ПРАКТИЧНА РОБОТА № 5, 6
ПОБУДОВА КОРЕЛЯЦІЙНОГО ПОЛЯ І РІВНЯННЯ РЕГРЕСІЇ

1. Мета роботи та завдання

Освоїти методи статистичного дослідження кореляційного взаємозв'язку.
Завданнями роботи є:
- Надбання навичок побудови кореляційного поля і рівняння регресії;
- Закріплення навичок знаходження коефіцієнтів рівняння регресії і умовного середнього.

2. Теоретичні відомості

Аналіз статистичних даних дозволяє виявити взаємозв'язок досліджуваних явищ. При цьому спостерігаються два види зв'язку: причинно-наслідкові зв'язки (зміни в одному явищі є причиною змін в іншому) і кореляція (зміни в обох явищах відбуваються одночасно і викликані загальною причиною). Кореляційна залежність виглядає як розкид точок щодо лінії на діаграмі розсіювання. Модель взаємозв'язку відображає кількісні відносини і будується методами кореляційного і регресійного аналізу. Кореляційний аналіз дозволяє досліджувати тісноту зв'язку, тобто ступінь розкиду точок від лінії. Регресійний аналіз дозволяє побудувати рівняння зв'язку. Для правильної інтерпретації моделі необхідний етап якісного аналізу досліджуваного явища, його природи і внутрішніх механізмів.

Тіснота лінійного зв'язку оцінюється за допомогою коефіцієнта лінійної кореляції:

$$r_{xy} = \frac{\overline{yx} - \bar{y} * \bar{x}}{\sigma_x \sigma_y}$$

Коефіцієнт кореляції приймає значення від -1 до +1, включно; його знак вказує на зворотній або прямий зв'язок показників. Величина коефіцієнта характеризує тісноту лінійного зв'язку (див. Табл. 2.1).

Таблиця 2.1

Оцінка тісноти лінійного зв'язку

Величина коефіцієнту кореляції	Характер зв'язку
$\|r\| < 0,3$	Майже відсутній
$0,3 \le \|r\| < 0,5$	Слабкий
$0,5 \le \|r\| < 0,7$	Помірний
$0,7 \le \|r\| < 1,0$	Сильний
$\|r\| = 1,0$	Функціональний

Низьке значення коефіцієнта говорить про відсутність лінійного зв'язку. Фактичною причиною можуть бути повна відсутність зв'язку, високий рівень випадкових відхилень, або наявність істотно нелінійного зв'язку. Тіснота нелінійного зв'язку може оцінюватися за допомогою коефіцієнтів рангової кореляції.

Модель зв'язку зазвичай будується в формі рівняння регресії. Парна регресія (зв'язок двох показників) може описуватися рівняннями:

Прямої: $\overline{y_x} = a_1 x + a_0$;

Параболи: $\overline{y_x} = a_2 x^2 + a_1 x + a_0$

Кубічного рівняння: $\overline{y_x} = a_3 x^3 + a_2 x^2 + a_1 x + a_0$

Невідомі коефіцієнти $a_0, a_1 \ldots . a_k$ можуть бути знайдені методом найменших квадратів(МНК), шляхом мінімізації суми квадратів:

$$\sum (\overline{y_x} - y_x)^2 \to \min$$

Системи рівнянь для обчислення коефіцієнтів регресії для поліномів різних ступенів виглядають наступним чином:

$$\begin{cases} \sum y = a_0 n + a_1 \sum x \\ \sum yx = a_0 \sum x + a_1 \sum x^2 \end{cases}$$ для прямої;

$$\begin{cases} \sum y = a_0 n + a_1 \sum x + a_2 \sum x^2 \\ \sum yx = a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 \\ \sum yx^2 = a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 \end{cases}$$ для параболи;

$$\begin{cases} \sum y = a_0 n + a_1 \sum x + a_2 \sum x^2 + a_3 \sum x^3 \\ \sum yx = a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 + a_3 \sum x^4 \\ \sum yx^2 = a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 + a_3 \sum x^5 \\ \sum yx^3 = a_0 \sum x^3 + a_1 \sum x^4 + a_2 \sum x^5 + a_3 \sum x^6 \end{cases}$$ для кубічного рівняння.

Загальний вигляд системи рівнянь у матричному записі: Y = Z A.

