} \frac{1}{x_i} n_i + b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \overline{y_{x_i}} n_i \end{cases}$$

для незгрупованих вибіркових даних:

$$\begin{cases} a\sum_{i=1}^{n} \frac{1}{x_i^2} + b\sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{n} \frac{1}{x_i} y_i \\ a\sum_{i=1}^{n} \frac{1}{x_i} + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Рівняння показникової регресії:

$$\overline{y_x} = ba^x.$$

Система для знаходження параметрів:	
для згрупованих вибіркових даних:	для незгрупованих вибіркових даних:
$\begin{cases} \lg a\sum\limits_{i=1}^{k} x_i^2 n_i + \lg b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \lg \overline{y_{x_i}} \\ \lg a\sum\limits_{i=1}^{k} x_i n_i + \lg b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \lg \overline{y_{x_i}} \end{cases}$	$\begin{cases} \lg a\sum\limits_{i=1}^{n} x_i^2 + \lg b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i \lg y_i \\ \lg a\sum\limits_{i=1}^{n} x_i + n\lg b = \sum\limits_{i=1}^{n} \lg y_i \end{cases}$

Перевірка статистичної значущості нелінійної регресійної моделі також здійснюється за *F*-статистикою. При цьому для параболічної регресії кількість параметрів $l = 3$, для гіперболічної і показникової – $l = 2$.

Приклад 4.2. Дано розподіл однотипних підприємств за об'ємом виробництва *X* (тис. од.) і собівартістю одиниці продукції *Y* (грн.) (табл. 4.6). Знайти регресійну модель, що описує залежність собівартості продукції від об'єму виробництва.

Таблиця 4.6

X \ Y	10	15	20	25
25	–	-	1	2
50	–	2	2	–
75	–	5	3	1
100	1	3	–	–
125	3	1	1	–

Розв'язок. Для проведення регресійного аналізу за даними табл. 4.6 побудуємо кореляційну таблицю (табл. 4.7).

Таблиця 4.7

y_j \ x_i	25	50	75	100	125	n_j
10	0	0	0	1	3	4
15	0	2	5	3	1	11
20	1	2	3	0	1	7
25	2	0	1	0	0	3
n_i	3	4	9	4	5	$n = 25$

За даними кореляційної таблиці побудуємо ряд, що відображає залежність середнього значення *Y* від *X* (табл. 3.2), для чого знайдемо середні значення $\overline{y}_{x_i}$ для кожного значення x_i, $i = \overline{1,5}$ і заповнимо табл. 4.8:

$$\overline{y}_{x_1} = \frac{y_1 n_{11} + y_2 n_{12} + y_3 n_{13} + y_4 n_{14}}{n_1} = \frac{20 \cdot 1 + 25 \cdot 2}{3} \approx 23{,}33;\ \overline{y}_{x_2} = \frac{15 \cdot 2 + 20 \cdot 2}{4} = 17{,}5;$$

$$\overline{y}_{x_3} = \frac{15 \cdot 5 + 20 \cdot 3 + 25 \cdot 1}{9} = 17{,}78;\ \overline{y}_{x_4} = \frac{10 \cdot 1 + 15 \cdot 3}{4} = 13{,}75;$$

$$\overline{y}_{x_5} = \frac{10 \cdot 3 + 15 \cdot 1 + 20 \cdot 1}{5} = 13.$$

Таблиця 4.8

x_i	25	50	75	100	125
$\overline{y}_{x_i}$	23,33	17,5	17,78	13,75	13
n_i	3	4	9	4	5

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.4).

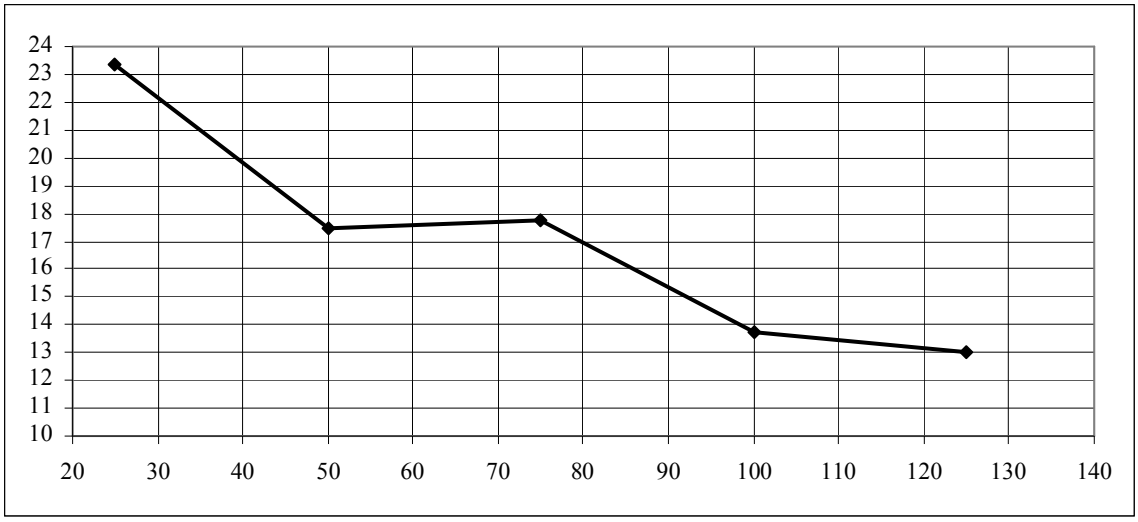

Рисунок 4.4. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до гіперболи, то висуваємо гіпотезу про гіперболічну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $\overline{y}_x = \frac{a}{x} + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему для згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.9), в останньому рядку якої знайдемо відповідні стовпцям суми.

Таблиця 4.9

x_i	$\overline{y}_{x_i}$	n_i	$\frac{1}{x_i}$	$\frac{1}{x_i} n_i$	$\frac{1}{x_i^2} n_i$	$\overline{y}_{x_i} n_i$	$\frac{1}{x_i} \overline{y}_{x_i} n_i$
25	23,33	3	0,04	0,12	0,0048	69,99	2,7996
50	17,5	4	0,02	0,08	0,0016	70	1,4
75	17,78	9	0,0133	0,12	0,0016	160,02	2,1336
100	13,75	4	0,01	0,04	0,0004	55	0,55
125	13	5	0,008	0,04	0,0003	65	0,52
			Суми	0,4	0,0087	420,01	7,4032

Отже, складемо систему для знаходження параметрів рівняння регресії та розв’яжемо її за правилом Крамера:

$$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y_{x_i}}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i \end{cases} \Rightarrow \begin{cases} 0,0087a + 0,4b = 7,4032 \\ 0,4a + 25b = 420,01 \end{cases}.$$

Головний визначник системи: $\Delta = \begin{vmatrix} 0,00872 & 0,4 \\ 0,4 & 25 \end{vmatrix} = 0,00872 \cdot 25 - 0,4^2 = 0,058$.

Допоміжні визначники: $\Delta a = \begin{vmatrix} 7,4032 & 0,4 \\ 420 & 25 \end{vmatrix} = 7,4032 \cdot 25 - 0,4 \cdot 420 = 17,076$;

$$\Delta b = \begin{vmatrix} 0,00872 & 7,4032 \\ 0,4 & 420 \end{vmatrix} = 0,00872 \cdot 420 - 7,4032 \cdot 0,4 = 0,701207.$$

Формули Крамера: $a = \frac{\Delta a}{\Delta} = \frac{17,076}{0,058} \approx 294,41$; $b = \frac{\Delta b}{\Delta} = \frac{0,7012207}{0,058} \approx 12,09$.

Отже, шукане рівняння регресії має вигляд $\overline{y}_x = \frac{294,41}{x} + 12,09$.

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для цього знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.10).

Знайдемо $\overline{y}$: $\overline{y} = \frac{\sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i}{n} \approx 16,8$.

Таблиця 4.10

x_i	y_i	n_i	$y_{i,\text{теор}}$	$\left(y_i - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - y_i\right)^2 n_i$
25	23,33	3	23,866	127,907	149,782	0,863
50	17,5	4	17,978	1,958	5,547	0,914
75	17,78	9	16,015	8,637	5,547	28,028
100	13,75	4	15,034	37,220	12,482	6,594
125	13	5	14,445	72,215	27,737	10,441
			Суми	247,936	201,096	46,840

Отже, $Q = 247,936$; $Q_p = 201,096$; $Q_o = 46,840$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $247,936 = 201,096 + 46,840$ і є тотожністю, тому рівняння регресії побудовано правильно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 25$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p(n-l)}{Q_o(l-1)} = \frac{201{,}096(25-2)}{46{,}840(2-1)} \approx 98{,}75.$$

Знайдемо $F_{кр}$: $F_{кр} = FРАСПОБР(0{,}001; 2-1; 25-2) \approx 14{,}20$. Розраховане значення F-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2: $R^2 = \frac{Q_p}{Q} = \frac{201{,}096}{247{,}936} \approx 0{,}81$. Значення коефіцієнта детермінації свідчить, що 81% варіації результативної ознаки Y пояснюється рівнянням регресії.

Висновок: Залежність собівартості одиниці продукції Y (грн.) від об'єму виробництва X (тис. од.) описується рівнянням $\overline{y_x} = \frac{294{,}41}{x} + 12{,}09$, яке є статистично значущим на рівні значущості 0,001 та описує 81% вибіркових даних.

