МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет "Львівська політехніка"

Інтелектуальний аналіз даних за допомогою програмного пакета WEKA та MS Excel.

Регресійний аналіз. Лінійні одно- та двофакторні моделі.

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторної роботи № 7

з курсу **"Системи інтелектуального аналізу та візуалізації даних"**

для студентів за освітньою програмою Комп'ютерні науки (Проектування і програмування інтелектуальних систем та пристроїв)

Затверджено на засіданні кафедри
"Системи автоматизованого проектування"

Протокол N 1 від 28.08.2023р.

ЛЬВІВ 2023

1. МЕТА РОБОТИ

Метою роботи є засвоєння методів графічного(побудова лінії регресії) та математичного (розрахунок рівняння регресії та обчислення коефіцієнту регресії) проведення регресійного аналізу даних із застосуванням Weka та табличного процесору MS Excel.

2. КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

2.1. Короткі теоретичні відомості.

Метод регресійного аналізу є найпростішим і, одним з найменш ефективних методів інтелектуального аналізу даних. Найпростіша модель аналізу використовує один **вхідний** (**незалежний**) параметр і один **результуючий** (**залежний**) параметр (прикладами такої моделі є точкові діаграми Excel). Безумовно, модель можна ускладнити, додавши кілька десятків вхідних параметрів, але у будь-якому випадку загальний підхід буде один і той самий: на підставі декількох незалежних змінних визначається один залежний результат. Таким чином, модель регресійного аналізу використовується для прогнозування значення однієї залежної змінної, виходячи з відомих значень декількох незалежних параметрів.

Регресією називають апроксимацію даних з врахуванням їх статистичних параметрів. Таке завдання постає при обробці даних, отриманих в результаті вимірювань процесів або фізичних явищ. Завданням регресійного аналізу є підбір математичних формул, які найкращим чином можуть описати заданий набір.

Регресія - та ж класифікація, тільки замість категорії передбачається число. Вартість автомобіля залежно від пробігу, кількість корків за часом доби, обсяг попиту на товар від зростання компанії тощо. Регресією ідеально вирішуються завдання, де є залежність від часу.

Якщо регресія формує пряму лінію - її називають лінійною, якщо криву - поліноміальною. Це два основних види регресії (рис.1.).

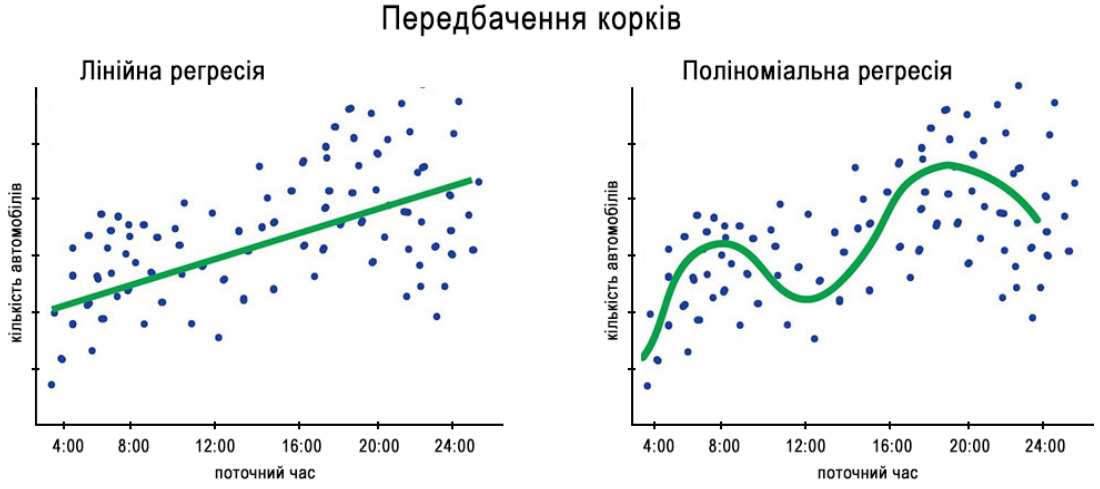


Рис. 1. Наочний приклад регресії

Схожість регресії і класифікації підтверджується ще і тим, що багато класифікаторів, після невеликих змін, перетворюються в регресійні моделі. Наприклад, можна не просто спостерігати до якого класу належить об'єкт, а запам'ятовувати, наскільки він близький - і тоді буде регресія.

Математична постановка задачі регресії полягає в наступному. Залежність величини певної властивості об’єкту Y від іншої змінної властивості або параметра X зареєстровано на множині точок множиною значень. В кожній точці зареєстровані значення відображено з випадковою похибкою. За сукупністю значень потрібно підібрати таку функцію, яка б з мінімальною похибкою відображала зареєстровані дані.

Види регресії називаються за типом апроксимуючих функцій: поліноміальна, експоненціальна, логарифмічна.

Вибірку даних, найчастіше, представляють у вигляді масиву, що складається з пар чисел (x_i, y_i). Тому, виникає завдання апроксимації дискретної залежності y(x) безперервною функцією f(x). Функція f(x), залежно від специфіки завдання, може відповідати різним вимогам:

- f(x) повинна проходити через точки (x_i, y_i) , тобто f (x_i) = yi , i = 1 ... n. В цьому випадку говорять про **інтерполяцію** даних функцією f (x) між точками x_i, або **екстраполяцію** за межами інтервалу, що містить всі x_i.
- f(x) повинна певним чином (наприклад, у вигляді певної аналітичної залежності) наближати y (x_i), не обов'язково проходячи через точки (x_i, y_i). Таку постановку завдання регресії в багатьох випадках можна назвати **згладжуванням** функції.
- f(x) повинна наближати експериментальну залежність y (x_i), враховуючи, що дані (x_i, y_i) отримано з деякою погрішністю, що виражає шумову компоненту вимірювань. При цьому функція f(x), за допомогою того чи іншого алгоритму, зменшує похибку, що присутня в даних (x_i, y_i). Такого типу задачі називають **фільтрацією**.

На рисунку проілюстровано різні види побудови апроксимуючої залежності f (x). Тут, вихідні дані позначено точками, інтерполяція - пунктиром, лінійна регресія (згладжування) - похилою прямою лінією, а фільтрація – грубою гладкою кривою (рис.2.).

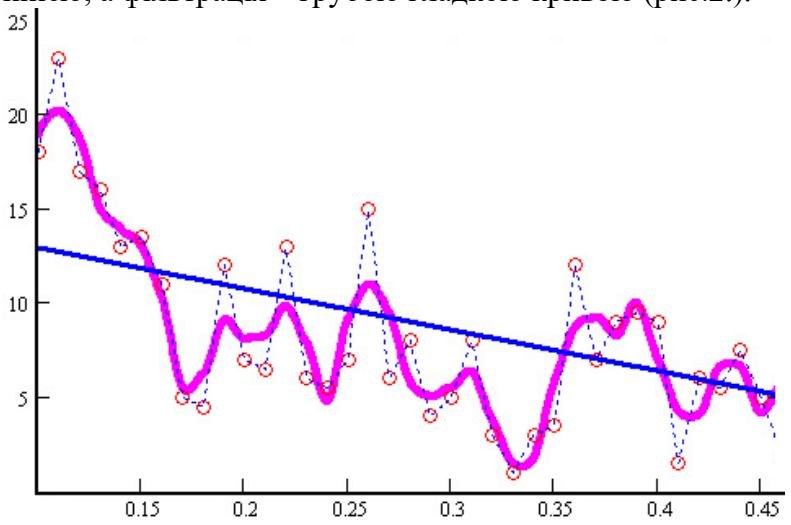


Рис.2. Різні види побудови апроксимуючої залежності f (x)

Для реалізації лінійної регресії часто використовують метод найменших квадратів. Тут вимірюється відстань по вертикалі від кожної точки до лінії. Необхідної лінією буде та конструкція, де сума відстаней буде мінімальною $\sum_{i=1}^{n}(y_i - \dot{\bar{y}}_i)^2$. Іншими словами, крива проводиться через точки, що мають нормально розподілене відхилення від істинного значення (рис.3.).

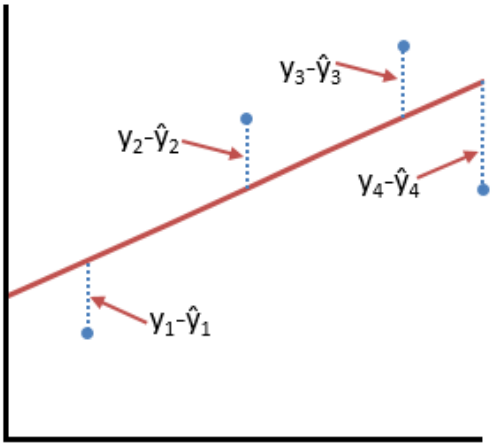


Рис.3. Наочний приклад методу найменших квадратів

Якщо лінійна функція може бути застосована для підбору даних, то метод найменших квадратів відноситься до типів метрики помилок, яка мінімізує похибки.

Найочевиднішим прикладом регресійного аналізу є визначення вартості будинку. Ціна на будинок (залежна змінна) визначається декількома незалежними параметрами: площею будинку і розміром ділянки, використанням в оформленні кухні гранітних плит, якістю і терміном служби сантехніки тощо.

Скористаємося моделлю регресійного аналізу для визначення ціни будинку і розберемо конкретний приклад. У таблиці 1 вказані фактичні параметри будинків, виставлених на продаж у певному районі. На підставі цих даних спробуємо оцінити вартість будинку в останньому рядку таблиці.