Для подальшої роботи доцільно обчислити проміжні значення, такі як $\sum x^2, \sum x^3$ і т.д. Використовуючи отримані суми, складаємо матриці для системи нормальних рівнянь. Наприклад, для побудови лінійного рівняння регресії будуть потрібні наступні матриці:

$$Y = \begin{pmatrix} \sum y \\ \sum yx \end{pmatrix}; \qquad Z = \begin{pmatrix} n & \sum x \\ \sum x & \sum x^2 \end{pmatrix}$$

Таким чином, для знаходження значень матриці коефіцієнтів регресії A треба знайти матрицю, зворотну Z (тобто Z^{-1}), і помножити її зліва на матрицю Y.

3. Методика виконання роботи

3.1. Генерація вихідних даних

Для початку роботи потрібно згенерувати значення двох змінних *x* і *y*, відповідно Табл. 2.2. Обсяг вибірки – 100 елементів.

Таблиця 2.2 – Варіанти завдань

№	Фактор (*x*)	Результат (*y*)
1	Заробітна плата (грн) 3000 – 10000	Споживання (грн) $y = 1500 + 0{,}5 \cdot x + 500 \cdot e$
2	Дохід (грн) 3500 – 11000	Заощадження (грн) $y = -1000 + 0{,}5 \cdot x + 300 \cdot e$
3	Кількість студентів 2000…10000	Кількість викладачів ВНЗ $y = 220 + 0{,}09 \cdot x + 50 \cdot e$
4	Ціна товару (грн) 15 – 50	Попит,кг $y = 200000 \cdot x^{-0{,}85} + 500 \cdot e$
5	Валовий національний продукт (млрд. грн.) 1 – 8	Особисті доходи (млн. грн.) $y = -0.4 + 0{,}95 \cdot x + 0{,}4 \cdot e$
6	Витрати на рекламу (млн. грн.) 0 – 10	Прибуток(млн. грн.) $y = 10 + 6 \cdot x - 0{,}3 \cdot x^2 + 2 \cdot e$
7	Грошова маса (млн. грн.) 100 – 350	Індекс цін (%) $y = 38 + 0{,}3 \cdot x + 7 \cdot e$
8	Індекс трудовитрат (%) 100 – 160	Індекс обсягу продукції (%) $y = 7 \cdot x^{0{,}6} + 2 \cdot e$

Згадувана в таблиці випадкова складова *e* має нормальний розподіл з одиничною дисперсією і нульовим математичним очікуванням. Значення *e* слід згенерувати окремо, за допомогою функції «*Генерація випадкових чисел*» статистичної надбудови. Цей же спосіб можна використовувати для генерації значень *x* (тип розподілу – «рівномірний», ліва і права межа – відповідно до варіанту завдання). Отримані значення *x* і *y* доцільно округлити до того чи іншого знака після коми (або до цілого), в залежності від порядку отриманих величин (залежить від варіанту). Для округлення використовується функція ОКРУГЛ (число; число розрядів). Приклад результату генерації даних і округлення можна бачити на Рис. 2.1. У подальшій роботі використовуються тільки округлені значення *x* і *y*.

E3 =ОКРУГЛ(C3:C102;0)

	A	B	C	D	E	F
1	*Сгенерированные значения*				*Округленные значения*	
2	**e**	**x2**	**x**	**y**	**Зарплата**	**Потребление**
3	-2,99478	9096256	3016,45	1510,83266	3016	1511
4	1,23524	17656804	4202,31	4218,77212	4202	4219
5	0,63869	18757561	4331,34	3985,01378	4331	3985
6	0,10819	84750436	9206,37	6157,27755	9206	6157
7	1,28046	11108889	3332,83	3806,646	3333	3807
8	0,25254	73719396	8586,41	5919,47741	8586	5919
9	-0,13763	33246756	5765,86	4314,11408	5766	4314
10	0,86556	88228449	9393,08	6629,31998	9393	6629
11	-0,44904	30902481	5559,5	4055,22664	5559	4055

Рис.2.1. – Приклад результату генерації і округлення даних.