4.4. Множинна лінійна регресія

У процесі аналізу діяльності економічного або соціального об'єкта часто виявляється, що на результативну ознаку цієї діяльності (наприклад, об'єм валової продукції, об'єм продаж, думку респондента відносно певного об'єкта та ін.) впливає декілька факторних ознак: час, вартість сировини і матеріалів, якість обладнання, продуктивність праці, соціальні установки, вплив зовнішніх і внутрішніх факторів та інше. Тоді як модель діяльності об'єкта використовують багатофакторну лінійну регресійну модель, на основі якої розробляються прогнози діяльності, вивчається вплив на діяльність різноманітних показників і виявляються ті показники, покращення яких суттєво збільшує її кінцевий продукт.

Загальний вигляд багатофакторної лінійної регресійної моделі:

$$Y = f(X_1, X_2, \ldots, X_m) + \varepsilon, \qquad (4.7)$$

де Y – результативна ознака,
$X_1, X_2, \ldots, X_m$ – факторні ознаки,
m – кількість факторних ознак,
ε – випадкова похибка моделі.

Зауваження 1. Задачі побудови багатофакторної регресійної моделі розв'язуються за умов, коли випадкова похибка ε має нормальний розподіл із нульовим математичним сподіванням, а випадкові похибки кожного вимірювання незалежні та мають однакові дисперсії. Кількість спостережень n повинна перевищувати величину $3(m + 1)$.

Крім того, для забезпечення статистичної значущості моделі необхідно дотримуватися основного правила її побудови: **„Факторні ознаки, які включено у модель, повинні бути тісно пов'язані із результативною ознакою і слабо пов'язані (або не мати зв'язку) між собою"**.

Тіснота зв'язку між результативною і факторними ознаками та зв'язку факторних ознак між собою визначається за аналізом парних і частинних коефіцієнтів кореляції (див. п. 3.5). В модель бажано включати тільки ті ознаки, що не мають статистично значущого зв'язку між собою, хоча й вважається, що сильний зв'язок між ними, зазвичай, не впливає на якість прогнозу за моделлю.

Якість моделі визначається за критерієм Фішера, тобто порівнянням статистики F моделі із критичним значенням $F_{\text{кр}}$, де $F_{\text{кр}}(\alpha, k_1, k_2)$ – табличне значення розподілу Фішера, що знаходиться за умов: $\alpha = 0{,}05;\quad k_1 = m - 1;\quad k_2 = n - m$. Якщо $F > F_{кр}$, то модель є достовірною на рівні значущості 0,05 (тобто 95% даних пояснюються побудованою моделлю, 5% – випадкові помилки моделі).

Відносну величину впливу факторних ознак на результативну можна оцінити за формулою:

$$r_{X_i}^{\,2} = \frac{t_{X_i}^{\,2} \cdot R^2}{\sum_{i=1}^{m} t_{X_i}^{\,2}}; \tag{4.8}$$

де $t_{X_i}^{\,2}$ – розраховане значення розподілу Стьюдента для ознаки X_i;

R^2 – загальний коефіцієнт детермінації моделі.

Обчислення, які необхідно провести для побудови багатофакторної регресійної моделі, дуже складні, але застосування засобу Excel ***Регрессия*** пакета ***Анализ данных*** значно полегшує цю роботу.

За допомогою засобу ***Регрессия*** отримують такі результати:

− параметри лінійної регресійної моделі виду $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$, де b_0, b_1, b_2, …, b_m – параметри моделі;

− коефіцієнт детермінації;

− критеріальну статистику для перевірки статистичної значущості моделі;

− теоретичні значення результативної ознаки, отримані за побудованою моделлю та залишки – тобто різниці між теоретичними та емпіричними значеннями цієї ознаки.

Зауваження 2. Засобом ***Регрессия*** можна також користуватися при побудові моделі, нелінійної відносно певної факторної ознаки X_j, але лінійної відносно коефіцієнта цієї ознаки. Наприклад, якщо необхідно побудувати модель виду $Y = b_0 + b_1 X_1 + b_2 X_2^3$, то як вхідні дані вказують трійки (y_i, x_{1i}, x_{2i}^3).

4.5. Регресія у Microsoft Excel

Пакет аналізу даних Microsoft Excel надає можливість будувати регресійні моделі, але тільки у випадку лінійної залежності результативної ознаки Y від факторної ознаки X і тільки для незгрупованих вибіркових даних.

Для побудови лінійної регресійної моделі необхідно:

1) Викликати ***Сервис – Анализ данных – Регрессия – ОК***. З'явиться вікно для надання вхідних даних (рис. 4.5).

2) У графі ***Входной интервал Y*** та ***Входной интервал X*** вказати відповідні стовпці даних; у графі ***Выходной интервал*** вказати ту чарунку, починаючи з якої будуть надаватися вихідні дані – параметри рівняння регресії та результати її статистичного аналізу.

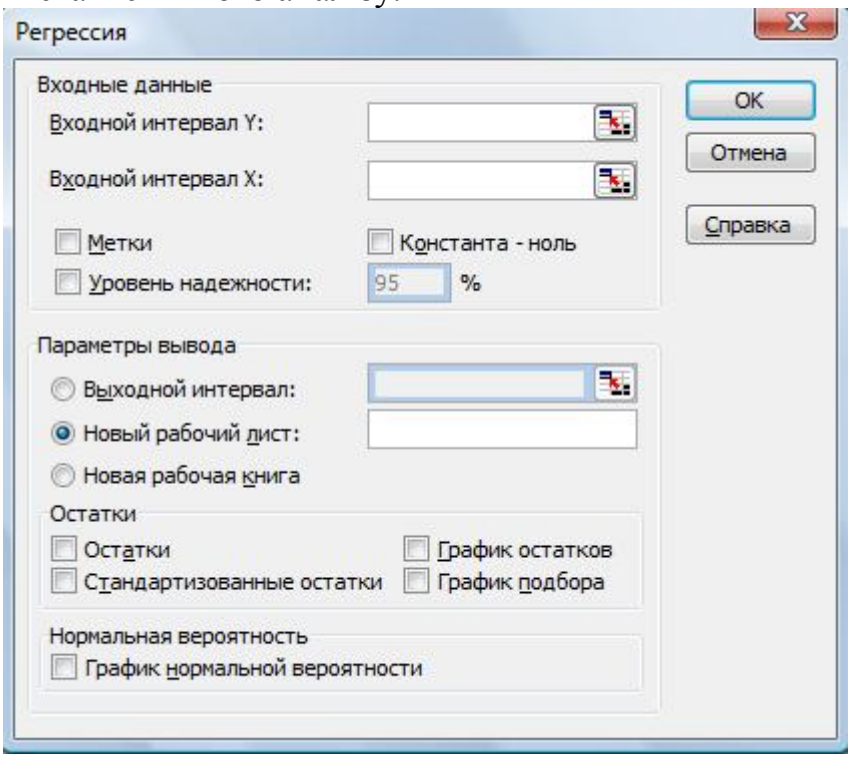

Рисунок 4.5. Діалогове вікно функції ***Регрессия***

Приклад і результати роботи функції ***Регрессия*** представлено на рис. 4.6.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

I23

	A	B	C	D	E	F	G	H	I	J
1					**Регресійний аналіз**					
2		Вхідні дані			Вихідні дані					
3	№	Значення			ВЫВОД ИТОГОВ					
4	**i**	**X**	**Y**							
5	1	1	328		*Регрессионная статистика*					
6	2	2	329		Множественный R	0,972633354				
7	3	3	329		R-квадрат	0,946015642				
8	4	4	345		Нормированный R-квадрат	0,937018249				
9	5	5	352		Стандартная ошибка	5,786032717				
10	6	6	370		Наблюдения	8				
11	7	7	377							
12	8	8	385		Дисперсионный анализ					
13						*df*	*SS*	*MS*	*F*	*Значимость F*
14					Регрессия	1	3520,005952	3520,005952	105,1433059	5,01935E-05
15					Остаток	6	200,8690476	33,4781746		
16					Итого	7	3720,875			
17										
18						*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	
19					Y-пересечение	310,6785714	4,508440371	68,91043151	6,28282E-10	
20					Переменная X 1	9,154761905	0,892804231	10,25394099	5,01935E-05	

Рисунок 4.6. Результати регресійного аналізу

В таблиці (рис. 4.6) у графі ***Коэффициенты*** вказані значення параметрів моделі a та b: b − в графі ***Y-пересечение***, a − в графі ***Переменная X1***. Отже, побудована лінійна регресійна модель має вигляд:

$$y = 69{,}15x + 310{,}68\,.$$

Для перевірки статистичної значущості моделі надається значення F-статистики у графі F: F = 105,14. Це значення обчислюється як відношення варіації регресії до варіації залишків (чарунки H14 та H15). В стовпці ***Значимость F*** надано критеріальну статистику. Якщо це значення менше, ніж, наприклад, 0,05, то рівняння регресії є значущим на рівні 0,05. У даному завданні рівняння регресії є значущим на рівні 0,00005.

У графі ***Множественный R-квадрат*** надано значення множинного коефіцієнта кореляції, який показує силу залежності результативної ознаки від факторної (або декілька факторних ознак). У нашому випадку він дорівнює 0,97, що означає сильний зв'язок між Y та X.

Коефіцієнт детермінації моделі R^2 виводиться у графі ***R-квадрат***, R^2= 0,97, тобто 97% даних описується рівнянням регресії.