Таблиця 1. Регресійна модель оцінки вартості будинку

Площа будинку (кв.футів)	Розмір ділянки	Кількість спалень	Гранітна обробка на кухні	Сучасне сантехнічне обладнання?	Ціна продажу
3529	9191	6	0	0	$ 205,000
3247	10061	5	1	1	$ 224,900
4032	10150	5	0	1	$ 197,900
2397	14156	4	1	0	$ 189,900
2200	9600	4	0	1 `	$ 195,000
3536	19994	6	1	1	$ 325,000
2983	9365	5	0	1	$ 230,000
3198	***9669***	***5***	***1***	***1***	***????***

Розглянута модель дає лише найзагальніше, досить поверхневе, уявлення про метод регресійного аналізу. Проте, поверхневого розгляду цілком достатньо для того, щоб зрозуміти основні принципи і створити модель регресійного аналізу за допомогою WEKA.

Розглянемо наступні поняття:

- метод найменших квадратів,
- нормальний розподіл,
- коефіцієнт детермінації R-квадрат.

Метод найменших квадратів (МНК, OLS, Ordinary Least Squares) - математичний метод, який застосовується для вирішення різних задач, заснований на мінімізації суми квадратів деяких функцій від шуканих змінних. Він може використовуватися зокрема для апроксимації точкових значень деякою функцією. МНК є одним з базових методів регресійного аналізу для оцінки невідомих параметрів регресійних моделей за вибірковими даними.

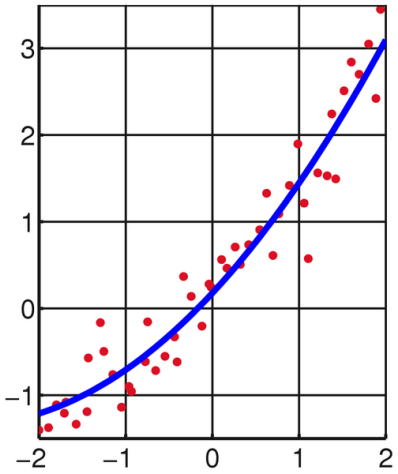


Рис. 4. Результат підгонки сукупності спостережень квадратичною функцією.

Нормальний закон розподілу (normal law of distribution) *(*який ще називається законом Гауса) відіграє виключно важливу роль у теорії імовірності і займає серед інших законів розподілу особливе положення. Це закон, який найчастіше зустрічається на практиці.

Більшість випадкових величин, таких, наприклад, як похибки вимірів, похибки гарматних стрільб тощо можуть бути подані як суми великої кількості малих доданків - елементарних похибок, кожна з яких визначається дією окремої причини, яка не залежить від інших. Яким би законам розподілу не підпорядковувались окремі елементарні похибки, особливості цих розподілів у сумі великої кількості доданків нівелюються і сума підпорядковується закону, що близький до нормального. Підсумовані похибки у загальній сумі повинні відігравати відносно малу роль.

Випадкова величина ξ нормально розподілена або підпорядковується закону розподілу Гауса, якщо її щільність розподілу має вигляд:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},$$

де a - довільне дійсне число, $\sigma > 0$.

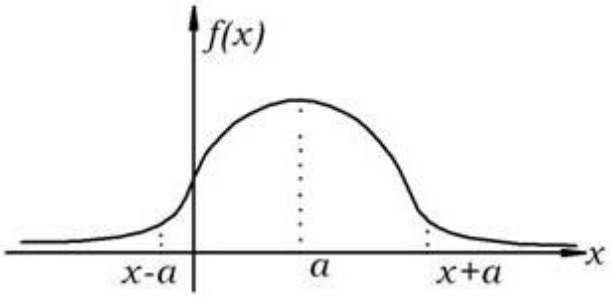


Рис. 5. Закон розподілу Гауса

Коефіцієнт детермінації (позначається як R^2 — R-квадрат) — статистичний показник, що використовується у статистичних моделях як міра залежності варіації залежної змінної від варіації незалежних змінних. Вказує наскільки отримані спостереження підтверджують модель.

2.2. Приклад регресійного аналізу у WEKA.

Для того щоб завантажити дані у WEKA, їх потрібно перетворити у формат, зрозумілий цьому програмному пакету. У моделях регресійного аналізу використовуються всього два типи даних: **NUMERIC та DATE**. Після того, як описані всі стовпці таблиці, потрібно додати дані по рядках, використовуючи комі як розділювач. Нижче наведено файл ARFF з даними про ціни на будинки, які використовуються для побудови тестової моделі. Зверніть увагу, що

у списку відсутній рядок з даними будинку, ціну для якого необхідно встановити. Зараз створюється регресійна модель на базі відомих параметрів і, отже, не можемо включити в неї параметри нашого будинку, оскільки ціна його невідома.

Файл даних для завантаження в WEKA

```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

Тепер, коли файл з даними готовий, його потрібно завантажити у WEKA. Запустіть WEKA і виберіть опцію **Explorer**. У результаті відкриється закладка **Preprocess** вікна Explorer. Клацніть на кнопці **Open File** і виберіть створений вами ARFF-файл. Вікно WEKA Explorer із завантаженими даними про будинках показано на рис. 6.

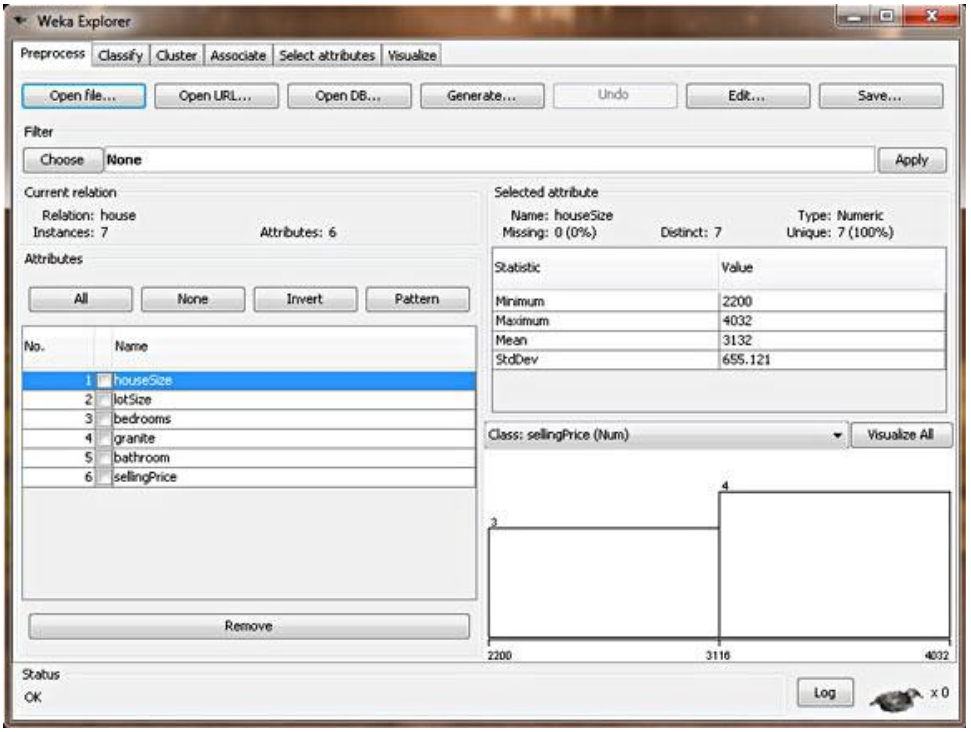


Рис. 6. Вікно WEKA Explorer із завантаженими даними про будинки

У цьому вікні можна перевірити дані, на підставі яких ми збираємося побудувати модель. У лівій частині вікна **Explorer** показані параметри об'єктів (**Attributes**), які відповідають заголовкам стовпців вхідної таблиці, а також вказана кількість об'єктів (**Instances**), тобто рядків таблиці. Якщо клацнути мишкою на одному із заголовків стовпців, то в правій панелі буде виведена повна інформація про набір даних у даному стовпці. Наприклад, якщо ми виберемо стовпець **houseSize** в лівій панелі (він обраний за замовчуванням), то у правій панелі

відобразиться додаткова статистична інформація з цього стовпця. Буде показано максимальне значення у стовпці (4032 кв.футів) і мінімальне значення (2200 кв.футів). Крім того, буде підраховано середнє значення (3131 кв.фут) і стандартне відхилення (655 кв.футів) (стандартне відхилення - статистичний показник розсіювання значень випадкової величини). Крім того, тут пропонується можливість візуального аналізу даних (кнопка **Visualize All**). Оскільки у таблиці даних не так багато, то їх візуальне відображення не дає такої наочної аналітичної картини, як у випадку використання сотень або тисяч показників.

Перейдемо від розгляду даних до створення моделі та визначимо, вартість будинку. Для того щоб створити модель, відкрийте закладку **Classify**. У якості першого кроку, нам потрібно вибрати тип моделі для аналізу (вказуємо WEKA, яким чином ми хочемо аналізувати наші дані, і яку модель побудувати):

1. Клацніть на кнопку **Choose** і розгорніть меню **functions**.
2. Виберіть опцію **LinearRegression**.

Таким чином, ми вказали WEKA, що хочемо створити модель регресійного аналізу. Як ви помітили, меню включає велику кількість інших моделей. Зверніть увагу: у меню включена опція **SimpleLinearRegression**, проте ми не використовуємо її, оскільки цей тип моделі визначає значення залежної змінної за значеннями одного незалежного параметра, а у нас їх шість. Якщо ви вибрали правильну модель, то вікно WEKA Explorer має виглядати так, як показано на рис. 7.

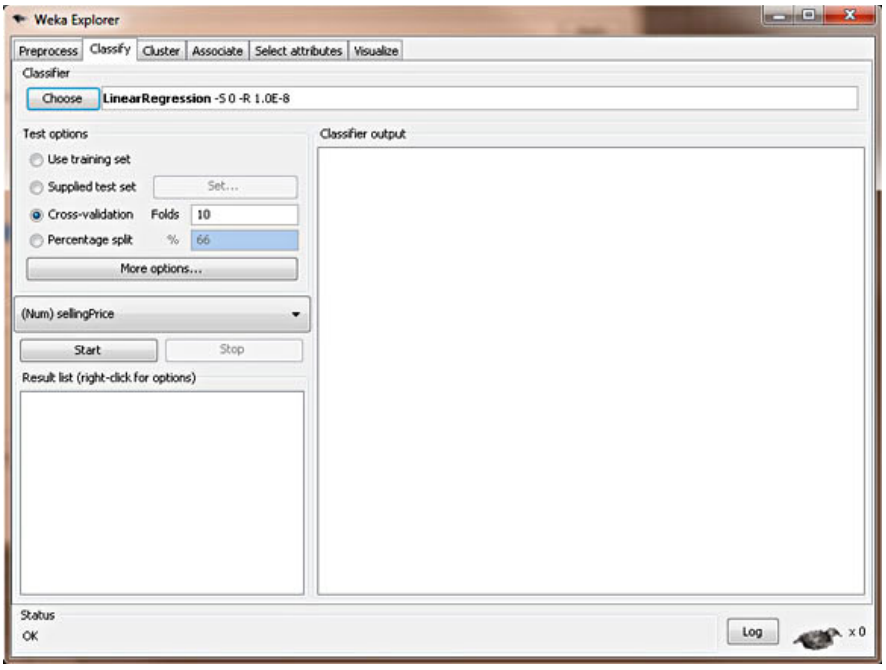


Рис. 7. Модель лінійного регресійного аналізу WEKA

Після того, як ми вибрали тип моделі, потрібно вказати WEKA, які дані повинні використовуватися для її створення. Незважаючи на те, що відповідь на це питання для нас цілком очевидна - потрібно взяти дані зі створеного нами ARFF-файлу - існує декілька інших, складніших можливостей надання даних для аналізу. Опція **Supplied test set** дозволяє вказати додатковий набір тестових даних для моделі, опція **Cross-validation** використовує кілька наборів даних, усереднює їх і будує модель на основі середніх значень, а опція **Percentage split** використовує в якості бази для моделі процентилі набору даних. Ці способи застосовуються для створення аналітичних моделей. У разі регресійного аналізу нам потрібна опція **Use training set**. У цьому випадку WEKA створить модель на базі даних із завантаженого ARFF-файлу.

Завершальний етап створення моделі - вибір залежної змінної (колонка, в якій знаходиться невідоме нам значення, яке потрібно розрахувати). У нашому прикладі - це ціна будинку, оскільки, саме це значення ми і хочемо дізнатися. Відразу після секції **Test options**

знаходиться список, що розкривається, в якому вам потрібно вибрати залежний параметр. Повинен бути вибраний атрибут **sellingPrice**. Якщо це не так, виберіть самі цей параметр.

Ми визначили всі параметри і можемо приступити до створення моделі. Натисніть кнопку **Start**. У результаті вікно WEKA має виглядати так, як показано на рис. 8.

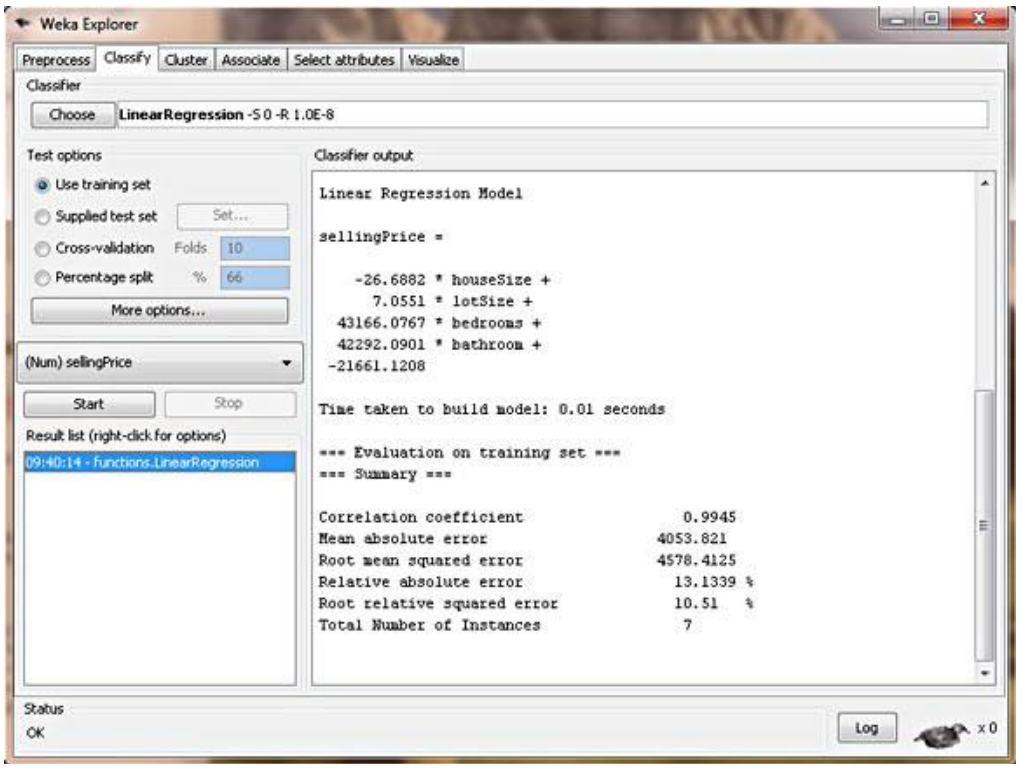


Рис. 8. Регресійна модель WEKA для розрахунку вартості будинку

Інтерпретація результатів регресійного аналізу. Розберемо, які дані включені в результуючий висновок:

```
sellingPrice = (-26.6882 * houseSize) +
               (7.0551 * lotSize) +
               (43166.0767 * bedrooms) +
               (42292.0901 * bathroom)
              - 21661.1208
```

Далі в отриману модель для визначення вартості підставляємо параметри нашого будинку.

Розрахунок вартості будинку на базі готової моделі

```
sellingPrice = (-26.6882 * 3198) +
               (7.0551 * 9669) +
               (43166.0767 * 5) +
               (42292.0901 * 1)
              - 21661.1208
sellingPrice = 219,328
```

Однак, можливості інтелектуального аналізу даних не обмежуються визначенням одного параметра. Основне завдання аналізу - виявлення залежностей і зв'язків у великих наборах даних. Інтелектуальний аналіз, як правило, використовується не для того, щоб визначити певне значення, а для того, щоб побудувати модель, що дозволяє аналізувати зв'язки між даними, прогнозувати результати і робити обґрунтовані висновки, які підтверджуються

зібраними статистичними даними. Тому не будемо обмежуватися розрахованою ціною будинку: розглянемо залежності між даними нашої моделі і постараємося зробити певні висновки щодо правил формування цін на нерухомість.

- **Гранітні елементи в оформленні кухні не впливають на ціну будинку** - WEKA використовує тільки ті дані, які, згідно зі статистикою, впливають на точність моделі (вплив кожного незалежного параметра на залежну змінну визначається за допомогою коефіцієнта детермінації). Таким чином, параметри, що не мають достатнього впливу на залежну змінну, в моделі не враховуються. Наша регресійна модель свідчить про те, що використання граніту на кухні не впливає на ціну будинку.
- **Стан ванних кімнат та сантехніки впливає на ціну будинку** - оскільки ми використовуємо значення 0 або 1 в якості показника модернізації ванних кімнат, то відповідний коефіцієнт в регресійній моделі демонструє нам, як сучасне сантехнічне обладнання впливає на ціну будинку, а саме додає 42292 $ до його ціни.
- **Велика площа будинку знижує його ціну** - Відповідно до моделі WEKA, у міру зростання площі будинків, ціна знижується. Це випливає з того, що модель включає змінну **houseSize** з негативним коефіцієнтом. Як це пояснити? Збільшення площі будинку на 1 кв.фут знижує його вартість на 26$. Подібне твердження здається очевидною нісенітницею. Проте, площа будинку не є незалежною величиною. Цей параметр пов'язаний, наприклад, із кількістю спалень - очевидно, що у великих будинках і кількість спалень більша. Так що наша модель, на жаль, не ідеальна, але ми можемо її спробувати покращити. Закладка **Preprocess** дозволяє видалити стовпці з набору даних.

У якості самостійної вправи, видаліть стовбець **houseSize** і створіть нову модель. Перевірте, як зміна набору даних відіб'ється на ціні будинку, і яка з двох моделей більше відповідає реальності (уточнена ціна будинку $ 217,894).

Розглянемо реальніший приклад. Для створення моделі скористаємося файлом даних, запропонованим в якості бази для регресійного аналізу на Web-сайті проекту WEKA. Теоретично, новий приклад буде дещо складнішим за нашу просту модель, що використовує дані про сім будинків. Запропонований файл призначений для створення регресійній моделі розрахунку витрати бензину (MPG - кількості миль на галон), виходячи з декількох параметрів автомобіля (дані збиралися з 1970 по 1982 рік). Модель враховує декілька параметрів машини - кількість циліндрів, робочий об'єм двигуна, його потужність, вагу автомобіля, час розгону, рік випуску, виробника і марку автомобіля. Цей набір даних містить 398 рядків і відповідає більшості вимог до статистичних даних, чого не можна сказати про наш попередній набір даних про будинки. Теоретично, модель на основі нового набору даних буде значно складнішою, і WEKA доведеться докласти більших зусиль на розробку нової моделі.

Для побудови моделі регресійного аналізу на основі нового набору даних вам слід виконати всі ті ж кроки, що і для моделі аналізу ціни будинку, так що ми не будемо приводити їх повторно. Висновок, який повинен вийти в результаті регресійного аналізу, показаний далі:

Модель регресійного аналізу для визначення MPG

```
class (aka MPG) =

-2.2744 * Cylinders = 6,3,5,4 +
-4.4421 * Cylinders = 3,5,4 +
 6.74 * cylinders = 5,4 +
 0.012 * displacement +
-0.0359 * Horsepower +
-0.0056 * Weight +
 1.6184 * model = 75,71,76,74,77,78,79,81,82,80 +
 1.8307 * model = 77,78,79,81,82,80 +
 1.8958 * model = 79,81,82,80 +
 1.7754 * model = 81,82,80 +
 1.167 * model = 82,80 +
 1.2522 * model = 80 +
```