3.2. Кореляційний аналіз

Для обчислення коефіцієнтів кореляції можна використовувати як функцію «*Кореляція*» статистичної надбудови, так і функцію КОРРЕЛ (*діапазон_x; діапазон_y*). Отримане значення можна округлити з урахуванням числа значущих розрядів у вихідних даних. Результати розрахунку наведені на Рис.2.2

=ОКРУГЛ(КОРРЕЛ(D3:D102;E3:E102);3)

C	D	E	F	G	H	I
е значения	*Округленные значения*					
y	**Зарплата**	**Потребление**				
1510,832661	3016	1511				
4218,772125	4202	4219				
3985,013782	4331	3985				
6157,277552	9206	6157				
3806,645996	3333	3807				
5919,477414	8586	5919				
4314,114081	5766	4314				
6629,319981	9393	6629				

	Зарплата	*Потребление*
Зарплата	1	
Потребление	0,915899763	1

Показатель	***Значение***
Регрессия	0,916
a0	
a1	

Рис.2.2 Приклад обчислення коефіцієнтів кореляції.

3.3. Регресійний аналіз

На Рис. 2.3 показаний приклад обчислення проміжних значень, таких як $\sum x^2, \sum x^3$ і т.д. Використовується функція СУММПРОИЗВ, яка дозволяє обчислити суму попарних добутків декількох стовпців.

f_x =СУММПРОИЗВ(D4:D103;D4:D103;E4:E103)

	C	D	E	F	G	H
					Вычисление промежуточных сумм	
1	2957,700105	3871	2958		n	100
1	6598,804167	9116	6599		Сумма *x*	628956
2	3353,568187	3997	3354		Сумма *y*	462708
3	2195,145292	3476	2195		Сумма *x*2	4426918638
7	6678,484727	8276	6678		Сумма *x*3	3,3934E+13
1	5978,923053	8847	5979		Сумма *x*4	2,76164E+17
3	3999,007407	3658	3999		Сумма *xy*	3152595054
5	5649,97021	6848	5650		Сумма *yx*2	2,3598E+13
2	5251,494479	8333	5251			

Рис.2.3 Приклад обчислення проміжних сум.

Для роботи з матрицями в пакеті *Excel* використовуються функції, що працюють з масивами. Матричні функції вводять в діапазон комірок, як описано нижче. Після введення матричних функцій, вони автоматично відображаються в фігурних дужках. На Рис. 2.4 наведено приклад матриць. Для знаходження оберненої матриці використовується функція МОБР (*матриця*_Z), для множення матриць – функція МУМНОЖ (*матриця*_Z-1; *матриця*_Y*)*.

Z		A
0,118880678	-1,67213E-05	=МУМНОЖ(H18:I19;G15:G16)
-1,67213E-05	2,56796E-09	

15	прямая					
16	Y	Z		Z-1		A
17	718089	100	718089	0,127976954	-1,64293E-05	-2,91038E-11
18	5593596543	718089	5593596543	-1,64293E-05	2,28792E-09	1
19						
20	парабола					
21	Y	Z		Z-1		A
22	718089	100	718089	5593596543	1,350055314	-0,000383858	2,53215E-08	0
23	5593596543	718089	5593596543	4,65092E+13	-0,000383858	1,12758E-07	-7,61314E-12	1
24	4,65092E+13	5,594E+09	4,65092E+13	4,06818E+17	2,53215E-08	-7,61314E-12	5,24664E-16	-6,93889E-18

Рис.2.4 Приклад роботи з матрицями

Дані функції повертають в якості результату не одне значення, а масиви чисел (діапазон комірок). Для того щоб отримати результат, виконайте наступні дії:

- оберіть діапазон комірок, в якому буде розташовуватися матриця, що є результатом обчислень матричної функції;
- введіть формулу в клітинку, що є лівим верхнім кутом обраного діапазону, натиснути *Enter*;
- виділіть область осередків (обраний діапазон), див. рис. 2.4;
- натисніть F2;
- натисніть Ctrl + Shift + Enter.

Коефіцієнти регресії можна також знайти за допомогою функції ЛИНЕЙН

При вивченні взаємозв'язків, необхідно побудувати діаграму розкиду (кореляційне поле): меню [*Вставка* → *Діаграма*]. На цій діаграмі вихідні дані (*x*, *y*)

показані точками. Сюди ж наноситься лінія регресії. Для цього необхідно сформувати допоміжні стовпці *x* і *y* для кожного виду регресії.

Стовпець допоміжних значень факторної ознаки *x* повинен містити кілька значень з постійним кроком від мінімального до максимального. Для цього в першу комірку вводимо початкове значення, обираємо діапазон значень і викликаємо [*Редагування → Заповнити → Прогресія*]. При цьому потрібно обрати вид заповнення – *За стовпцями*, вид прогресії – *Арифметична*, крок та граничне значення. Кількість допоміжних проміжних значень фактора вибирають таким чином, щоб отримати на графіку гладку криву лінію.