Крім того, можна задати: графік підбору – порівняльна діаграма, що містить емпіричну і теоретичну лінії регресії; таблиця залишків – різниць емпіричних і теоретичних значень Y (рис. 4.7).

Регресійний аналіз

Вхідні дані

№	Значення	
i	X	Y
1	1	328
2	2	329
3	3	329
4	4	345
5	5	352
6	6	370
7	7	377
8	8	385

Вихідні дані

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки
1	319,8333333	8,166666667
2	328,9880952	0,011904762
3	338,1428571	-9,142857143
4	347,297619	-2,297619048
5	356,452381	-4,452380952
6	365,6071429	4,392857143
7	374,7619048	2,238095238
8	383,9166667	1,083333333

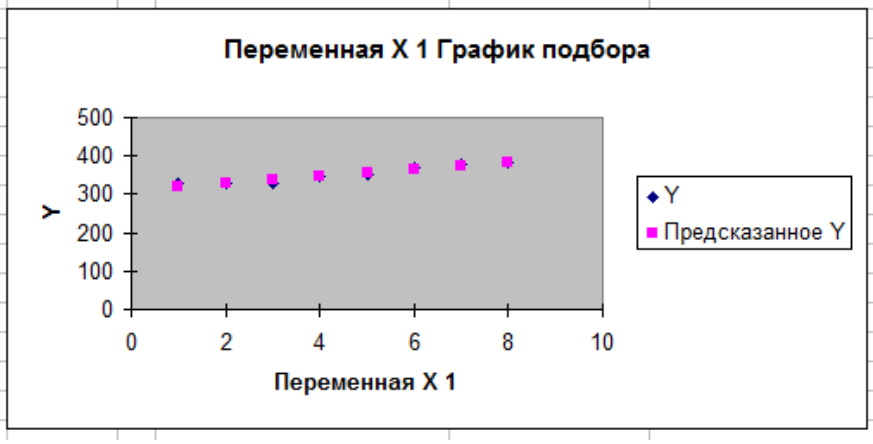

Рисунок 4.7. Додаткові результати регресійного аналізу

Розглянемо детальніше методику побудови множинної лінійної регресійної моделі засобами Microsoft Excel.

Приклад 4.3. В таблиці 4.11 вказано дані по консервному заводу с. Виноградне Одеської області за 12 місяців минулого року (табл. 4.11).

Умовні позначення:

X_1 – часовий фактор, порядковий номер місяця;

X_2 – фонди (тис. грн./робітника);

X_3 – фондовіддача (тис. грн. обсягу товарного продукту/тис. грн. основного фонду);

X_4 – продуктивність праці (тис. умовних банок/робітника);

Y – валова продукція (тис. умовних банок).

Таблиця 4.11

X_1	X_2	X_3	X_4	Y
1	328	0,054	0,3	397
2	329	0,101	0,6	670
3	329	0,099	1,2	1209
4	347	0,019	0,1	138
5	352	0,065	0,3	378
6	370	0,053	0,1	79
7	378	0,178	2,3	1883
8	385	0,174	2,6	2124
9	396	0,298	5,5	5069
10	399	0,195	2,4	2618
11	390	0,102	1,6	1265
12	378	0,138	0,6	562

Розробляється проект модернізації заводу, для чого необхідно: побудувати багатофакторну лінійну регресійну модель діяльності заводу; визначити вплив факторних ознак на об'єм валової продукції; виявити найвпливовіші ознаки для визначення напрямків майбутньої модернізації.

Розв'язок. За основним правилом побудови множинної регресійної моделі розв'язок задачі складається з трьох етапів: виявлення факторних ознак, які необхідно включити в модель; побудова моделі; аналіз якості моделі.

Етап 1. Виявимо факторні ознаки, що включаються в модель. Для чого:

− розрахуємо парні коефіцієнти кореляції; побудуємо кореляційну матрицю і проведемо її статистичний аналіз;

− на основі результатів статистичного аналізу побудуємо кореляційні плеяди і виявимо ознаки, які необхідно включити в модель;

− у разі необхідності (тобто у випадку існування неявного зв'язку між факторними ознаками) розрахуємо частинні коефіцієнти кореляції та проведемо їх статистичний аналіз.

Парні коефіцієнти кореляції обчислимо за допомогою вбудованих сервісних функцій Excel: перенесемо табл. 4.11 на сторінку Excel, викличемо ***Сервис – Аналіз данных – Корреляция – ОК***. У графі ***Входной интервал***

вкажемо масив даних табл. 4.11; у графі ***Группирование*** вкажемо ***По столбцам***, у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – парні коефіцієнти кореляції. Отримаємо табл. 4.12, яка є матрицею парних коефіцієнтів кореляції. Чарунки таблиці, розташовані вище головної діагоналі, незаповнені, оскільки таблиця симетрична відносно головної діагоналі.

Таблиця 4.12

	X_1	X_2	X_3	X_4	Y
X_1	1,00				
X_2	0,89	1,00			
X_3	0,56	0,69	1,00		
X_4	0,46	0,66	0,94	1,00	
Y	0,43	0,63	0,94	0,99	1,00

Розрахуємо критичне значення коефіцієнта кореляції $r_{кр}$ за формулою: $r_{кр} = \frac{t_{\alpha,k}}{\sqrt{t_{\alpha,k}^{\ 2} + n - 2}}$, α – рівень значущості, $\alpha = 0,05$; $t_{\alpha,k}$ знайдемо за допомогою вбудованої функції Excel. Викличемо ***Функции – Статистические – СТЬЮДРАСПОБР – Ок***. В графі ***Вероятность*** вкажемо 0,05 (рівень значущості); в графі ***Степени свободы*** вкажемо значення $n - 2 = 12 - 2 = 10$. Отримаємо $t_{\alpha,k} = 2,228$. Тоді:

$$r_{кр} = \frac{t_{\alpha,k}}{\sqrt{t_{\alpha,k}^{\ 2} + n - 2}} = \frac{2,228}{\sqrt{2,228^2 + 12 - 2}} \approx 0,57598.$$

Доповнимо табл. 4.12. Виділимо в ній елементи, які більші за $r_{кр}$ (це означає, що відповідні ознаки тісно пов'язані між собою). Отримаємо табл. 4.13.

Таблиця 4.13

	X_1	X_2	X_3	X_4	Y
X_1	1,00	**0,89**	0,56	0,46	0,43
X_2	**0,89**	1,00	**0,69**	**0,66**	**0,63**
X_3	0,56	**0,69**	1,00	**0,94**	**0,94**
X_4	0,46	**0,66**	**0,94**	1,00	**0,99**
Y	0,43	**0,63**	**0,94**	**0,99**	1,00

Отже, тісно пов'язані між собою такі факторні ознаки:

X₁ та ***X₂*** оскільки $r(X_1, X_2) = 0,89 > 0,57598$;

X₂ та ***X₃*** оскільки $r(X_2, X3) = 0,69 > 0,57598$;

X₂ та ***X₄*** оскільки $r(X_2, X_4) = 0,66 > 0,57598$;

X₃ та ***X₄*** оскільки $r(X_3, X_4) = 0,94 > 0,57598$.

З факторних ознак тісно пов'язані із результативною ознакою (із ***У***):

X₂ оскільки $r(X_2, Y) = 0,63 > 0,57598$;

X₃ оскільки $r(X_3, Y) = 0,94 > 0,57598$;

X₄ оскільки $r(X_4, Y) = 0,99 > 0,57598$.

За результатами аналізу кореляційної матриці побудуємо кореляційні плеяди, тобто зобразимо достовірний зв'язок між факторними ознаками графічно (рис. 4.8).

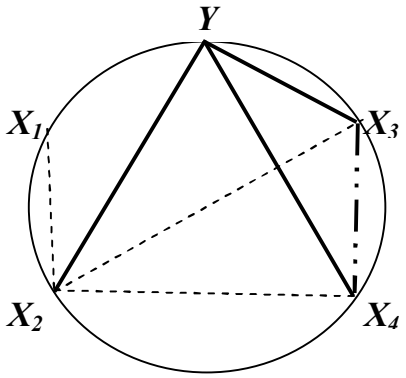

Рис.4.8. Кореляційний зв'язок між факторними ознаками

Перша кореляційна плеяда: ***Y, X₂ , X₃ , X₄*** вказує, що в модель необхідно включити ознаки ***X₂ , X₃ , X₄***, оскільки вони мають зв'язок (тобто впливають) на результативну ознаку ***Y***. Тобто із всіх факторних ознак в модель не потрібно включати ознаку ***X₁***.

Друга кореляційна плеяда: ***X₂, X₁ , X₃ , X₄*** вказує, що в модель можна включити тільки одну з ознак ***X₂, X₁ , X₃ , X₄***, оскільки вони пов'язані між собою. Однак наявність дуже сильного (0,94) зв'язку між ***X₃*** та ***X₄*** свідчить про те, що зв'язок може існувати між ***X₂*** та ***X₃*** , а між ***X₂*** та ***X₄*** він може бути тільки наслідком зв'язку ***X₃ − X₄***. Або навпаки, зв'язок може існувати між ***X₂*** та ***X₄*** , а між ***X₂*** та ***X₃*** він може бути тільки наслідком зв'язку ***X₃ − X₄***. Тому, можливо, в модель потрібно включати ***X₂*** та одну із ознак ***X₃*** та ***X₄***. Для того, щоб виявнити це, скористуємось частинними коефіцієнтами кореляції.