```
2.1363 * origin = 2,3 +
37.9165
```

З точки зору виконання обчислень, створення потужних регресійних моделей на базі великих масивів даних, не викликає особливих проблем. Модель для визначення MPG може здатися набагато складнішою, ніж модель для визначення вартості будинку, тим не менш, це не так. Наприклад, перший рядок моделі, -2.2744 * cylinders = 6,3,5,4 означає, що якщо у машини 6-цілінрового двигун, то потрібно в формулу підставити 1, а якщо 8-циліндровий двигун - то 0. Давайте підставимо в модель реальні дані (наприклад, з рядка 10) і перевіримо, наскільки результат обчислень буде відповідати реальному показнику.

Обчислення показника MPG

```
      data = 8,390,190,3850,8.5,70,1,15
class (aka MPG) =
    -2.2744 * 0 +
    -4.4421 * 0 +
     6.74 * 0 +
     0.012 * 390 +
    -0.0359 * 190 +
    -0.0056 * 3850 +
     1.6184 * 0 +
     1.8307 * 0 +
     1.8958 * 0 +
     1.7754 * 0 +
     1.167 * 0 +
     1.2522 * 0 +
     2.1363 * 0 +
    37.9165

Expected Value = 15 mpg
Regression Model Output = 14.2 mpg
```

Таким чином, при використанні випадково вибраних даних, результат роботи нашої моделі (14.2 MPG) виявився досить близькою до реального показника (15 MPG).

Таблиця 2. Параметри налаштування методів

Метод	**Параметр**
LinearRegression	*attributeSelectionMethod* – метод відбору атрибутів. *eliminateColinearAttributes* – виключити колінеарні атрибути. *ridge* – штраф за великі значення коефіцієнтів регресії (регуляризація Тихонова).

2.3. Регресійний аналіз в Excel

Excel містить декілька функцій, які допоможуть вам обчислити оцінки за методом найменших квадратів. Два з них показані в таблиці 3.

Таблиця 3

Функція	**Опис**
INTERCEPT(*y*, *x*)	Обчислює оцінку за методом найменших квадратів, а, для відомих значень y і x.

SLOPE(y, x)	Обчислює оцінку за методом найменших квадратів b для відомих значень y і x.

Наприклад, якщо значення y знаходяться в діапазоні клітинок A2:A11, а значення x знаходяться в діапазоні B2:B11, тоді функція INTERCEPT(A2:A11, B2:B11) відобразить значення a, а функція SLOPE(A2:A11, B2:B11) відобразить значення b.

Аркуш «Рак молочної залози» містить дані дослідження 1965 року, що аналізує взаємозв'язок між середньою річною температурою та рівнем смертності жінок із певним типом раку молочної залози. Суб'єкти були з 16 різних регіонів Великобританії, Норвегії та Швеції. У таблиці 4 представлені дані.

Таблиця 4

Назва діапазону	**Діапазон**	**Опис**
Регіон	A2:A17	Число, що вказує на регіон, де були зібрані дані
Температура	B2:B17	Середньорічна температура по регіону
Смертність	C2:C17	Індекс смертності від новоутворень жіночої молочної залози по регіону

Визначимо чи є докази лінійної залежності між середньорічною температурою в регіоні та індексом смертності. Чи відрізняється індекс смертності для жінок, які проживають в регіонах з різними температурами?

	A	B	C	D	E
1	Region	Temperature	Mortality		
2	1	31.8	67.3		
3	2	34.0	52.5		
4	3	40.2	68.1		
5	4	42.1	84.6		
6	5	42.3	65.1		
7	6	43.5	72.2		
8	7	44.2	81.7		
9	8	45.1	89.2		
10	9	46.3	78.9		
11	10	47.3	88.6		
12	11	47.8	95.0		
13	12	48.5	87.0		
14	13	49.2	95.9		
15	14	49.9	104.5		
16	15	50.0	100.4		
17	16	51.3	102.5		
18					
19					
20					

Mortality DataRegion

Рис.9. Аркуш «Рак молочної залози»

Перш ніж обчислювати будь-яку статистику регресії, ви завжди повинні побудувати діаграму розсіювання. Діаграма розсіювання може швидко вказати на очевидні проблеми при припущенні, що лінійна модель відповідає нашим даним (можливо, діаграма розсіювання покаже, що значення даних не падають уздовж прямої лінії). Точкові діаграми в Excel також дозволяють накладати лінію регресії на графік разом із рівнянням регресії. З цієї інформації можна отримати хороше уявлення про те, чи відповідає пряма лінія нашим даним чи ні.

Як створювати діаграми розсіювання: https://support.microsoft.com/en-gb/topic/present-your-data-in-a-scatter-chart-or-a-line-chart-4570a80f-599a-4d6b-a155-104a9018b86e#:~:text=Click%20the%20Insert%20tab%2C%20and%20then%20click%20X%20Y%20Scatter%2C%20and,predefined%20sets%20of%20chart%20elements

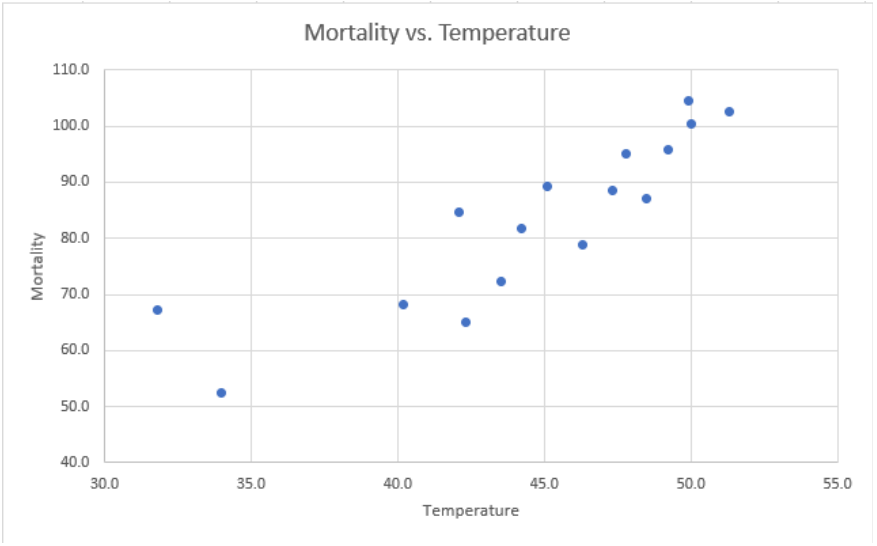


Рис.10. Діаграма розсіювання залежності індексу смертності від середньорічної температури

Щоб додати лінію регресії зробіть наступне :

1. Клацніть правою кнопкою миші будь-яку точку даних на графіку та виберіть у меню Add Trendline. Дивіться рисунок 11.