Тип діаграми для ліній регресії – *Точкова діаграма зі значеннями*, з'єднаними гладкими лініями без маркерів, див. Рис. 2.5.

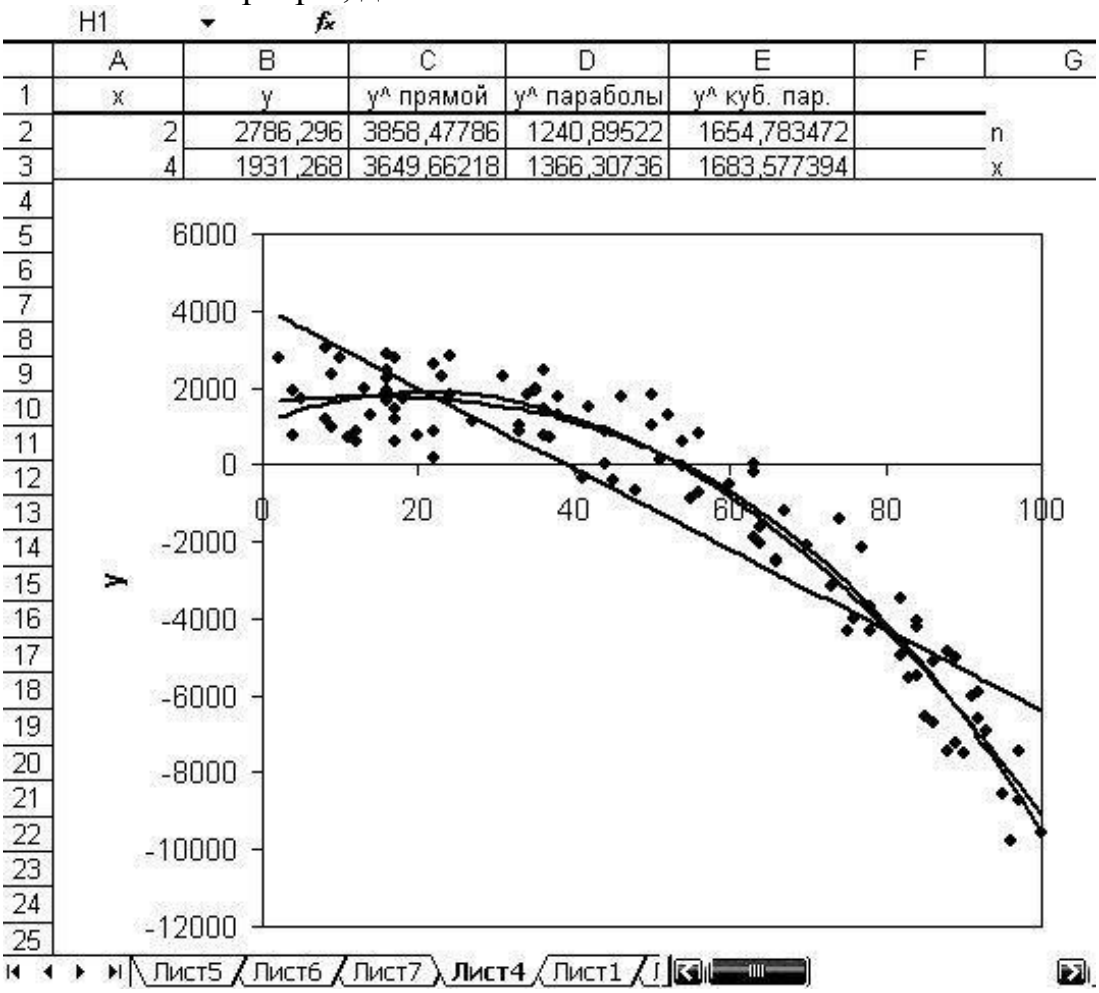

Рис.2.5.Приклад кореляційного поля з лініями регресії

3.4. Умовне середнє

Умовне середнє $\bar{y}/x$ – це середнє арифметичне значень результативної ознаки *y* за умови, що відповідні значення факторної ознаки *x* потрапляють в заданий інтервал. Додайте інтервали за *x*, які обираються за загальними правилами групування даних (див. Лабораторну роботу №1).