Розрахуємо частинні коефіцієнти кореляції між факторними ознаками ***X₂*** і ***X₃***, та між ***X₂*** і ***X₄***. Величина цих частинних коефіцієнтів кореляції дозволить визначити „чистий" зв'язок між вказаними ознаками, тобто зв'язок, що не залежить від впливу всіх останніх факторних ознак.

Частинний коефіцієнт кореляції між факторними ознаками ***X₂*** і ***X₃*** розраховуємо за формулою:

$$R_{23} = \frac{-A_{23}}{\sqrt{A_{22} A_{33}}},$$

де A_{23} – алгебраїчне доповнення елемента r_{23},

A_{22} – алгебраїчне доповнення елемента r_{22},

A_{33} – алгебраїчне доповнення елемента r_{33}.

Алгебраїчне доповнення A_{23} – це визначник матриці, отриманої із матриці A викреслюванням 2-го рядка і 3-го стовпця, помножений на $(-1)^{2+3}$. Аналогічно A_{22} – це визначник матриці, отриманої із матриці A викреслюванням 2-го рядка і 2-го стовпця, помножений на $(-1)^{2+2}$; A_{33} – це визначник матриці, отриманої із матриці A викреслюванням 3-го рядка і 3-го стовпця, помножений на $(-1)^{3+3}$. Визначники обчислюємо за допомогою вбудованих функцій Excel: викликаємо ***Функции – Математические –***

МОПРЕД, у графі ***Массив*** вказуємо матрицю, визначник якої потрібно знайти. Отримаємо:

$$A_{23}=\begin{vmatrix}1 & 0,89 & 0,46 & 0,43\\ 0,56 & 0,69 & 0,94 & 0,94\\ 0,46 & 0,66 & 1 & 0,99\\ 0,43 & 0,63 & 0,99 & 1\end{vmatrix}=0,00028; \quad A_{22}=\begin{vmatrix}1 & 0,56 & 0,46 & 0,43\\ 0,56 & 1 & 0,94 & 0,94\\ 0,46 & 0,94 & 1 & 0,99\\ 0,43 & 0,94 & 0,99 & 1\end{vmatrix}=0,001005;$$

$$A_{33}=\begin{vmatrix}1 & 0,89 & 0,46 & 0,43\\ 0,89 & 1 & 0,66 & 0,63\\ 0,46 & 0,66 & 1 & 0,99\\ 0,43 & 0,63 & 0,99 & 1\end{vmatrix}=0,001442; \quad R_{23}=\frac{-A_{23}}{\sqrt{A_{22}A_{33}}}=\frac{-0,0028}{\sqrt{0,001005\cdot 0,001442}}\approx -0,24;$$

$$A_{24}=\begin{vmatrix}1 & 0,89 & 0,56 & 0,43\\ 0,56 & 0,69 & 1 & 0,94\\ 0,46 & 0,66 & 0,94 & 0,99\\ 0,43 & 0,63 & 0,94 & 1\end{vmatrix}=-0,00041; \quad A_{44}=\begin{vmatrix}1 & 0,89 & 0,56 & 0,43\\ 0,89 & 1 & 0,69 & 0,63\\ 0,56 & 0,69 & 1 & 0,94\\ 0,43 & 0,63 & 0,94 & 1\end{vmatrix}=0,008578;$$

$$R_{24}=\frac{-A_{24}}{\sqrt{A_{22}A_{44}}}=\frac{0,00041}{\sqrt{0,001005\cdot 0,008578}}\approx 0,1411.$$

Оскільки $|R_{23}|>R_{24}$, то в модель необхідно включати факторну ознаку ***X_4.***

Висновок з етапу 1: шукана багатофакторна лінійна регресійна модель має вигляд: $Y=b_0+b_1X_2+b_2X_4$.

Етап 2. Побудуємо вказану модель. Для знаходження b_0, b_1, b_2 викликаємо ***Сервис – Анализ данных – Регрессия – ОК***. У графі ***Входной интервал Y*** вкажемо відповідний стовпчик даних табл. 4.11; у графі ***Входной интервал X*** вкажемо стовпчики ***X_2*** та ***X_4*** табл 4.11; у графі ***Выходной интервал*** вкажемо ту чарунку, починаючи з якої будуть виводитися вихідні дані – рівняння регресії. Отримаємо таблицю з результатами регресійного аналізу (рис. 4.9).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный	0,992676928
R-квадрат	0,985407483
Нормированный	0,982164702
Стандартная оши	191,3107107
Наблюдения	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	22243684,82	11121842,41	303,8772358	5,47754E-09
Остаток	9	329398,0921	36599,78801		
Итого	11	22573082,92			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	657,6665422	998,224854	0,658836072	0,526495682	-1600,474958	2915,808042	-1600,474958	2915,808042
Переменная X 1	-1,760422463	2,859983	-0,615535881	0,553445728	-8,230154606	4,709309679	-8,230154606	4,709309679
Переменная X 2	927,0473673	48,915520	18,9520088	1,4588E-08	816,3927733	1037,701961	816,3927733	1037,701961

Рисунок 4.9. Результати регресійного аналізу

В таблиці (рис. 4.9) у графі ***Коэффициенты*** вказані значення параметрів моделі b_0, b_1, b_2: b_0 − в графі ***Y*-пересечение**, b_1 − в графі **Переменная *X1***, b_2 − в графі **Переменная *X2*.** Отже, b_0=657,67; b_1= −1,76; b_2=927,05; багатофакторна лінійна регресійна модель має вигляд:

$$Y = 657,67 - 1,76X_2 + 927,05X_4 .$$

Етап 3. Перевіримо якість побудованої моделі. Скористуємось результатами регресійного аналізу (рис. 4.9).

Перевіримо статистичну значущість моделі. Значення *F*-статистики моделі подано в таблиці у графі *F*: *F* = 303,877. Критичне значення $F_{кр}$ знайдемо за допомогою статистичної функції Excel $FРАСПОБР(\alpha, k_1, k_2)$, де

$\alpha = 0,05;\quad k_1 = m - 1 = 2 - 1 = 1;\quad k_2 = n - m = 12 - 2 = 10$.

Отже, $F_{кр}$=4,96; $F > F_{кр} \Rightarrow$ рівняння регресії є значущим, модель є достовірною на рівні значущості 0,05. Крім того, в стовпчику ***Значимость F*** є критеріальна статистика, яка показує, що рівняння регресії є значущим на рівні 0,000000005.

У графі ***Множественный R-квадрат*** подано значення множинного коефіцієнта кореляції – 0,99, який показує сильну залежність результативної ознаки від обраних факторних ознак. У графі ***R-квадрат*** бачимо коефіцієнт детермінації моделі R^2=0,985, тобто 98,5% даних описуються рівнянням регресії.

Для наочності висновків зобразимо емпіричну і теоретичну лінії регресії (рис. 4.10).

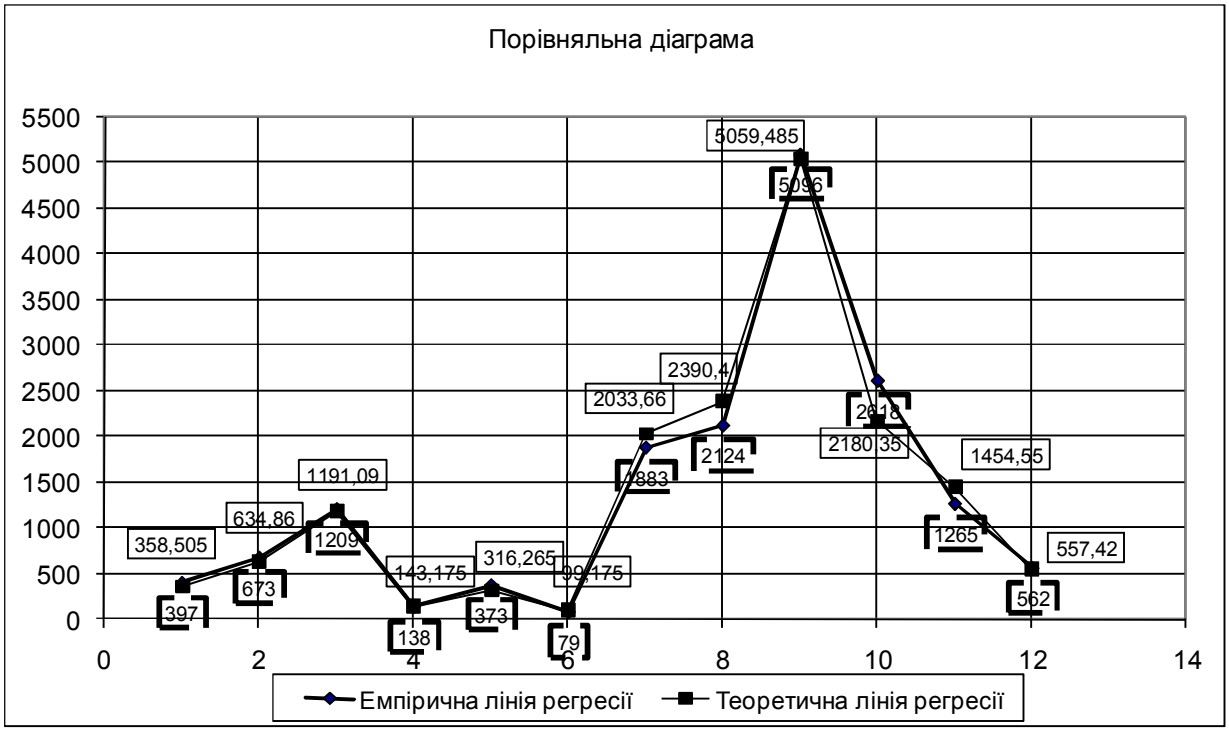

Рисунок 4.10. Порівняльна діаграма за результатами регресійного аналізу

Висновок: на об'єм валової продукції ***Y*** значно впливають факторні ознаки X_2 – фонди (тис. грн./робітника) та X_4 – продуктивність праці (тис. умовних банок/робітника), що й визначає напрями модернізації заводу.

