

Самостійна робота 9.2. Приклад розв'язання

На основі даних умови попередньої самостійної роботи 9.1 виконати розрахунки за допомогою вбудованих в MS Excel функцій:

1) Побудувати базову таблицю дисперсійного аналізу (ANOVA-таблицю) та визначити дисперсії (функції: *VAR.P* (ДИСПР), *DEVSQ* (КВАДРОТКЛ), *SUMXMY2* (СУММКВРАЗН));

2) Визначити коефіцієнти детермінації (*RSQ* (КВПІРСОН)) та кореляції (*CORREL* (КОРРЕЛ));

3) Побудувати економетричну модель за допомогою функції *LINEST* (ЛИНЕЙН).

Розв'язання

1. Побудова базової таблиці дисперсійного аналізу (ANOVA-таблиці) та визначення дисперсії.

Для виконання завдання самостійної роботи 9.2 в тому ж файлі, в якому виконували самостійну роботу 9.1 відкрити новий аркуш, який треба назвати за номером роботи – «9.2». На цей новий робочий аркуш 9.2 переносимо частину даних з самостійної роботи 9.1, а саме: знайдені значення оцінок параметрів a_0 та a_1 ; умову та розраховане теоретичне значення (рис. 1).

	A	B	C	D	E	F
4						
5			ОЗ	Прибуток		
6		№ під-ва	X_i	Y_i	$Y_{теор}$	
7		1	2	3	7	
8		1	2,50	1,20	1,13	
9		2	2,80	1,50	1,49	
10		3	5,00	4,50	4,16	
11		4	3,60	2,20	2,46	
12		5	3,00	1,90	1,74	
13		6	4,20	3,10	3,19	
14		7	4,50	3,40	3,55	
15		8	5,60	4,80	4,89	
16		9	3,90	2,80	2,83	
17		10	6,00	5,40	5,37	
18						

Рис. 1. Дані, перенесені з самостійної роботи 9.1

Для визначення

Згадаємо правило додавання дисперсій:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9.18)$$

де $\sum_{i=1}^n (y_i - \bar{y})^2$ – загальна сума квадратів, яка позначається через *TSS* (*total sum squares*); вона відображає дисперсію величини y_i (емпіричне або фактичне значення) відносно її середнього значення;

$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ – сума квадратів, що пояснює регресію та позначається через *ESS* (*explained sum squares*); відображає дисперсію оціненої (теоретичної) величини $\hat{y}_i$ відносно середнього значення $\bar{y}$;

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$ – сума квадратів помилок, яка позначається через *RSS* (*residual sum squares*); відображає залишкову або нез’ясовану дисперсію величини y_i щодо лінії регресії $\hat{y}_i$ або просто залишкова сума квадратів.

Вираз (9.18) запишемо у скороченому вигляді:

$$TSS = ESS + RSS \quad (9.19)$$

Формула (9.19) показує, що загальна варіація спостережуваних величин Y щодо їх середнього значення (*TSS*) може бути розбита на дві частини, одна з яких відповідає лінії регресії (пояснена сума квадратів – *ESS*), а інша – випадковим відхиленням (сума квадратів залишків – *RSS*), оскільки не всі спостережувані Y лежать на лінії регресії. На рис. 2 це розбиття пояснене геометрично.

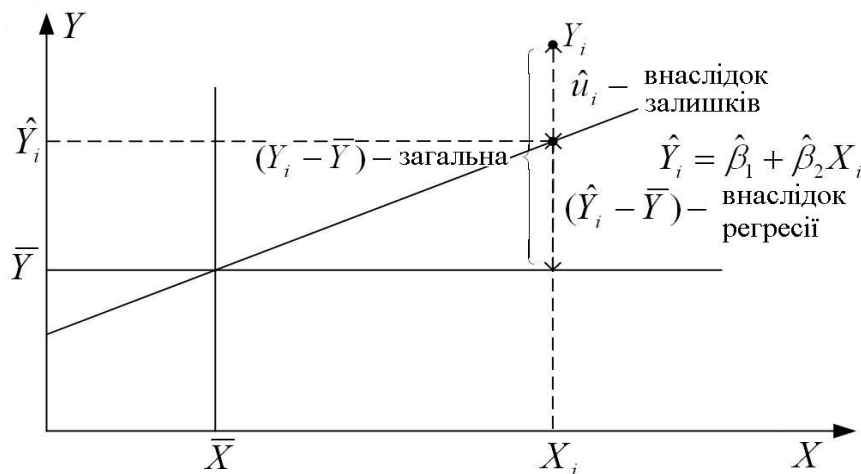


Рис. 2. Розбиття варіації Y_i на дві компоненти

Таким чином, ми розклали загальну дисперсію на дві частини: дисперсію, яка пояснює регресію, та дисперсію помилок (або дисперсію випадкової величини).