Для знаходження умовного середнього можна використовувати функцію СУММЕСЛИ, яка дозволяє обчислити суму при виконанні заданої умови. Формат функції наступний:

СУММЕСЛИ (*діапазон*; *критерій*; *діапазон_суммування*).

Діапазон - комірки, значення яких перевіряються за допомогою умови;

Критерій - умови підсумовування, наприклад, "<=" & W2;

Діапазон_суммування - комірки, значення яких складають при виконанні умови.

Отримана сума ділиться на кількість елементів, що потрапляють в діапазон. Для цього використовується функція СЧЕТЕСЛИ.

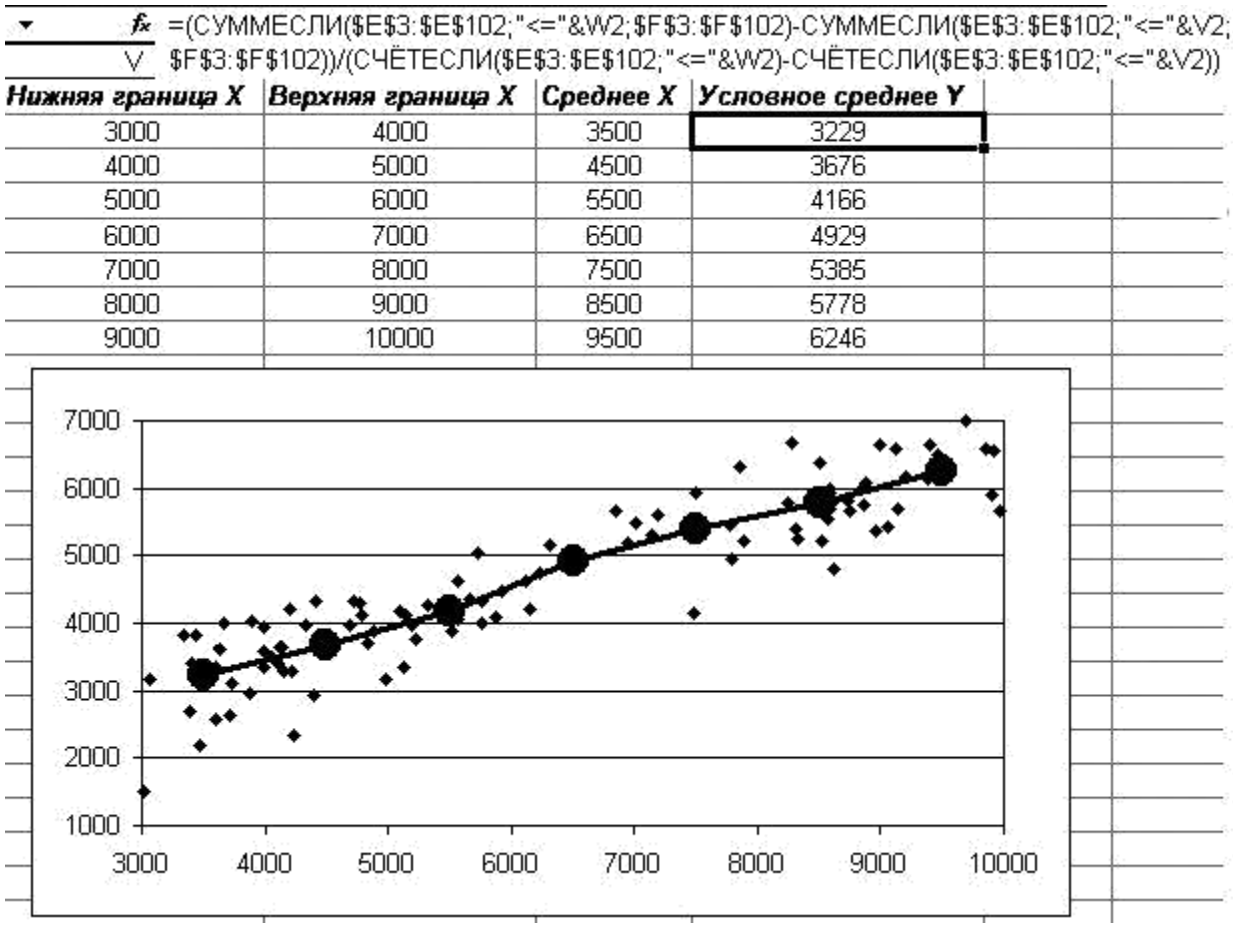

Рис.2.6.Приклад кореляційного поля з лінією умовного середнього

Формула для розрахунку умовного середнього може бути побудована в такий спосіб:

=(СУММЕСЛИ($E$3:$E$102;"<="&W2;$F$3:$F$102)-
СУММЕСЛИ($E$3:$E$102;"<="&V2;$F$3:$F$102))/
(СЧЁТЕСЛИ($E$3:$E$102;"<="&W2)-СЧЁТЕСЛИ($E$3:$E$102;"<="&V2))

Лінія умовного середнього (емпірична регресія) наноситься на кореляційне поле, див. Рис. 2.6. Як значення *x* беруться середини інтервалів, точки з'єднуються прямими лініями.

3.5. Вправа 5. Аналіз якості моделі зв'язку

Для аналізу отриманої моделі зв'язку використовують показник залишкової дисперсії:

$$\sigma^2_{\text{зал}} = \frac{\sum(y_i - \hat{y}(x_i))^2}{n - k - 1}$$

де *n* – обсяг вибірки, *k* – число коефіцієнтів рівняння регресії.

Залишки – це різниця між фактичним значенням (Точками на графіку) і теоретичним прогнозом (лінією регресії). Облік числа коефіцієнтів *k* компенсує поступове наближення лінії регресії до початкових точок на кореляційному полі за рахунок підвищення порядку моделі. Рекомендується обирати рівняння регресії, що дає найменшу залишкову дисперсію.

Вимоги до змісту та оформлення звіту

Звіт повинен бути продемонстрований як на паперовому носії, що містить статистичні показники, коефіцієнти рівнянь регресії, представлені у вигляді таблиці, кореляційне поле з лініями регресії, так і в електронній формі у вигляді файлу із заповненою таблицею та графіками.

Висновки за результатами аналізу взаємозв'язку соціально-економічних явищ можуть містити наступні положення:

- Яка інформація досліджувалася і якими методами;
- Чи є лінійними зв'язки між показниками;
- Чи збігаються результати, отримані різними способами;
- Які прогнози і рекомендації можна зробити.

Титульний аркуш звіту повинен містити всю інформацію, необхідну для однозначної ідентифікації авторів і роботи. Для цього на титульному аркуші вказують назву дисципліни, тему і номер роботи, варіант завдання, номер групи, прізвища та ініціали студентів, посаду, прізвище та ініціали викладача та інше.

Порядок виконання робот:

1.Ознайомтесь з описом наступних функцій Excel:
КОРРЕЛ, ОКРУГЛ, СУММПРОИЗВ, МОБР, МУМНОЖ, ЛИНЕЙН,СУММЕСЛИ.
2.Згенеруйте вихідні дані.
3.Розрахуйте коефіцієнт кореляції за допомогою надбудови і функції КОРРЕЛ.
4.Зробіть висновок про тісноту зв'язку ознак.
5.Розрахуйте коефіцієнти рівнянь регресії першого, другого і третього порядків за допомогою матричних функцій, функції ЛИНЕЙН і надбудови.
6. Побудуйте кореляційне поле.
7. Нанесіть на кореляційне поле лінії регресії.
8. Нанесіть на кореляційне поле лінію емпіричної регресії.
9. Розрахуйте значення залишкової дисперсії для кожного рівняння регресії.
10. Оформіть звіт відповідно до вимог.

Контрольні питання

1. Що таке кореляційна залежність?
2. Що вивчає кореляційний аналіз?

3. Як коефіцієнт кореляції характеризує взаємозв'язок параметрів?
4. Що вивчає регресійний аналіз?
5. Що таке регресія?
6. Як визначити параметри рівняння регресії?
7. Як визначити оптимальний вид рівняння регресії?
8. Що таке емпірична регресія?
9. Що таке умовне середнє?
10. Що таке кореляційне поле?

Критерії результативності виконання роботи

Лабораторна робота вважається виконаною в тому випадку, якщо студент:

1. Виконав всі зазначені завдання, дотримуючись порядку виконання роботи;
2. Освоїв методику виконання типових завдань і здатний продемонструвати роботу програми;
3. Представив звіт, що містить статистичні показники і коефіцієнти рівнянь регресії і кореляційне поле з лініями регресії;
4. Відповів на всі контрольні і додаткові питання.