4.6. Регресійний аналіз засобами SPSS

Щоб знайти та дослідити рівняння лінії регресії, варто побудувати емпіричну лінію регресії та визначити її вид. За допомогою SPSS можна вивчати кілька видів регресійного аналізу, які зображені на рис. 4.11.

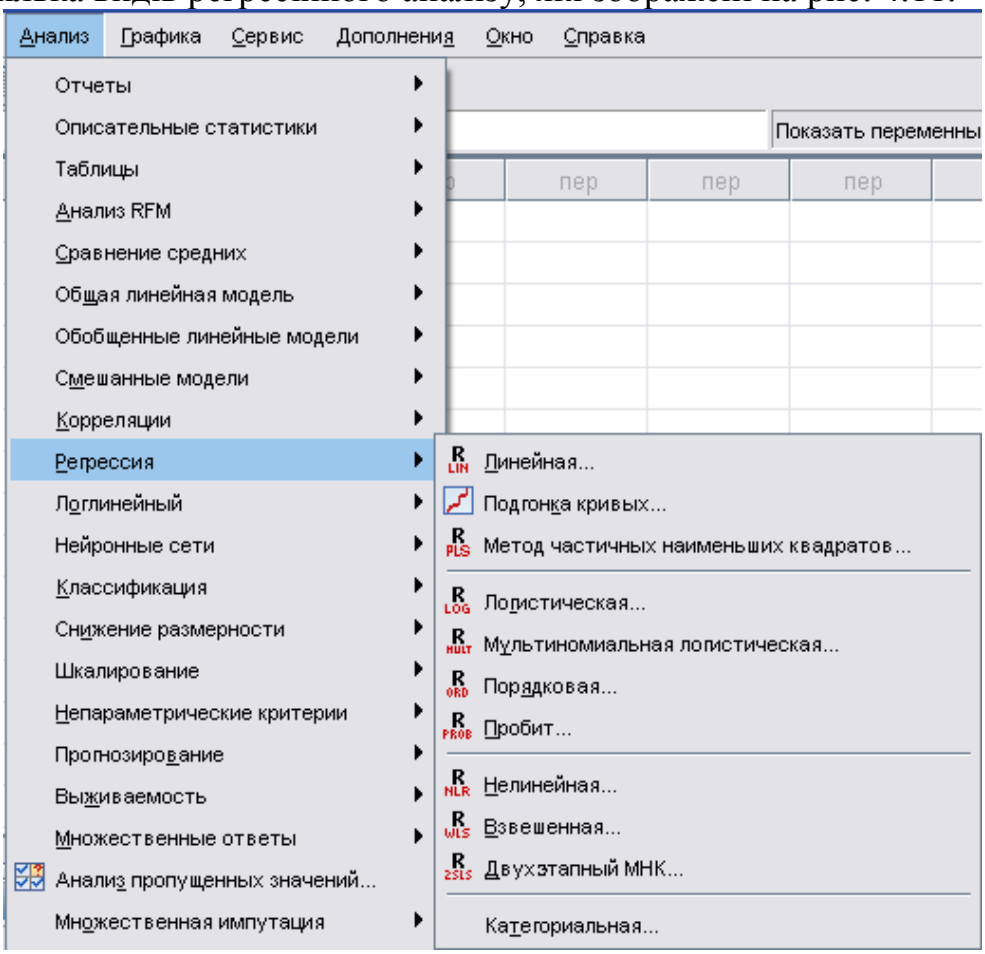

Рисунок 4.11. Види регресій у SPSS

4.6.1. Лінійна регресія

Знайдемо рівняння лінії регресії та побудуємо регресійну пряму, які характеризують залежність сумарних виробничих затрат Y (тис. грн.) від обсягів виробництва X (тис. од.) за статистичними даними, представленими у табл. 4.2 із прикладу 4.1.

1) Введемо стрічкові дані табл. 4.2 у два стовпчики вкладки *Набор данных* і побудуємо регресійну пряму. Виберемо в меню ***Графика*** – ***Рассеяния/точки***, відкриється діалогове вікно у якому виберемо вид ***Простая диаграмма рассеяния*** і натиснемо ***Задать***. На вісь OX перенесемо змінну X, на OY – змінну Y.

Отримаємо графік (рис. 4.12), який дає підстави спрогнозувати лінійну залежність Y від X. Отже, вивчаємо лінійну регресію.

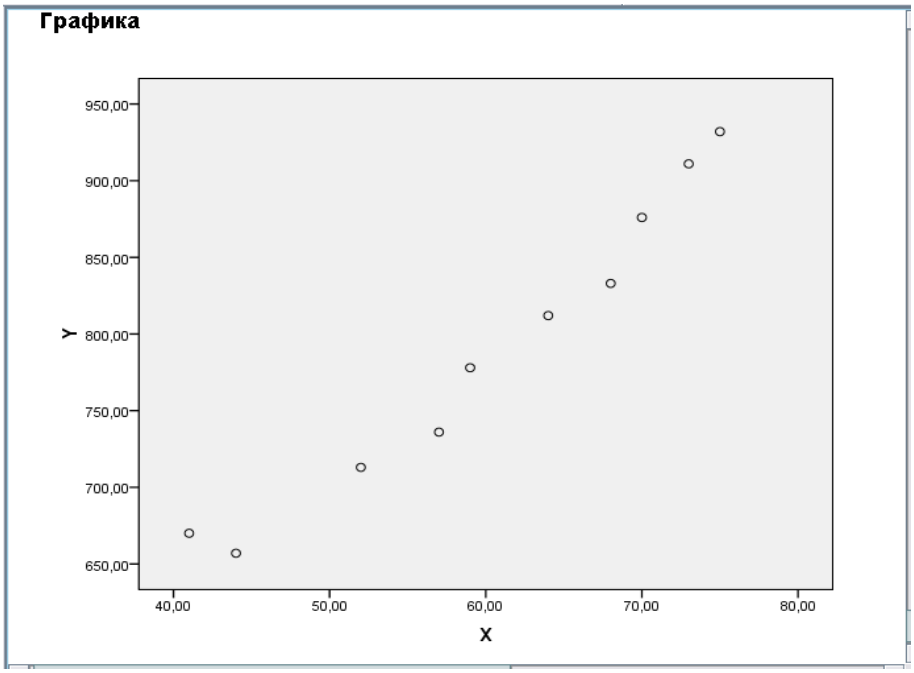

Рисунок 4.12. Графік емпіричної лінії регресії засобами SPSS

2) Виберемо в меню послідовно ***Анализ − Регрессия − Линейная***. У діалоговому вікні ***Линейная регрессия*** перенесемо змінну X у поле ***Независимые переменные***, а змінну Y − в поле ***Зависимые переменные*** (рис. 4.13) та перейдемо у вікно виведення результатів;

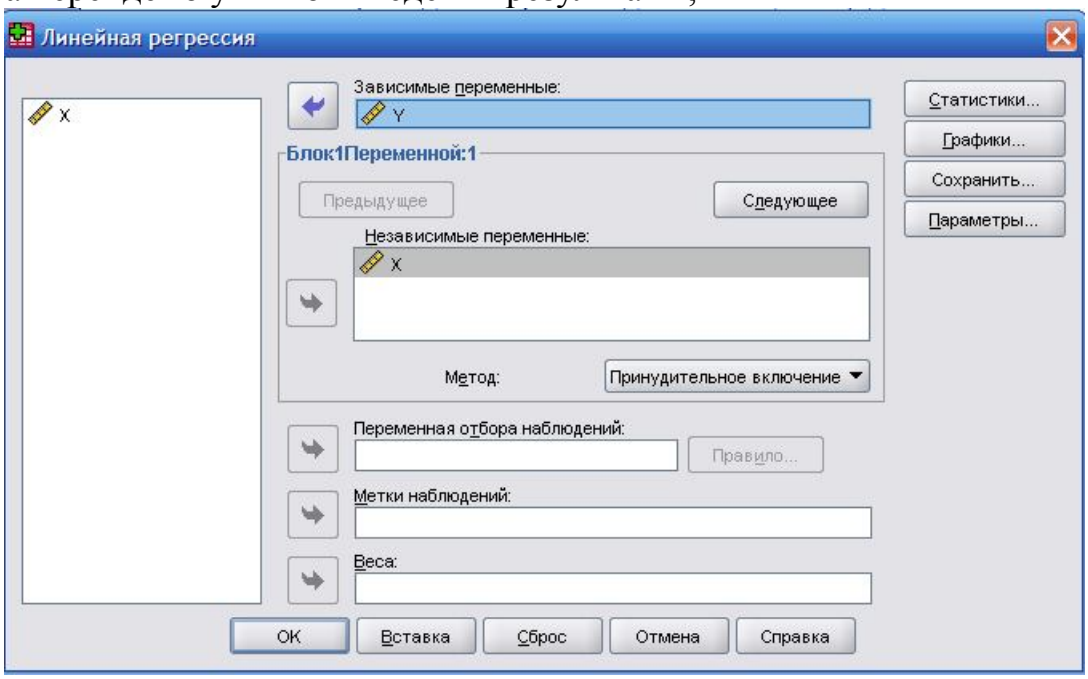

Рисунок 4.13. Діалогове вікно вибору параметрів лінійної регресії

3) В таблиці ***Сводка для модели*** (рис. 4.14) під назвою ***R-квадрат*** зберігається значення коефіцієнта детермінації 0,959, який свідчить про те, що рівняння регресії описує 95,9% вибіркових даних. У таблиці ***Коэффициенты*** в першому стовпчику знаходяться коефіцієнти рівняння регресії, а саме: $y = 8,059x - 305,832$.