Проведемо аналіз дисперсії з погляду регресії (табл. 9.1) (*ANOVA*, *analysis variance*):

ANOVA-таблиця для регресійної моделі

Джерело варіації	Ступені свободи, df	Сума квадратів відхилень, SS	Середні суми квадратів відхилень, MS
Регресії	$k_1=m-1$	$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MSE = \frac{ESS}{k_1}$
Залишків	$k_2=n-m$	$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSR = \frac{RSS}{k_2}$
Загальної змінної	$n-1$	$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$	$MST = \frac{TSS}{n-1}$
SS – сума квадратів (sum squares) MS – середня сума квадратів (mean sum squares) n – кількість спостережень, m – кількість змінних в моделі			

На робочому аркуші за допомогою меж виокремити діапазон для розрахунків як відображено на рис. 3.

Джерело варіації	Ступені свободи, df	Сума квадратів відхилень, SS	Середні суми квадратів відхилень, MS
Регресії			
Залишків			
Загальної змінної			

Рис. 3

Заповнюємо дані ANOVA-таблиці для нашої регресійної моделі $\hat{y} = -1,8994 + 1,2115 \cdot x$.

Визначаємо ступені свободи:

n – обсяг вибіркової сукупності або кількість спостережень, в нашому випадку $n=10$;

m – кількість змінних моделі – $m=2$;

$k_1=m-1=2-1=1$;

$k_2=n-m=10-2=8$

Суми квадратів відхилень, SS , можна автоматично отримати за допомогою вбудованих в MS Excel функцій.

Для визначення ESS використовуємо функцію *DEVSQ* або *КВАДРОТКЛ*. В аргументі функції «Число 1» треба зробити посилання на діапазон даних, де знаходяться теоретичні значення результативної ознаки Y ($Y_{\text{теор}}$ з рис. 1), як це відображено на рис. 4. Тому, що ESS відображає дисперсію оціненої (теоретичної) величини $\hat{y}_i$ відносно середнього значення $\bar{y}$:

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

$ESS = KBAДPOTKЛ(\hat{Y}) = 18,48$ – англomовна назва функції *DEVSQ*

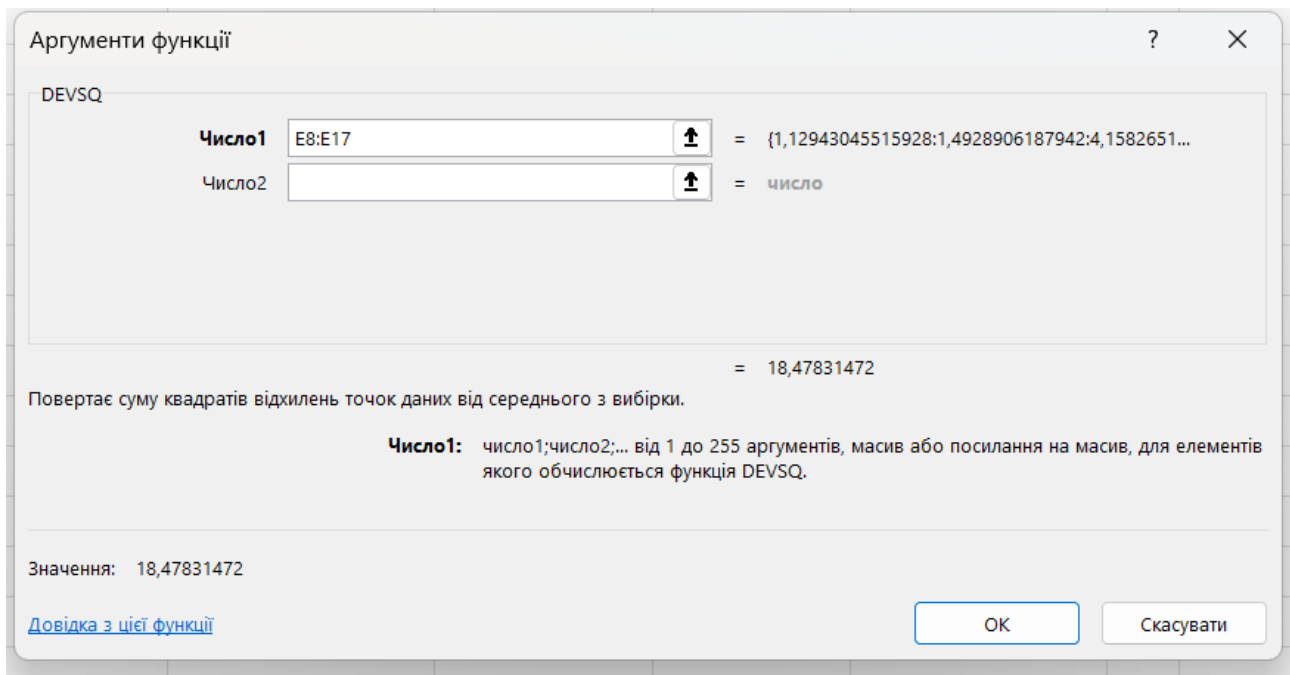


Рис. 4. Вікно аргументів функції *DEVSQ* або *KBAДPOTKЛ*

Для визначення *RSS* використаємо функцію *SUMXMY2* або *СУММКВРАЗН*. В аргументах функції «Масив_х» треба зробити посилання на діапазон даних, де знаходяться фактичні (емпіричні) значення змінної Y_i , а «Масив_у» – теоретичні значення результативної ознаки $Y_{теор}$, як це відображено на рис. 5. Тому, що *RSS* відображає залишкову або нез'ясовану дисперсію величини y_i щодо лінії регресії $\hat{y}_i$ або просто залишкову сума квадратів:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