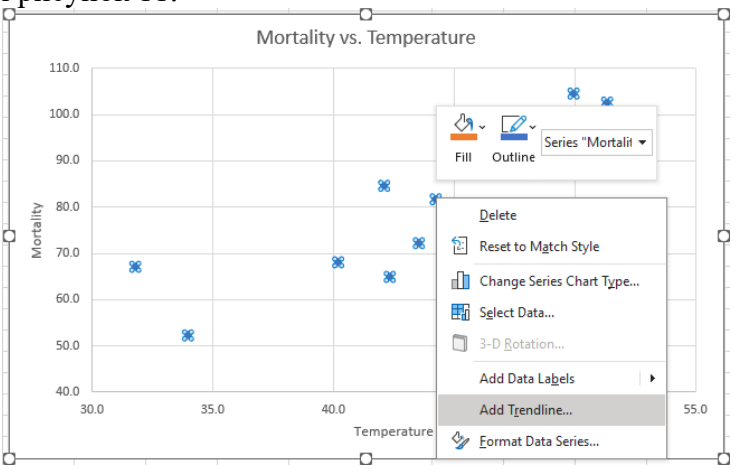


Рис.11. Вибір у меню Add Trendline

2. Excel відображає список можливих ліній регресії та тренду. Переконайтеся, що вибрано параметр «Лінійна регресія», як показано на рисунку 12.

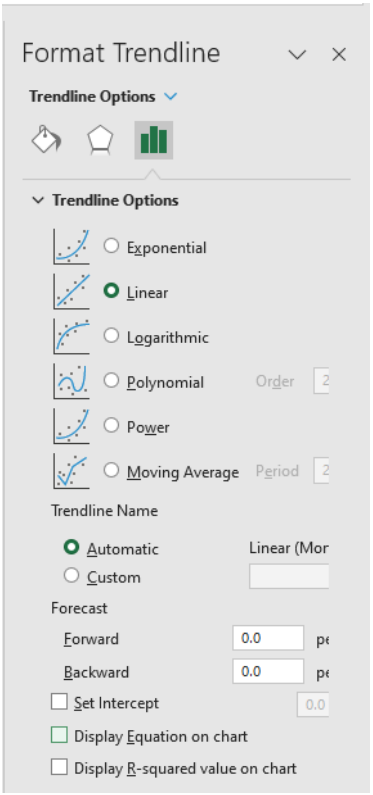


Рис.12. Вибір серед списку можливих ліній регресії та тренду

3. Установіть прапорці **Display Equation on chart** та **Display R-squared value on chart**, а потім натисніть кнопку **Close**.

Excel додає лінію регресії до графіка разом із рівнянням регресії та значенням R^2

4 Перетягніть текст, що містить рівняння регресії та значення R^2, у точку над графіком. Дивіться рис. 13.

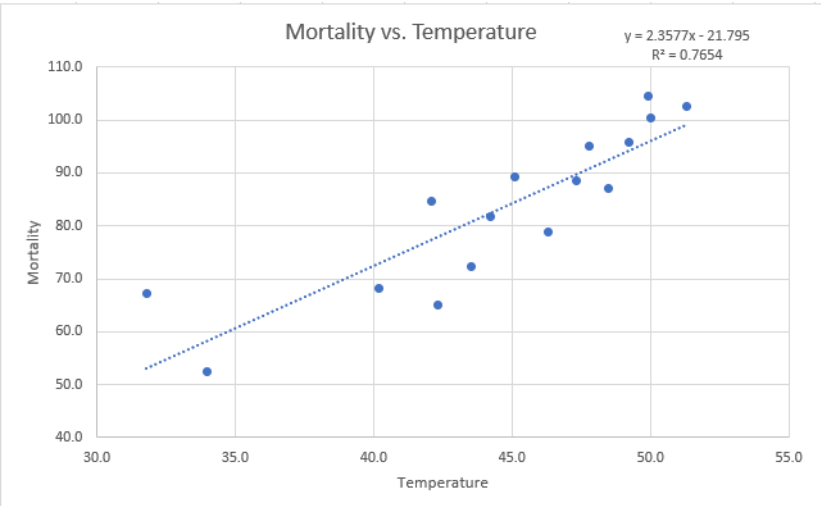


Рис. 13. Розташування рівняння регресії та R^2

Рівняння регресії для даних про смертність має вигляд y = - 21,795 + 2,3577x.

Це означає, що на кожен градус підвищення середньорічної температури в цих регіонах індекс смертності від раку молочної залози збільшувався приблизно на 2,3577 пункту.

Я можна інтерпретувати постійний член у цьому рівнянні (-21,7952)

На перший погляд, це точка перетину y, і це означає, що якщо середньорічна температура дорівнює 0, значення індексу смертності буде -21,795. Зрозуміло, що це абсурд; індекс смертності не може опускатися нижче нуля. Насправді будь-яка середньорічна температура нижче 9,24 градуса за Фаренгейтом призведе до негативної оцінки індексу смертності. Це не означає, що лінійне рівняння марне, але це означає, що ми повинні бути обережними, роблячи

будь-які прогнози для значень температури, які знаходяться за межами діапазону спостережених даних.

Значення R^2 становить 0,7654. Що це значить? Значення R^2, також відоме як коефіцієнт детермінації, вимірює відсоток варіації значень залежної змінної (у цьому випадку індексу смертності), який можна пояснити зміною незалежної змінної (температури). Значення R^2 змінюються від 0 до 1. Значення 0,7654 означає, що 76,54% варіації індексу смертності можна пояснити зміною середньорічної температури. Припускається, що решта 23,46% варіації пов'язані з випадковою змінністю.

Розрахунок регресійної статистики. Рівняння регресії на діаграмі розсіювання є корисною інформацією, але воно не говорить нам, чи є регресія статистично значущою. На цьому етапі у вас є дві гіпотези на вибір.

H_0: Не існує лінійної залежності між індексом смертності та середньорічною температурою.

H_a: Між індексом смертності та середньорічною температурою існує лінійна залежність.

Лінійний зв'язок, який ми перевіряємо, виражається через рівняння регресії.

Для виконання регресійного аналізу можна використовувати інструмент регресії (1) Data Analysis/Regression або (2) статистичні функції.

Створіть таблицю регресійної статистики:

1. Поверніться до аркуша даних про смертність.
2. Натисніть **Data Analysis** у групі Analysis на вкладці Data, щоб відкрити діалогове вікно Data Analysis.
3. Прокрутіть униз список інструментів аналізу даних і натисніть **Regression**, а потім натисніть кнопку **OK**.

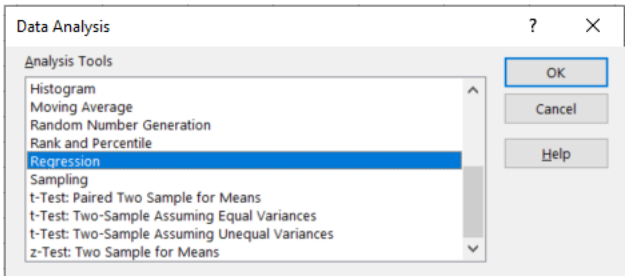


Рис. 14. Вибір Регресійного аналізу

4. Введіть діапазон клітинок **C1:C17** у полі «Діапазон введення Y» (це можна зробити, вибравши діапазон на робочому аркуші).
5. Введіть діапазон клітинок **B1:B17** у полі «Діапазон введення X».
6. Оскільки перша клітинка в цих діапазонах містить текстову мітку, клацніть прапорець **Labels**.
7. Натисніть на **New Worksheet Ply** кнопка опції та наберіть **Regression Statistics** у супровідному текстовому полі.
8. Установіть усі чотири прапорці у Залишках (Residuals).

З'явитися діалогове вікно регресії, як показано на рис. 15.

Зауважте, що ми не встановили прапорець Нормальні графіки ймовірностей (Normal Probability Plots). Ця опція створює звичайний графік залежної змінної. У більшості ситуацій цей графік не потрібний для регресійного аналізу.

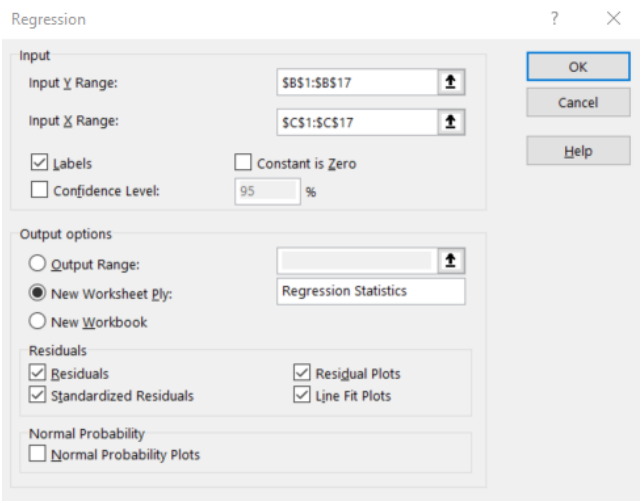
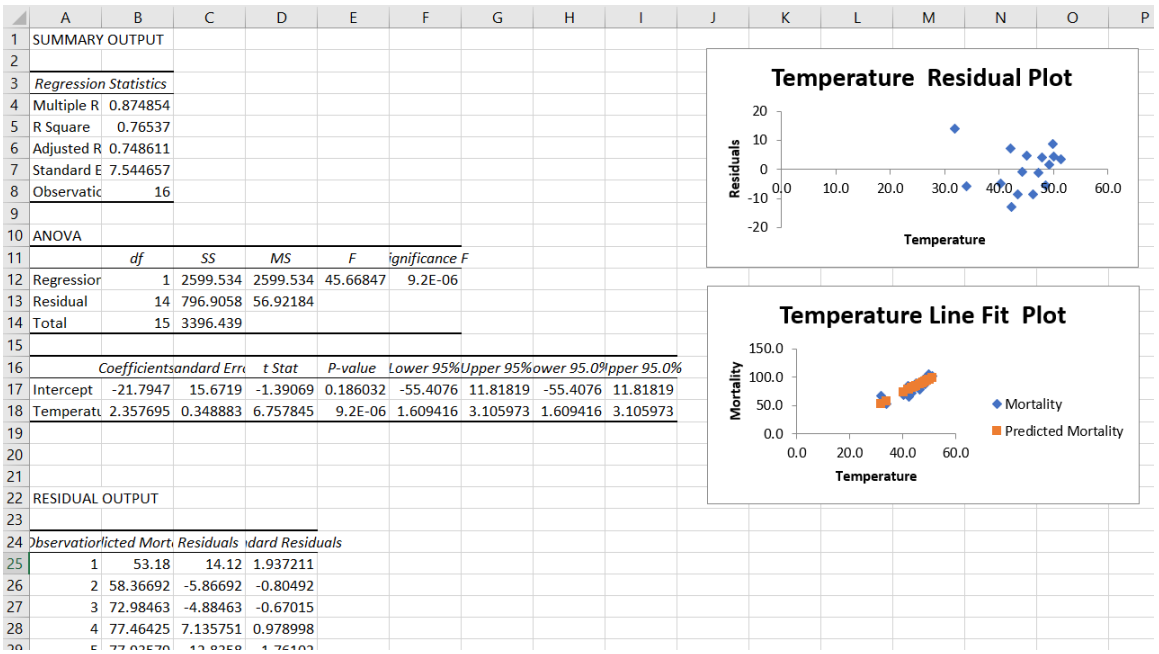


Рис. 15. Діалогове вікно регресії

9. Клацніть **OK.**

Excel генерує вихідні дані, показані на рис. 16.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.874854							
5	R Square	0.76537							
6	Adjusted R	0.748611							
7	Standard E	7.544657							
8	Observatic	16							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*ignificance F*			
12	Regressior	1	2599.534	2599.534	45.66847	9.2E-06			