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,979[a]	,959	,954	20,84459

a. Предикторы: (конст) X

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	82103,626	1	82103,626	188,963	,000[a]
	Остаток	3475,974	8	434,497		
	Всего	85579,600	9			

a. Предикторы: (конст) X

b. Зависимая переменная: Y

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	305,832	35,962		8,504	,000
	X	8,059	,586	,979	13,746	,000

a. Зависимая переменная: Y

Рисунок 4.14. Результати розрахунку рівняння лінійної регресії

4.6.2. Нелінійна регресія

Знайдемо рівняння лінії регресії та побудуємо регресійну пряму, які характеризують залежність собівартості одиниці продукції Y (грн.) від обсягів виробництва X (тис. од.) за статистичними даними, представленими у табл. 4.6 із прикладу 4.2.

Для зручності, скористаємось елементами вище проведеного дослідження, де, згідно графіка, припускається, що емпірична лінія регресії наближається до гіперболи і висувається гіпотеза про гіперболічну залежність Y від X:

$$\overline{y_x} = \frac{a}{x} + b\,.$$

Введемо стрічкові дані таблиці 4.8 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

1) Виберемо в меню послідовно ***Анализ*** − ***Регрессия*** − ***Нелинейная***. У діалоговому вікні ***Нелинейная регрессия*** (рис. 4.15) перенесемо змінну Y у поле ***Зависимые переменные***, у полі ***Выражение, задающее модель*** задаємо вираз: $a/X+b$ та активуємо кнопку ***Параметры***, перейшовши у вікно ***Нелинейная регрессия***: ***Параметры*** (рис. 4.15);

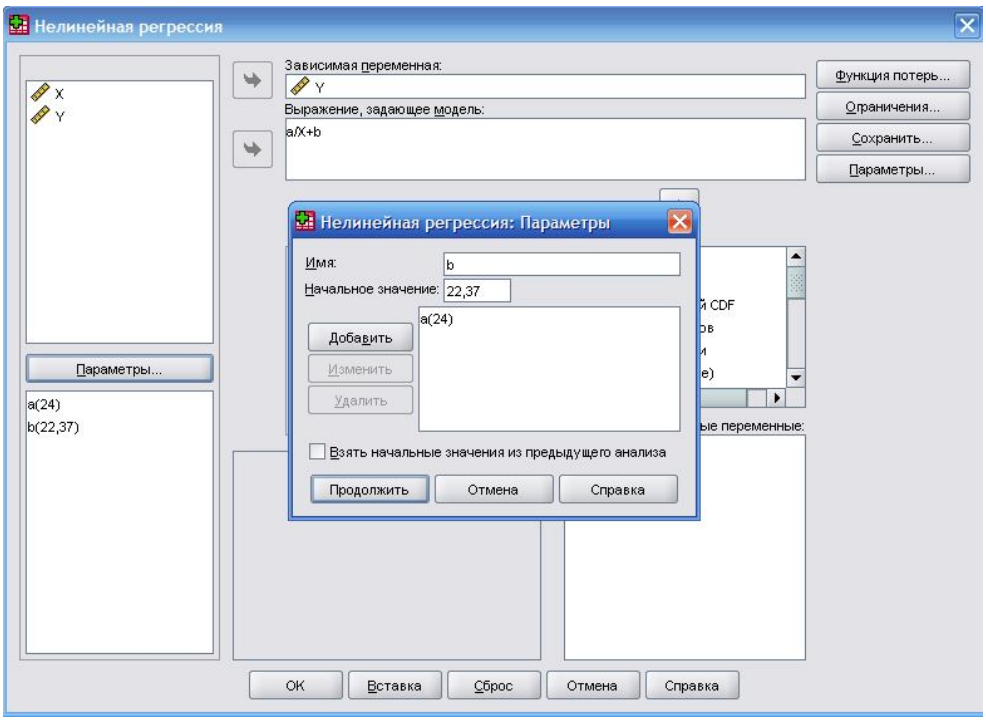

Рисунок 4.15. Діалогове вікно вибору параметрів нелінійної регресії

2) В полі ***Имя*** (діалогове вікно на першому плані рис. 4.15) введемо a, його початкове значення має «покривати» максимальне значення y_i із табл. 4.8. Тому вибираємо значення − 24. Натискаємо ***Добавить*** і задаємо значення b − 22,37 (так як $x_1 = 25,\ a = 24,\ y_1 = 23,33$ і $b = y_1 - \frac{a}{x_1}$). Переходимо у вікно виведення результатів (рис. 4.16).

Оценки параметра

Параметр	Оценка	Стд. Ошибка	Доверительный интервал 95 %	
			Нижняя граница	Верхняя граница
a	300,739	56,971	119,432	482,045
b	11,579	1,233	7,655	15,502

Корреляции оценок параметров

	a	b
a	1,000	-,844
b	-,844	1,000

Дисперсионный анализ[a]

Источник	Сумма квадратов	ст.св.	Средние квадраты
Регрессия	1518,173	2	759,086
Остаток	6,557	3	2,186
Нескорректированный итог	1524,730	5	
Скорректированный итог	67,464	4	

Зависимая переменная: Y

a. R в квадрате = 1 - (остаточная сумма квадратов) / (скорректированная сумма квадратов) = ,903.

Рисунок 4.16. Результати розрахунку рівняння параболічної регресії

Із першого стовпчика таблиці ***Оценки параметра*** знаходимо коефіцієнти і складаємо рівняння: $y = \frac{300,7}{x} + 11,6$, яке описує 90,3% вибіркових даних, про що свідчить коефіцієнт детермінації.

Якщо гіпотетично неможливо зробити припущення про вид лінії регресії, то варто використати можливості програми щодо підбору виду кривої.

Приклад 4.4. Побудувати регресійну модель, що характеризує залежність обсягу продажу деякої продукції в день *Y* (тис. грн.) від кількості днів рекламної компанії *X* (дні). Дані наведено у табл. 4.14.

Таблиця 4.14

X	11	12	22	23	25	30	34	56	78	90
Y	540	530	505	490	483	465	470	485	484	470

Розв'язок. Для спрощення дослідження про вид та рівняння емпіричної лінії регресії, необхідно:

1) Ввести дані табл. 4.14 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

2) Вибрати в меню послідовно ***Анализ − Регрессия − Подгонка кривых***. У діалоговому вікні ***Подгонка кривых*** перенести змінну *X* у поле ***Независимая переменная***, а змінну *Y* у поле ***Зависимые*** і активувати усі моделі у полі ***Модели***, відзначивши їх галочками (рис. 4.17);

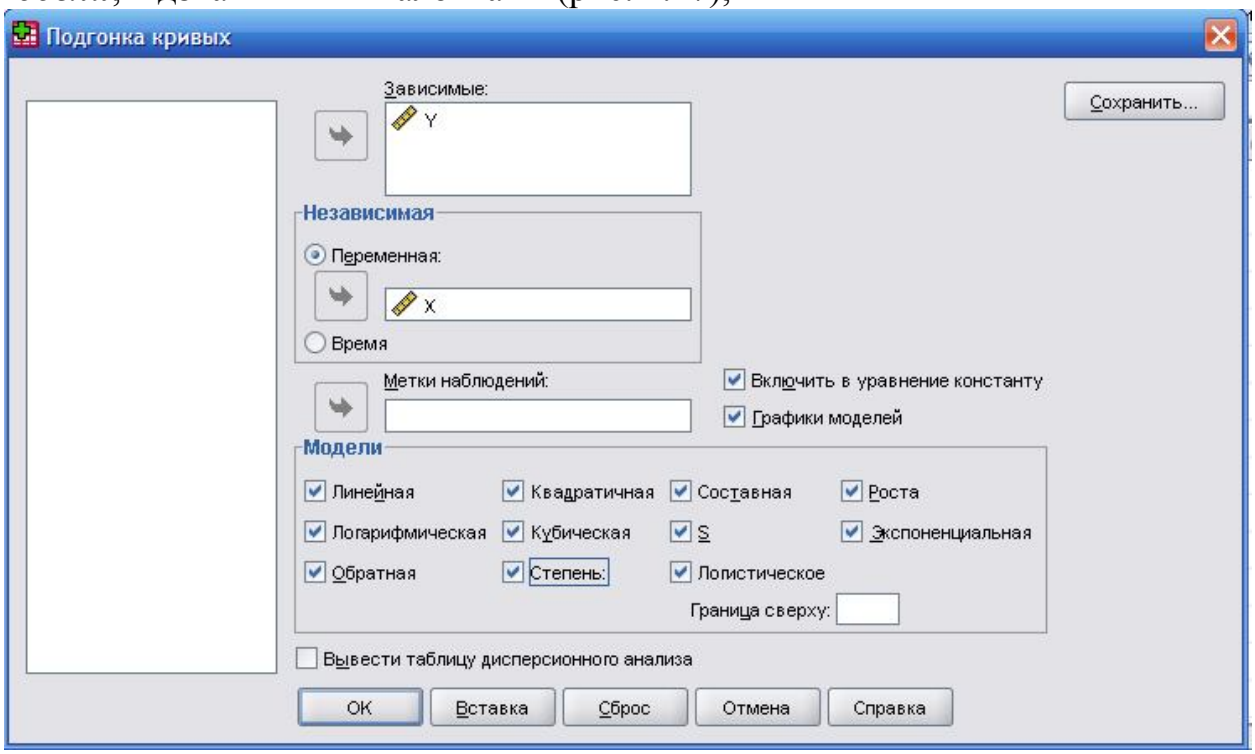

Рисунок 4.17. Діалогове вікно вибору гіпотетичної моделі лінії регресії

3) Проаналізувати дані таблиць вікна виводу результатів (рис. 4.18).