$RSS = СУММКВРАЗН(Y, \hat{Y}) = 0,26$ – англomовна назва функції *SUMXMY2*

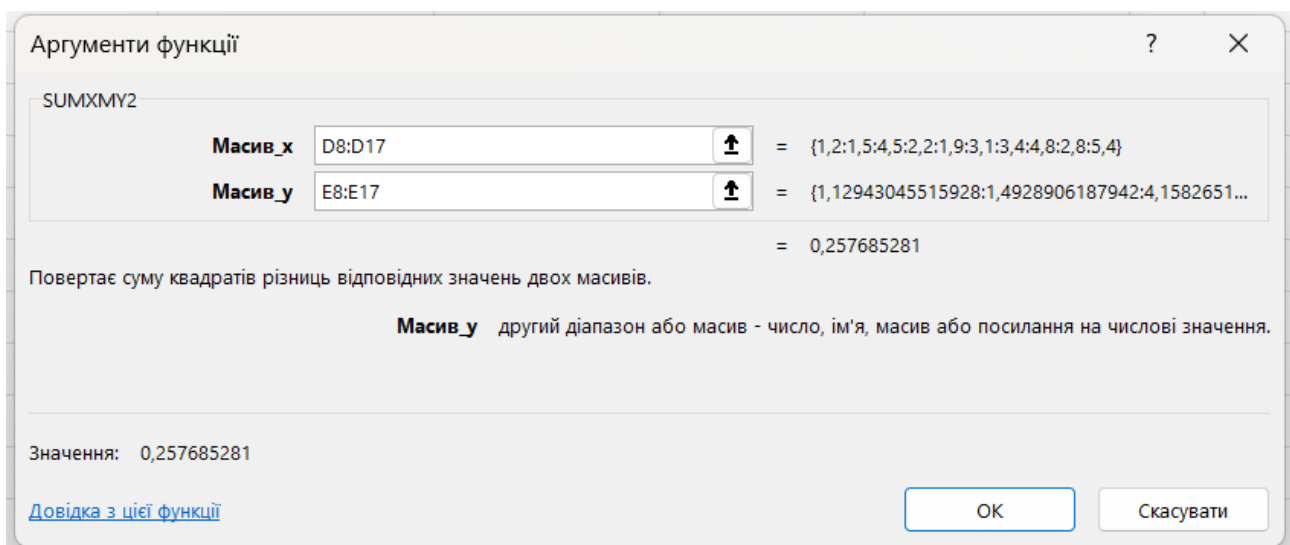


Рис. 5. Вікно аргументів функції *SUMXMY2* або *СУММКВРАЗН*

Для визначення TSS використаємо функцію $DEVSQ$ або $КВАДРОТКЛ$. В аргументі функції «Число 1» треба зробити посилання на діапазон даних, де знаходяться фактичні (емпіричні) значення змінної (Y_i з рис. 1), як це відображено на рис. 6. Тому, що TSS відображає дисперсію емпіричного або фактичного значення величини y_i відносно її середнього значення:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$TSS = КВАДРОТКЛ(Y) = 18,74 \text{ – англomовна назва функції } DEVSQ$$

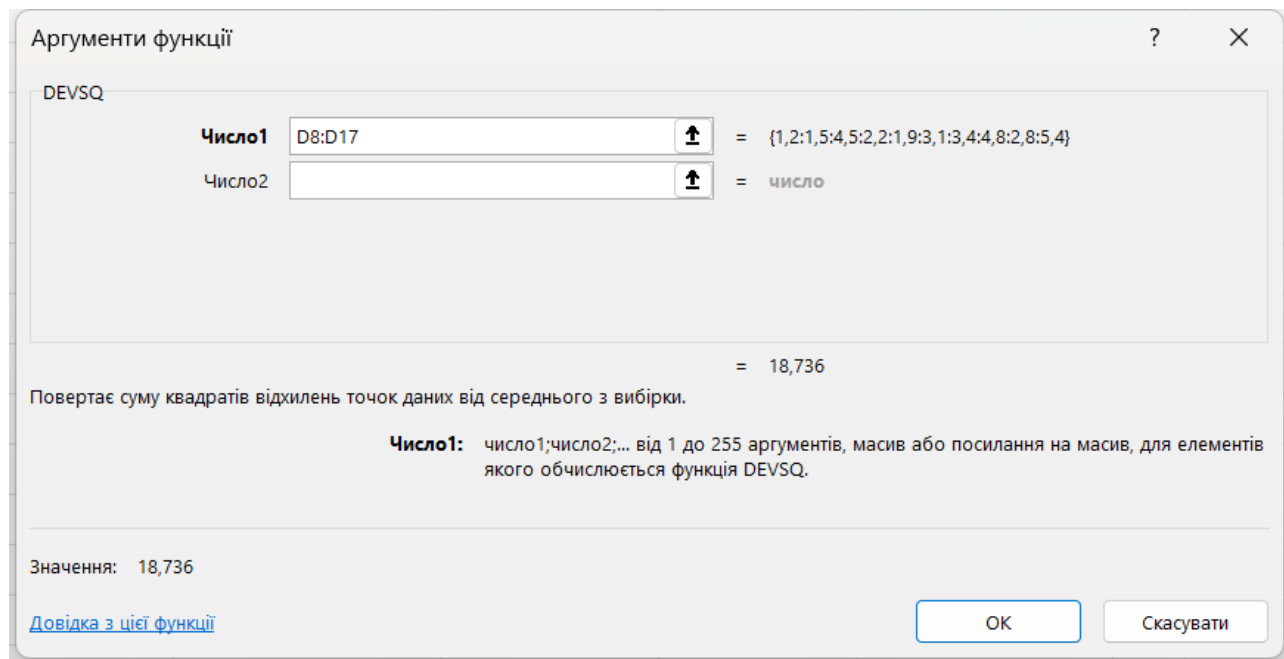


Рис. 6. Вікно аргументів функції $DEVSQ$ або $КВАДРОТКЛ$

Перевіримо розрахунки, відповідно до теореми додавання дисперсій:
 $18,74 = 18,48 + 0,26$

Для розрахунку середніх сум квадратів відхилень, MS скористаємось формулами, наведеними в табл. 9.1.