13	Residual	14	796.9058	56.92184					
14	Total	15	3396.439						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-21.7947	15.6719	-1.39069	0.186032	-55.4076	11.81819	-55.4076	11.81819
18	Temperatu	2.357695	0.348883	6.757845	9.2E-06	1.609416	3.105973	1.609416	3.105973
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	*Observatior*	*licted Mort*	*Residuals*	*ndard Residuals*					
25	1	53.18	14.12	1.937211					
26	2	58.36692	-5.86692	-0.80492					
27	3	72.98463	-4.88463	-0.67015					
28	4	77.46425	7.135751	0.978998					

Рис. 16. Результат проведеного регресійного аналізу.

Вихідні дані розділені на шість областей: регресійна статистика, дисперсійний аналіз (ANOVA), оцінки параметрів, залишковий вихід, ймовірнісний вихід (не показано на рис. 16) і графіки. Розглянемо ці сфери детальніше.

Команда Regression не форматує вивід для нас, тому ми можемо захотіти зробити це самостійно на аркуші.

Інтерпретація дисперсійного аналізу. На рис. 17 показано результат команди регресії у таблиці ANOVA.

		df	*SS*	*MS*	*F*	*ignificance F*
12	Regressior	1	2599.534	2599.534	45.66847	9.2E-06
13	Residual	14	796.9058	56.92184		
14	Total	15	3396.439			

Рис. 17. Таблиці ANOVA

Таблиця ANOVA аналізує мінливість індексу смертності. Мінливість ділиться на дві частини: перша – це мінливість, зумовлена лінією регресії, а друга – випадкова мінливість.

Значення в колонці *df* таблиці вказують кількість ступенів свободи для кожної частини. Загальна кількість ступенів свободи дорівнює числу спостережень мінус 1. У цьому випадку загальна кількість ступенів свободи становить 15. З цих 15 ступенів свободи 1 ступінь свободи відноситься до регресії, а решта 14 ступенів свободи приписуються випадковій мінливості.

Стовпець *SS* містить суми квадратів. Загальна сума квадратів – це сума квадратів відхилень індексу смертності від загального середнього. Ця сума також ділиться на дві частини. Перша частина, позначена в таблиці як сума квадратів регресії, є сумою квадратів відхилень між лінією регресії та загальним середнім. Друга частина, позначена залишковою сумою квадратів, дорівнює сумі квадратів відхилень індексу смертності від лінії регресії. Це значення, ми хочемо зробити якомога меншим у рівнянні регресії. У цьому прикладі загальна сума квадратів становить 3396.44, з яких 2599.53 приписується регресії, а 796.91 приписується помилці.

Який відсоток від загальної суми квадратів можна віднести до регресії?

У цьому випадку це 2599.53/3396.44 = 0.7654, або 76.54%. Це дорівнює значенню R^2, яке вимірює відсоток мінливості, поясненої регресією. Зауважте також, що загальна сума квадратів (3396.44), поділена на загальну кількість ступенів свободи (15), дорівнює 226.43, що є дисперсією індексу смертності. Квадратний корінь із цього значення є стандартним відхиленням індексу смертності.

Стовпець *MS* (середній квадрат) відображає суму квадратів, поділену на ступені свободи. Зверніть увагу, що середній квадрат залишку дорівнює квадрату стандартної помилки в комірці B7 ($7.5447^2 = 56.9218$). Таким чином, ви можете використовувати середній квадрат для залишку, щоб отримати стандартну помилку.

Наступний стовпець відображає відношення середнього квадратичного для регресії до середньої квадратичної помилки залишків. Ця величина називається коефіцієнтом *F*. Велике співвідношення *F* вказує на те, що регресія може бути статистично значущою. У цьому прикладі коефіцієнт становить 45.7. Значення p відображається у наступному стовпці та дорівнює 0,0000092. Оскільки значення *p* менше 0.05, регресія є статистично значущою. Ви дізнаєтеся більше про дисперсійний аналіз та інтерпретацію таблиць ANOVA в наступному розділі.

Оцінки параметрів і статистика. Таблиця вихідних даних, створена за допомогою команди Analysis ToolPak Regression, відображає оцінки параметрів регресії разом із статистичними даними, що вимірюють їх значимість (див. рис.18.)

16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-21.7947	15.6719	-1.39069	0.186032	-55.4076	11.81819	-55.4076	11.81819
18	Temperatu	2.357695	0.348883	6.757845	9.2E-06	1.609416	3.105973	1.609416	3.105973

Рис.18. Оцінка параметрів регресії разом із статистичними даними.