Сводка модели и оценки параметров

Зависимая переменная:Y

Уравнение	Сводка для модели					Оценки параметра			
	R-квадрат	F	ст.св.1	ст.св.2	Знч.	Константа	b1	b2	b3
Линейный	,335	4,039	1	8	,079	512,657	-,537		
Логарифмическая	,577	10,916	1	8	,011	584,680	-27,090		
Обратная	,795	30,985	1	8	,001	458,412	828,724		
Квадратичный	,665	6,939	2	7	,022	557,874	-3,225	,027	
Кубический	,941	31,955	3	6	,000	621,955	-9,391	,183	-,001
Составная	,334	4,014	1	8	,080	512,038	,999		
Степенная	,572	10,711	1	8	,011	590,673	-,054		
S	,786	29,307	1	8	,001	6,131	1,642		
Роста	,334	4,014	1	8	,080	6,238	-,001		
Экспоненциальная	,334	4,014	1	8	,080	512,038	-,001		
Логистическая	,334	4,014	1	8	,080	,002	1,001		

Независимой переменной является X.

Рисунок 4.18. Результати підбору виду регресії

Дані першого стовпчика таблиці ***Сводка модели и оценки параметров*** (рис. 4.18) ***R-квадрат*** є коефіцієнтами детермінації, які показують скільки відсотків вибіркових об'єктів охоплює кожний вид рівняння. Необхідно вибрати максимальне значення: в нашому випадку − 0,941.

Отже, емпіричною лінією регресії є кубічна парабола, яка охоплює 94,1% досліджуваних даних. Це підтверджує також найменше значення рівня значущості серед усіх інших видів рівняння p = 0,000 (у таблиці стовпчик ***Знч***).

Запишемо рівняння: $\overline{y}_x = 621{,}96 - 9{,}391x + 0{,}183x^2 - 0{,}001x^3$, графік якого, поряд з іншими, зображений на рис. 4.19.

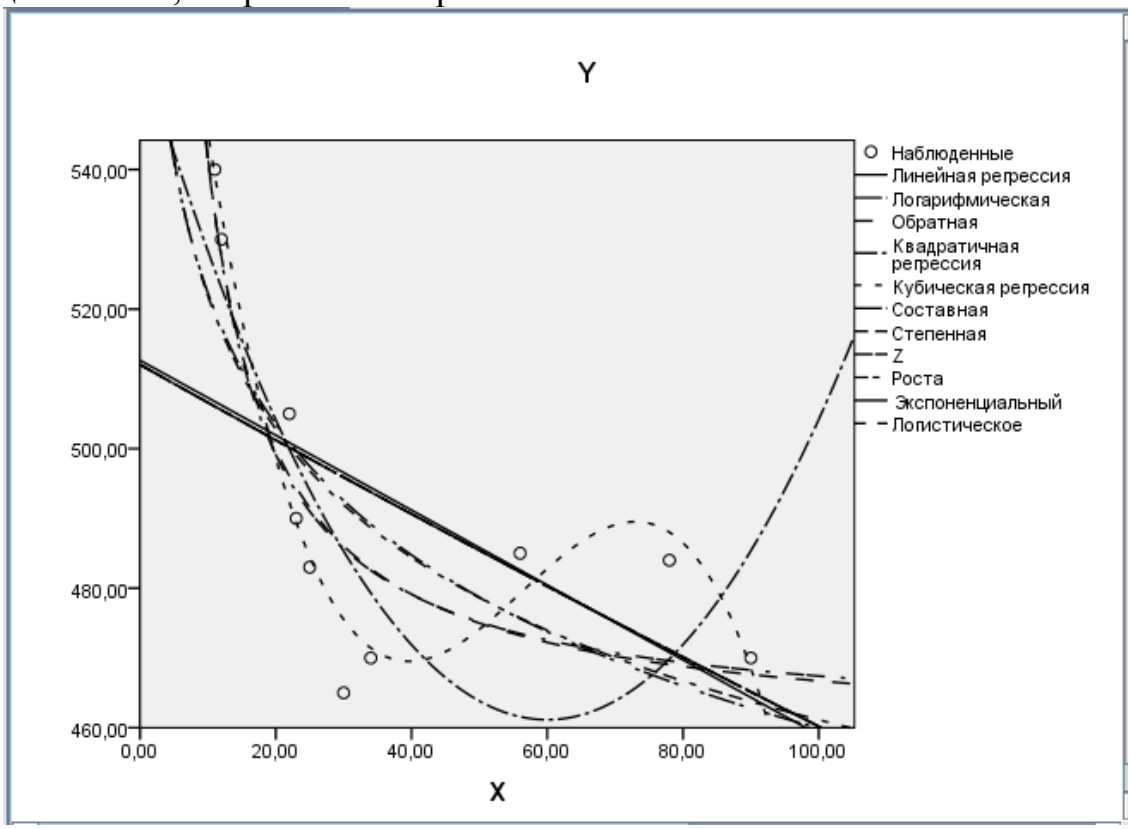

Рисунок 4.19. Графіки підгонки кривих

Висновок: Залежність обсягу продажу деякої продукції в день Y (тис. грн.) від кількості днів рекламної компанії X (дні) описується рівнянням $\bar{y}_x = 621{,}96 - 9{,}391x + 0{,}183x^2 - 0{,}001x^3$, яке характеризує 94,1% варіації результативної ознаки Y з ймовірністю випадковості отриманого результату $p = 0{,}000$.

4.6.3. Множинна лінійна регресія

Множинний регресійний аналіз передбачає вивчення залежності між кількома незалежними ознаками. Ознаки можуть належати до інтервальної або порядкової шкал. Якщо ж ознака відноситься до номінальної шкали і може бути дихотомічною, то її можна розписати на кілька дихотомічних змінних. Наприклад, ознаку *Освіта (середня, середня професійна, неповна вища, вища)* можна представити як: *Освіта1 (1 − середня, 0 − не середня), Освіта2 (1 − середня професійна, 0 − не середня професійна); Освіта3 (1 − неповна вища, 0 − не неповна вища)* і т. д.

Множинний аналіз лінійної регресії в SPSS можна провести з допомогою кількох методів. Автоматично активований метод *Принудительное включение*, який не варто використовувати для множинного аналізу. Даний метод передбачає одночасну обробку усіх незалежних ознак, тому об'єктивними можуть бути лише результати аналізу рівняння регресії з однією ознакою (змінною). Для множинного аналізу варто вибрати один із покрокових методів. Використовуючи прямий метод, незалежні змінні, які мають найбільші значення коефіцієнтів частинної кореляції з залежною змінною, поетапно вносяться у рівняння регресії. Обернений метод передбачає вилучення незалежних змінних з найменшими значеннями частинних коефіцієнтів кореляції із гіпотетичного рівняння лінії регресії, яке містить усі змінні. Процес продовжується до того часу, поки відповідний регресійний коефіцієнт не виявиться незначущим (у даному випадку рівень значущості дорівнює 0,1).

Приклад 4.5. У 10 компаніях вивчається взаємозв'язок між середньорічними цінами на рекламу X_1 (млн. грн.), рівнем затрат на проведення реклами X_2 (% до вартості реалізованої продукції) та вартістю реалізованої рекламної продукції Y (млн. грн.). Дані наведено у таблиці 4.15

Таблиця 4.15

№ компанії	X_1	X_2	Y
1	3	4	20
2	3	3	25
3	5	3	20
4	6	5	30
5	7	10	32
6	6	12	25
7	8	12	29
8	9	11	37
9	9	15	36
10	10	15	40

Вважаючи, що між показниками існує лінійна залежність, визначити параметри рівняння регресії та оцінити адекватність обраної моделі.

Розв'язок. Для знаходження коефіцієнтів рівняння емпіричної лінії регресії, необхідно:

1) Ввести дані табл. 4.15 у стовпчики вкладки ***Данные*** редактора ***Набор данных***.

2) Вибрати в меню послідовно ***Анализ − Регрессия − Линейная…*** У діалоговому вікні ***Линейная регрессия*** перенести змінні X_1, X_2 у поле ***Независимые переменные***, а змінну Y у поле ***Зависимые***. Вибрати один із обернених покрокових методів: ***Удалить*** (рис. 4.20).