Заповнюємо $ANOVA$ таблицю для нашого прикладу:

Таблиця 9.2

$ANOVA$ таблиця для регресійної моделі $\hat{y} = -1,8994 + 1,2115 \cdot x$

Джерело варіації	Ступені свободи, df	Сума квадратів відхилень, SS	Середні суми квадратів відхилень, MS
Регресії	1	18,48	$MSE = 18,48/1 = 18,48$
Залишків	8	0,26	$MSR = 0,26/8 = 0,032$
Загальної змінної	9	18,74	$MST = 18,74/9 = 2,0818$

Використовуючи $ANOVA$ -таблицю визначимо:

- дисперсію регресії
 $MSE = 18,48$
- незміщену оцінку дисперсії залишків

$$MSR=0,032$$

- стандартизовану похибку залишків
 $s = \sqrt{MSR}=0,1794$ або 17,94%
- дисперсію залежної змінної (загальну дисперсію)
 $MST=2,0818$

Чим менша стандартна помилка залишків, тим краще підібрана функція.

2. Визначення коефіцієнтів детермінації та кореляції .

За допомогою даних ANOVA-таблиці визначаємо **коефіцієнт детермінації**

$$R^2 = \frac{ESS}{TSS} = \frac{18,48}{18,74} = 0,9862$$

$$\text{або } R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{0,2577}{18,74} = 0,9862$$

Перевірити одержаний результат можна за допомогою статистичної функції (*RSQ (КВПИРСОН)*). В аргументах функції треба вказати діапазони на робочому аркуші емпіричних значень відповідних змінних X_i та Y_i (рис. 7).

$$R^2 = \text{КВПИРСОН}(Y; X) = \text{RSQ}(Y; X) = 0,9862$$

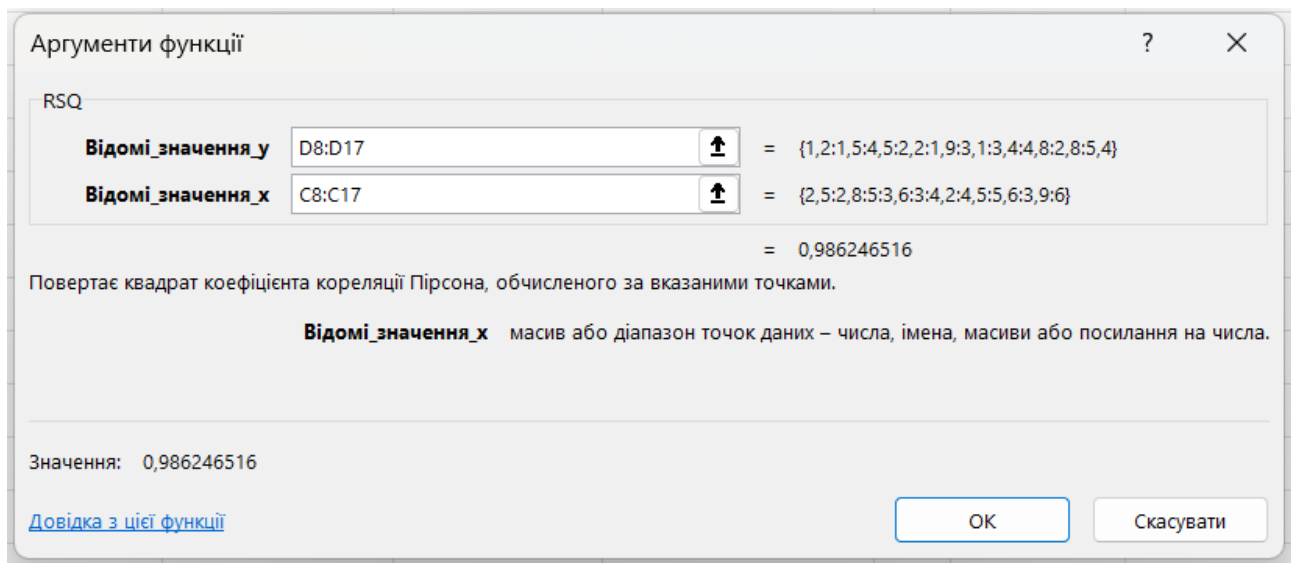


Рис. 7. Вікно аргументів функції *RSQ* або *КВПИРСОН*

Оскільки коефіцієнт детермінації дорівнює 0,9862, то це означає, що теоретична пряма пояснює 98,62% або можна сказати, що 98,62% варіації прибутку підприємства відбувається під впливом вартості основних засобів. А решта 1,38% - припадає на інші фактори та випадкові величини.