Як ви вже бачили, постійний коефіцієнт, або перетин, дорівнює приблизно –21.79, а нахил на основі змінної температури становить приблизно 2.36. Стандартні похибки для цих значень показано в стовпці Standard Error і становлять 15.672 і 0.349 відповідно. Відношення оцінок параметрів до їхніх стандартних помилок відповідає *t*-розподілу з *n*-2 або 14 ступенями свободи. Співвідношення для кожного параметра показано в стовпці *t* Stat, а відповідні двосторонні значення *p* – у стовпці *P* value. У цьому прикладі значення *p* для терму перетину становить 0,186, а значення *p* для терму нахилу (позначеного як температура) становить 9,2 x 10^{-6}, або 0,0000092 (зверніть увагу, що це те саме значення *p*, яке з'явилося в таблиці ANOVA).

Остання частина цієї таблиці відображає 95% довірчий інтервал для кожного з термів. У цьому випадку 95% довірчий інтервал для терму перетину становить приблизно (255.41, 11.82), а 95% довірчий інтервал для нахилу становить (1.61, 3.11).

Примітка. У вихідних даних довірчі інтервали можуть з'являтися двічі. Перша пара, інтервал 95%, завжди з'являється. Друга пара з'являється завжди, але з рівнем достовірності,

який ви вказуєте в діалоговому вікні Regression. У цьому випадку ви використали стандартне значення 95%, тому цей інтервал відображається в обох парах.

Що ви дізналися з регресійної статистики? Перш за все, ви вирішили відхилити нульову гіпотезу та прийняти альтернативну гіпотезу про те, що існує лінійна залежність між індексом смертності та температурою. На основі довірчого інтервалу для параметра нахилу ви можете повідомити з 95% упевненістю, що для кожного градуса підвищення середньорічної температури індекс смертності для регіону збільшується від 1.61 до 3.11 балів.

Залишки та прогнозовані значення. Остання частина результату команди регресії Analysis ToolPak складається з залишків і прогнозованих значень. Дивіться рис. 19 (значення переформатовано, щоб їх було легше переглядати).

24	*Observation*	*Predicted Mortality*	*Residuals*	*Standard Residuals*
25	1	53.17999556	14.12000444	1.937211191
26	2	58.36692354	-5.866923536	-0.804919714
27	3	72.98462964	-4.884629644	-0.670152708
28	4	77.46424926	7.135750743	0.978998007
29	5	77.93578816	-12.83578816	-1.761021577
30	6	80.7650216	-8.565021604	-1.175088562
31	7	82.41540778	-0.715407777	-0.098151241
32	8	84.53733286	4.662667142	0.639700293
33	9	87.3665663	-8.466566298	-1.161580866
34	10	89.72426083	-1.124260831	-0.154244333
35	11	90.9031081	4.096891902	0.56207807
36	12	92.55349427	-5.553494271	-0.761918405
37	13	94.20388044	1.696119555	0.232701186
38	14	95.85426662	8.645733382	1.186161912
39	15	96.09003607	4.309963929	0.591310746
40	16	99.15503896	3.344961035	0.458916

Рис. 19. Залишки і прогнозовані значення

Залишки — це різниці між спостережуваними значеннями та лінією регресії (прогнозовані значення). У вихідні дані також включені стандартизовані залишки. Зі значень, наведених на рис.19, ви бачите, що є одна нев'язка, яка здається більшою за інші, вона знайдена у першому спостереженні та має стандартизоване залишкове значення 1,937. Стандартизовані залишки - це залишки, стандартизовані за загальною шкалою, незалежно від вихідної одиниці вимірювання. Стандартизований залишок, значення якого вище 2 або нижче -2, є потенційним викидом. Є багато способів обчислення стандартизованих залишків. Excel обчислює за такою формулою:

$$\text{Standardized residual} = \frac{\text{Residual}}{\sqrt{\text{Sum of squared residuals}/(n-1)}}$$

тут n – кількість спостережень у наборі даних. У цьому наборі даних значення першого стандартизованого залишку становить:

$$\frac{14.12}{\sqrt{796.9058/15}} = 1.937$$

Залишки відіграють важливу роль у визначенні відповідності регресійної моделі.

Перевірка регресійної моделі. Як і в будь-якій статистичній процедурі, для статистичного висновку щодо регресії роблять деякі важливі припущення. Їх є чотири:

1. Правильною є прямолінійна модель.
2. Терм помилки ε зазвичай розподіляється із середнім значенням 0.

3. Помилки мають постійну дисперсію.

4. Помилки незалежні одна від одної.

Кожного разу, коли ви використовуєте регресію ви повинні враховувати ці припущення. На щастя, регресія є дещо надійнішою, тому припущення не потребують повного виконання.

Один момент, який не можна підкреслити занадто сильно, полягає в тому, що значна регресія не є доказом того, що ці припущення не були порушені. Щоб переконатися, що ваші дані не порушують ці припущення, потрібно пройти серію тестів, які називаються **діагностикою**.