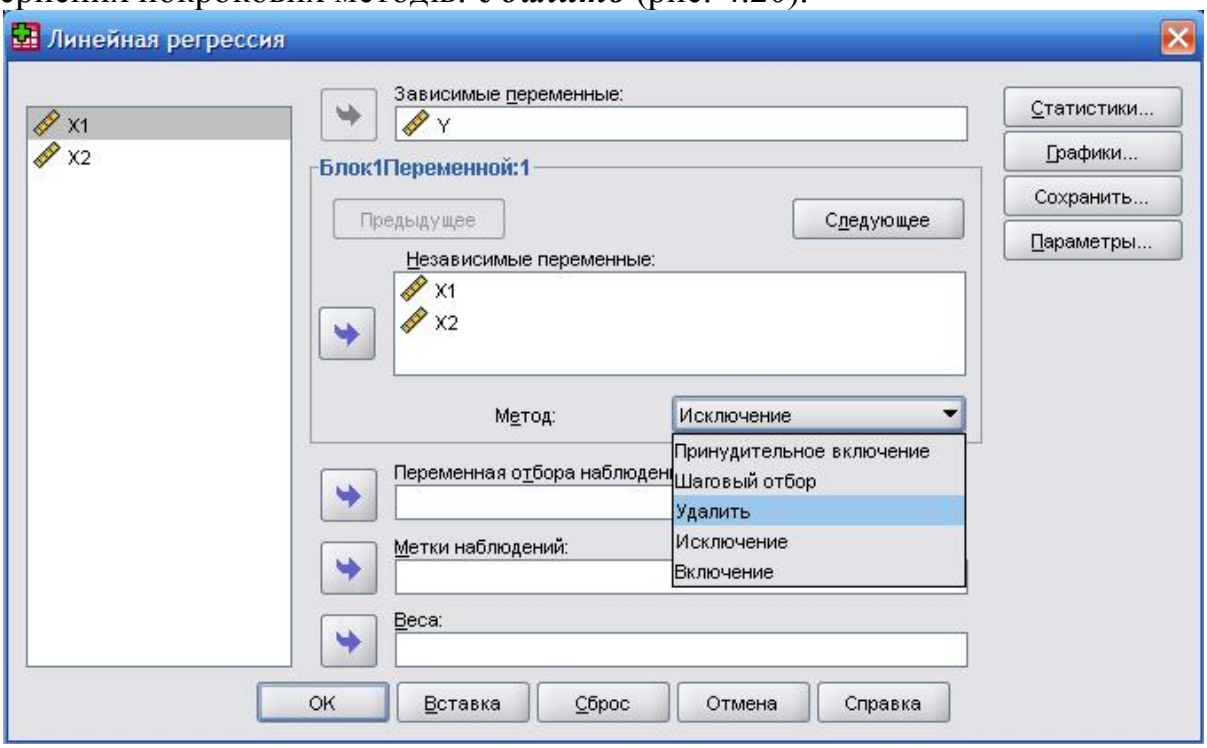

Рисунок 4.20. Діалогове вікно множинної лінійної регресії

3) Проаналізувати інформацію вікна виведення результатів (рис. 4.21) та зробити відповідні висновки.

Дані стовпчика ***R-квадрат*** таблиці ***Сводка для модели*** є коефіцієнтами детермінації, які вказують на степінь відповідності між регресійними моделями і вхідними даними. Коефіцієнт детермінації для моделі 1 дорівнює 0,798. Отже, 80% досліджуваних об'єктів описує перша модель емпіричної лінії регресії.

У стовпчику ***Нестандартизованные коэффициенты: B*** таблиці ***Коэффициенты*** подано значення коефіцієнтів рівняння регресії, а саме:

$$Y_{X_1X_2} = 12{,}51 + 2{,}67X_1 - 0{,}083X_2 .$$

Даними стовпчика ***Стандартизованные коэффициенты Бета*** таблиці ***Коэффициенты*** є регресійні коефіцієнти, які вказують важливість незалежних ознак, використаних в рівнянні лінії регресії. Значення 0,943 показує, наскільки важливі показники середньорічних цін на рекламу (X_1) для визначення вартості реалізованої рекламної продукції (Y).

У таблиці ***Исключенные переменные*** дані стовпчика ***Частная корреляция*** показують тісний зв'язок між незалежними і залежною змінними ($r_{X_1Y} = 0,89$, $r_{X_2Y} = 0,78$).

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,893[a]	,798	,741	3,546
2	,000[b]	,000	,000	6,963

a. Предикторы: (конст) X2, X1

b. Предиктор (константа)

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	12,507	3,595		3,479	,010
	X1	2,672	1,027	,943	2,601	,035
	X2	-,083	,525	-,057	-,157	,879
2	(Константа)	29,400	2,202		13,351	,000

a. Зависимая переменная: Y

Исключенные переменные[b]

Модель		Бета включения	t	Знч.	Частная корреляция	Статистики коллинеарности
						Толерантность
2	X1	,893[a]	5,614	,001	,893	1,000
	X2	,777[a]	3,488	,008	,777	1,000

a. Предиктор (константа)

b. Зависимая переменная: Y

Рисунок 4.21. Результати розрахунку коефіцієнтів рівняння множинної регресії

Висновок: на вартість реалізованої рекламної продукції (Y) значно впливають середньорічні ціни на рекламу (X_1) та рівень затрат на проведення реклами (X_2). Залежність від згаданих факторів можна описати наступним рівнянням регресії:

$$Y_{X_1X_2} = 12,51 + 2,67X_1 - 0,083X_2 .$$

Завдання для самостійного виконання

4.1. Відомо дані про обсяг виробництва сільськогосподарської продукції (грн.) на 1 особу АР Крим (табл. 4.11). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.11

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг виробництва с/г продукції	1000	995	949	926	1049	1112	1596

4.2. Відомо дані про чисельність наукових та науково-технічних працівників, що припадають на 1000 осіб (табл. 4.12). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.12

Рік	2001	2002	2003	2004	2005	2006	2007
Чисельність наукових та науково-технічних працівників	1,1	1,1	1,0	1,0	1,0	1,0	0,9

4.3. Відомо дані про інвестиції в основний капітал в розрахунку на 1 особу (табл. 4.13). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.13

Рік	2001	2002	2003	2004	2005	2006	2007
Інвестиції в основний капітал	376,8	600,2	735,7	955,2	1376,2	1704,1	2375,6

4.4. Відомо дані про обсяг інноваційної продукції в розрахунку на 1 особу (табл. 4.14). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.14

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг інноваційної продукції	38,4	139,0	264,5	172,3	313,1	469,9	282,2

4.5. Відомо дані про обсяг експорту товарів в розрахунку на 1 особу (табл. 4.15). Побудувати регресійну модель за даними таблиці, оцінити її статистичну значущість та адекватність.

Таблиця 4.15

Рік	2001	2002	2003	2004	2005	2006	2007
Обсяг експорту товарів	84,6	107,3	109,2	158,1	137,1	178,1	201,7

4.6 – 4.15. Знайти рівняння ліній регресії, які описують залежність Y від X за даними кореляційних табл. 4.16–4.17, оцінити їх статистичну значущість та адекватність.

Таблиця 4.16

Y	X					
	10	20	30	40	50	60
5	*a*	*b*				
10		*c*	*d*			
15			*e*	*f*	*g*	
20			*h*	*k*	*m*	
25				*n*	*p*	*q*

Таблиця 4.17

	4.6	4.7	4.8	4.9	4.10	4.11	4.12	4.13	4.14	4.15
a	2	2	4	1	4	2	2	2	3	4
b	3	6	2	5	2	4	4	4	3	2
c	7	4	6	5	6	3	6	6	5	5
d	3	4	4	3	2	7	2	3	4	3
e	2	7	6	9	5	5	3	6	20	5
f	50	35	45	40	40	30	50	45	22	45
g	2	8	2	2	5	10	2	4	8	5
h	1	2	2	4	2	7	1	2	5	2
k	10	10	8	11	8	10	10	8	10	8
m	6	8	6	6	7	8	6	6	6	7
n	4	5	4	4	4	5	4	4	4	4
p	7	6	7	7	7	6	7	7	7	7
q	3	3	4	3	8	3	3	3	3	3

Питання для самоконтролю

1. Що називається регресійним аналізом?
2. Що називається регресійною моделлю?
3. Що називається факторними ознаками? Результативною ознакою?
4. Які види регресійних моделей Ви знаєте?
5. Як повинні бути задані вхідні дані для регресійного аналізу?
6. Як сформулювати гіпотезу про вид регресійної моделі?
7. Що таке теоретичні та емпіричні значення результативної ознаки?
8. Що називається емпіричною лінією регресії? Теоретичною лінією регресії?
9. Як побудувати емпіричну лінію регресії? Теоретичну лінію регресії?
10. Як знайти параметри регресійної моделі?
11. Як будується розрахункова таблиця у регресійному аналізі?
12. Як перевірити правильність побудованої регресійної моделі?
13. Як перевірити адекватність побудованої регресійної моделі вхідним даним?
14. Що називається коефіцієнтом детермінації і як він використовується у статистичному моделюванні?
15. Як перевірити статистичну значущість побудованої регресійної моделі?
16. Що називається багатофакторною лінійною регресією?
17. Які етапи побудови багатофакторної лінійної регресійної моделі?
18. Як обґрунтовується вибір факторних ознак для побудови моделі?
19. Що називається кореляційними плеядами?
20. Як оцінюється вплив факторних ознак на результативну?
21. Як здійснюється прогноз за багатофакторною лінійною регресійною моделлю?
22. Як визначити вид нелінійної регресійної моделі?