Визначимо **коефіцієнт кореляції Пірсона** за допомогою статистичної функції (*CORREL (КОРРЕЛ)*). В аргументах функції «Масив1» та «Масив2» треба також вказати діапазони на робочому аркуші емпіричних значень змінних X_i та Y_i (рис. 8). Оскільки ми шукаємо кореляційний зв'язок порядок змінних не впливає на значення коефіцієнта кореляції.

$$R^2 = \text{КОРРЕЛ}(X; Y) = \text{CORREL}(X; Y) = 0,9931$$

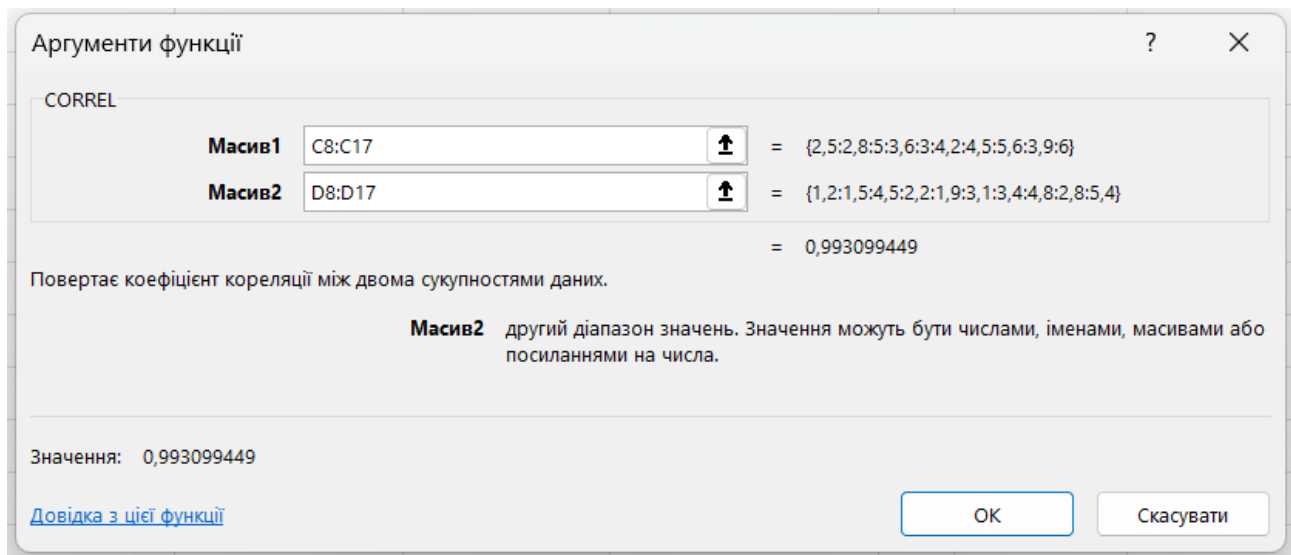


Рис. 8. Вікно аргументів функції *CORREL* або *KORPEЛ*

Коефіцієнт кореляції 0,9931 свідчить, що між вартістю основних засобів підприємства та його прибутком є дуже щільний прямий лінійний зв'язок. Оскільки значення коефіцієнта кореляції близьке до 1, то це означає, що залежність між змінними x та y практично лінійна.

3. Побудова економетричної моделі за допомогою функції *LINEST* (*ЛИНЕЙН*).

Для парної регресії отримаємо дані у вигляді таблиці (рис. 1).

$\hat{a}_1$	$\hat{a}_0$
$S_{\hat{a}_1}$	$S_{\hat{a}_0}$
R^2	σ_u
F	$k = n - m$
$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2$	$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$

Рис. 9

Перший рядок результатів розрахунку містить оцінки параметрів моделі a_0 та a_1 .

Другий рядок містить стандартні похибки оцінок параметрів моделі.

Третій рядок містить коефіцієнт детермінації та стандартну помилку залишків.

Четвертий рядок містить F-критерій Фішера та число ступенів свободи.

П'ятий рядок таблиці результатів містить суми квадратів.

Виділяємо на робочому аркуші область розміром 5 рядків та 2 стовпці та застосовуємо до неї функцію.

В аргументах функції *LINEST* або *ЛИНЕЙН* «Відомі значення y » та «Відомі значення x » треба вказати діапазони на робочому аркуші емпіричних значень відповідних змінних X_i та Y_i (рис. 10).

Аргументи функції

LINEST

Відомі_значення_y: D8:D17 = {1,2;1,5;4,5;2,2;1,9;3,1;3,4;4,8;2,8;5,4}

Відомі_значення_x: C8:C17 = {2,5;2,8;5;3,6;3;4,2;4,5;5,6;3,9;6}

Конст: 1 = TRUE

Статистика: 1 = TRUE

Повертає параметри лінійного наближення, обчислені за методом найменших квадратів.

Статистика логічне значення: якщо TRUE, повертається додаткова статистика регресії; якщо FALSE або не вказано, повертаються лише коефіцієнти m і константа b.

Значення: 1,211533879

[Довідка з цієї функції](#) OK Скасувати

Рис. 10. Вікно аргументів функції *LINEST* або *ЛИНЕЙН*

ЗАУВАЖЕННЯ: Після вводу всіх аргументів функції, як це наведено на рис. 10 замість того аби натиснути кнопку **ОК**, треба натиснути послідовно клавіші **Ctrl + Shift + Enter**. Це зауваження стосується старших версій MS Excel/ В сучасних версіях натискаємо «ОК».

У виокремленому діапазоні отримаємо наступні дані:

1,211534	-1,899404242
0,050583	0,215503573
0,986247	0,179473285
573,6708	8
18,47831	0,257685281

Рис. 11

Синтаксис функції *ЛИНЕЙН* (*LINEST*)

ЛИНЕЙН (відомі значення *y*; відомі значення *x*; константа; статистика) – повертає параметри лінійного наближення за методом найменших квадратів, тобто визначає оцінки параметрів лінійної регресії:

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \dots + \hat{a}_k x_k.$$

«Відомі значення y » – множина значень y . Якщо масив y має один стовпець, то кожен стовпець масиву «відомі значення x » інтерпретуються як окрема змінна. Якщо масив «відомі значення y » має один рядок, то кожен рядок «відомих значень x » інтерпретується як окрема змінна.

«Відомі значення x » – множина значень x , що враховує або одну (парна регресія), або кілька змінних (множинна регресія). Якщо «відомі значення x » пропустили, то вважається, що це масив $\{1;2;3;\dots\}$ такого самого розміру, як n «відомих значень y ».

«Константа» – логічне значення. Якщо «константа» має значення «ложь», то $a_0 = 0$. Якщо «константа» має значення «истина», то a_0 обчислюється традиційно (модель з вільним членом).

«Статистика» – логічне значення. Якщо «статистика» має значення «истина», то функція додатково обчислює регресійну статистику:

$\hat{a}_k$	$\hat{a}_{k-1}$	...	$\hat{a}_1$	$\hat{a}_0$
$S_{\hat{a}_k}$	$S_{\hat{a}_{k-1}}$	...	$S_{\hat{a}_1}$	$S_{\hat{a}_0}$
R^2	σ_u			
F	$k = n - m$			
$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$			

де $\hat{a}_j$ – оцінка параметра a_j , $j = \overline{1, k}$;

$\hat{a}_0$ – оцінка вільного члена регресії;

$S_{\hat{a}_j}$ – стандартна похибка оцінки параметра;

R^2 – коефіцієнт детермінації;

σ_u – стандартна похибка залишків;

F - критерій Фішера;

$k = n - m$ – ступінь свободи, де n – кількість спостережень,
 m – кількість змінних у моделі.

$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ – сума квадратів відхилення, що

пояснюється регресією;

$SSE = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$ – сума квадратів, що пояснюється

похибкою u .

Якщо «статистика» має значення «ложь», то функція
ЛИНЕЙН обчислює лише коефіцієнти $\hat{a}_j$ та константу $\hat{a}_0$.