Перевірка прямолінійного припущення. Щоб перевірити, чи правильна прямолінійна модель, вам слід спочатку створити діаграму розсіювання даних, щоб візуально перевірити, чи дані якимось чином відхиляються від цього припущення. На рис.20 показано класичну проблему, яку ви можете побачити у своїх даних.

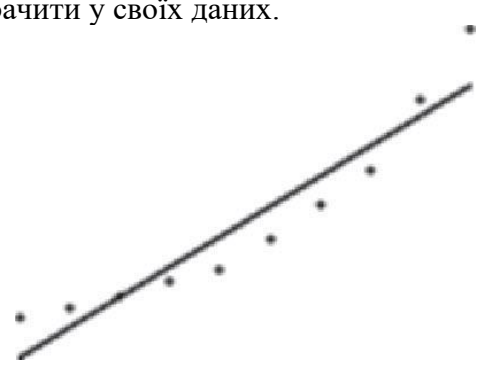


Рис. 20. Викривлені відносини

Інший точніший спосіб побачити, чи дані йдуть по прямій лінії, полягає в тому, щоб підібрати лінію регресії, а потім побудувати графік залишків регресії проти значень змінної предиктора. U-подібний (або перевернутий U-подібний) візерунок на графіку, як показано на рис. 21, є хорошим свідченням того, що дані мають викривлену залежність і що припущення прямої лінії є неправильним.

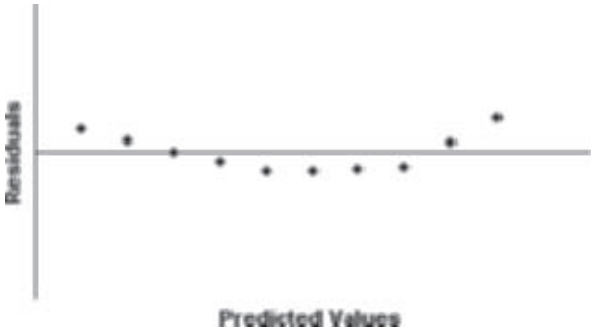


Рис. 21. Залишки, що показують викривлену залежність

Застосуємо цю діагностику до даних індексу смертності. Команда Regression створює цей графік для вас, але його може бути важко прочитати через його розмір. Перемістимо графік на аркуш діаграми та переформатуємо осі для зручності перегляду.

Щоб створити графік залежності залишків від змінної предиктора:

1. Перейдіть до комірки J1 на аркуші статистики регресії та клацніть **Temperature Residual Plot** правою клавішею.

2. Виберіть команду **Move Chart.**

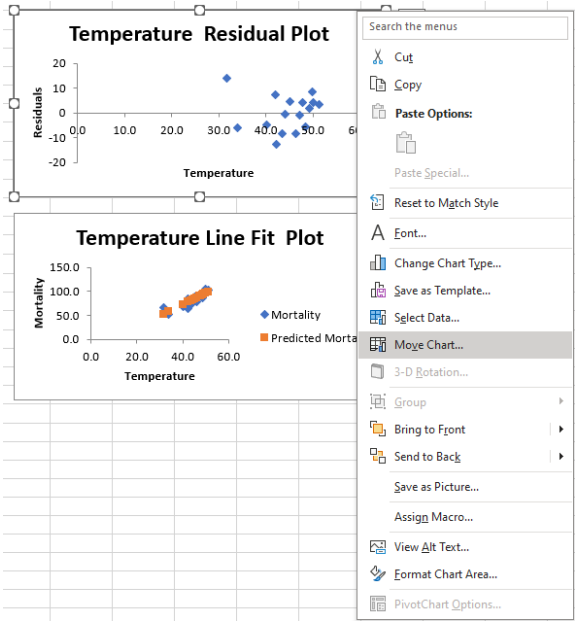


Рис. 22. Вибір Move Chart..

3. Натисніть кнопку параметра **New sheet**, а потім введіть Residuals vs.Temperature і натисніть OK.

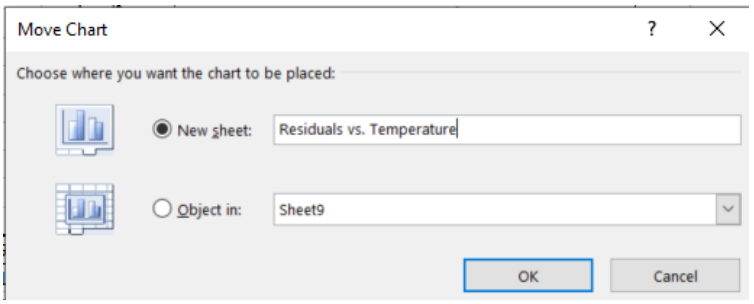


Рис.23. Вікно Move chart

4. Змініть масштаб горизонтальних осей змінної температури так, щоб нижня межа становила 30. Переглянутий графік відображається на рис. 22.

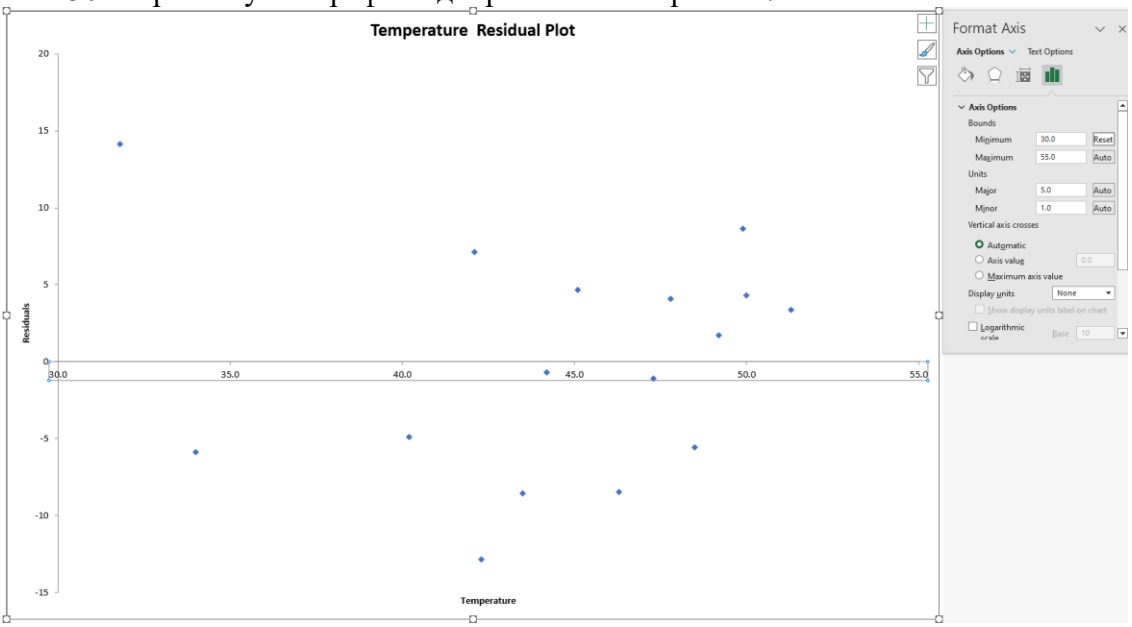


Рис. 24. Залишки проти значень температури

Діаграма показує, що більшість позитивних залишків, як правило, знаходяться при нижчих і вищих температурах, а більшість негативних залишків зосереджені при середніх

температурах. Це може вказувати на криву в даних. Велике перше спостереження має тут вплив. Без нього було б менше ознак кривої.

2.4. Використання статистичних формул для проведення статистичного аналізу

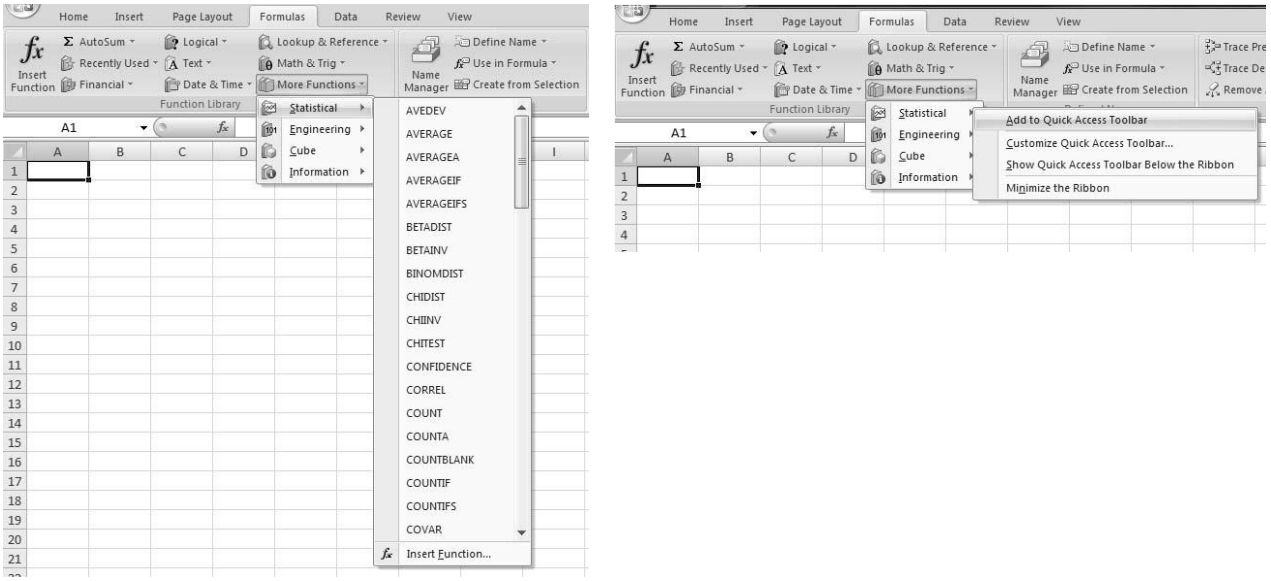


Рис.24.

Дані, які використовуються у прикладах:

Student	SAT	GPA
1	990	2.20
2	1150	3.20
3	1080	2.60
4	1100	3.30
5	1280	3.80
6	990	2.20
7	1110	3.20
8	920	2.00
9	1000	2.20
10	1200	3.60
11	1000	2.10
12	1150	2.80
13	1070	2.20
14	1120	2.10
15	1250	2.40
16	1020	2.20
17	1060	2.30
18	1550	3.90
19	1480	3.80
20	1010	2.00

Корисно: використовуйте в своїх обчисленнях діапазони комірок (cell range). Клацніть правою кнопкою миші на вибраному діапазоні:

Клацнувши правою кнопкою миші на вибраному діапазоні комірок, відкриється це спливаюче меню.

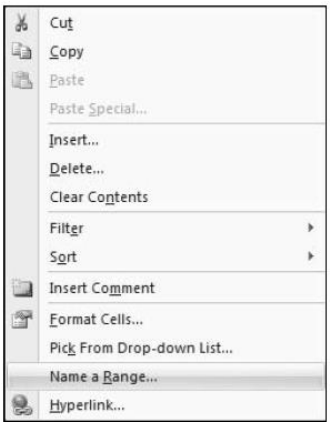


Рис.25.

Введіть ім’я діапазону (у прикладах це буде GPA та SAT):

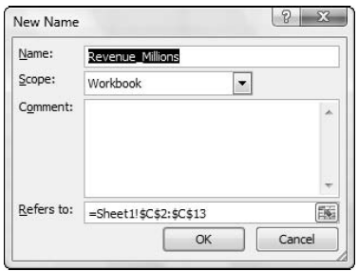


Рис.26.

Для обчислення кутового коефіцієнта прямої регресії за допомогою даних виберіть **SLOPE**.

Щоб розрахувати перетин із віссю y, виберіть **INTERCEPT**.

Щоб обчислити стандартну похибку оцінки, виберіть **STEYX**.

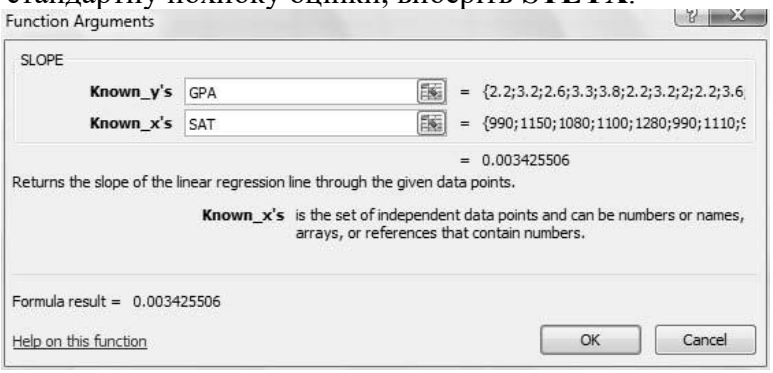


Рис.27.

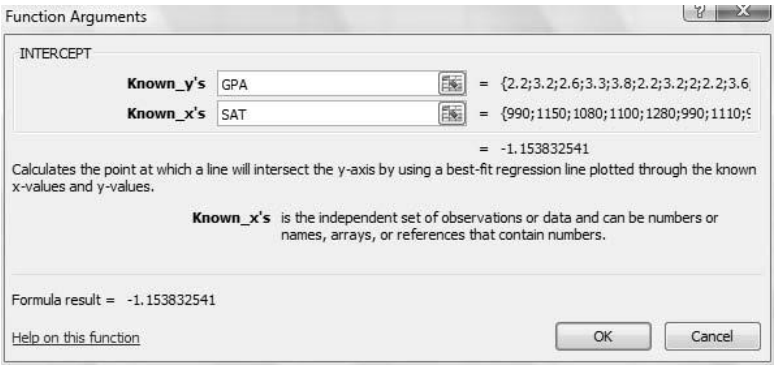


Рис.28.

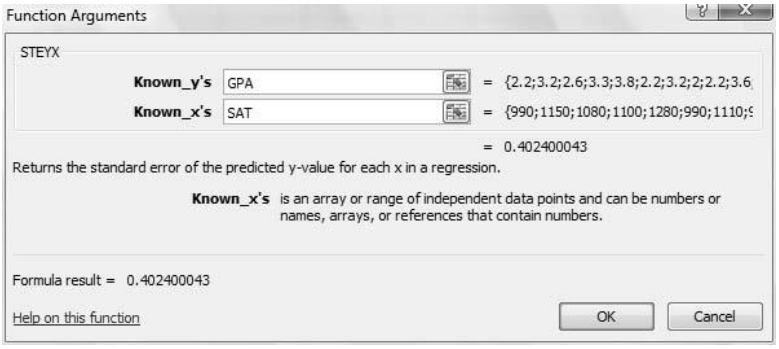


Рис.29.

FORECAST

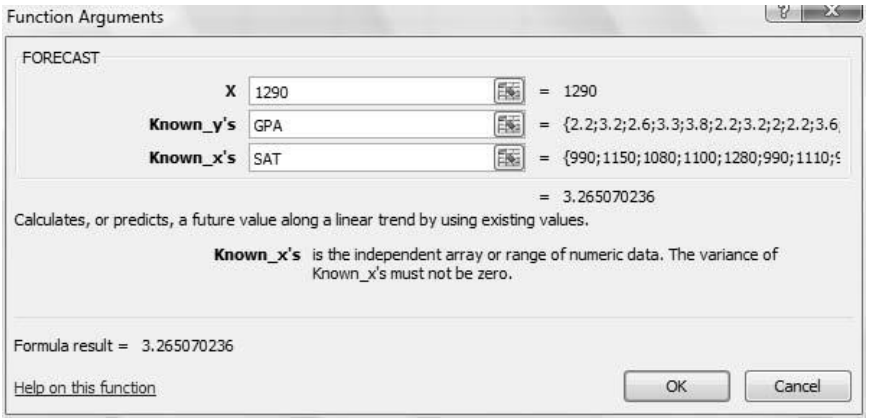


Рис.30.

TREND

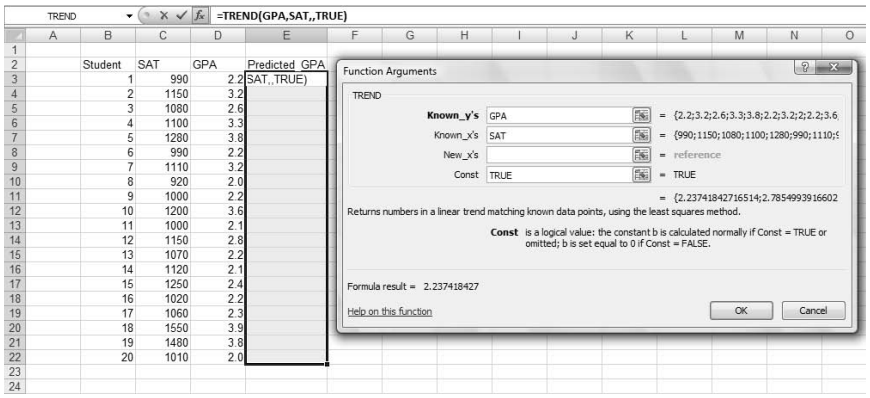


Рис.31.

	A	B	C	D	E	F
1						
2		Student	SAT	GPA	Predicted_GPA	
3		1	990	2.2	2.237418427	
4		2	1150	3.2	2.785499392	
5		3	1080	2.6	2.54571397	
6		4	1100	3.3	2.61422409	
7		5	1280	3.8	3.230815175	
8		6	990	2.2	2.237418427	
9		7	1110	3.2	2.648479151	
10		8	920	2.0	1.997633005	
11		9	1000	2.2	2.271673487	
12		10	1200	3.6	2.956774693	
13		11	1000	2.1	2.271673487	
14		12	1150	2.8	2.785499392	
15		13	1070	2.2	2.511458909	
16		14	1120	2.1	2.682734211	
17		15	1250	2.4	3.128049994	
18		16	1020	2.2	2.340183608	
19		17	1060	2.3	2.477203849	
20		18	1550	3.9	4.155701803	
21		19	1480	3.8	3.915916381	
22		20	1010	2.0	2.305928548	
23						

Рис.32.

Прогнозування нового набору y для нового набору x:

ВАЖЛИВО: НЕ натискайте кнопку ОК. Оскільки це функція масиву, натисніть Ctrl + Shift + Enter, щоб помістити відповіді **TREND** у вибраний масив.

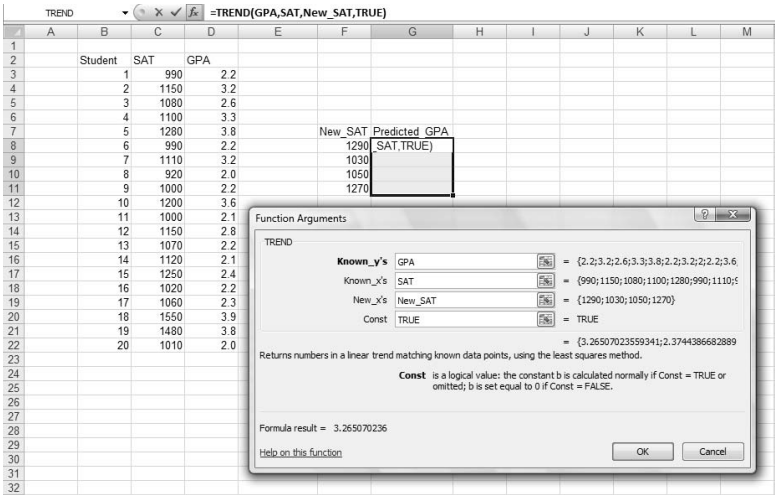
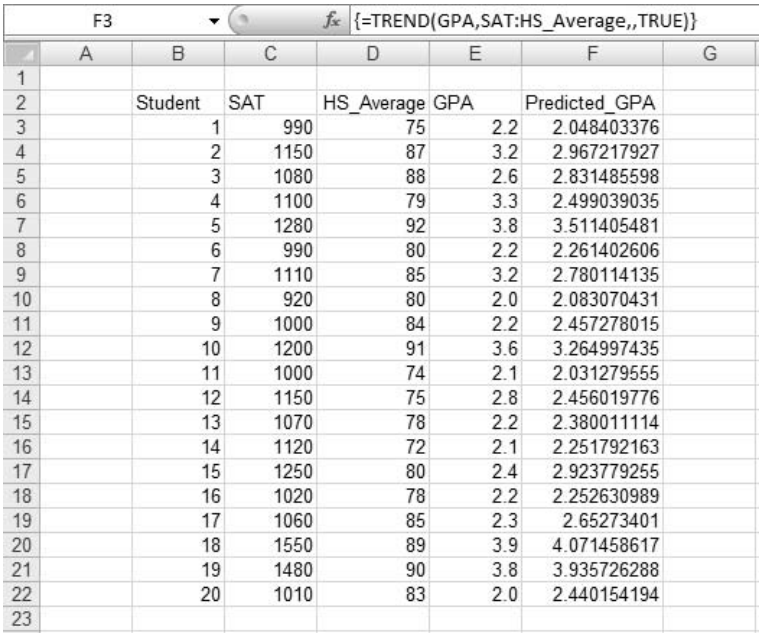


Рис.33.

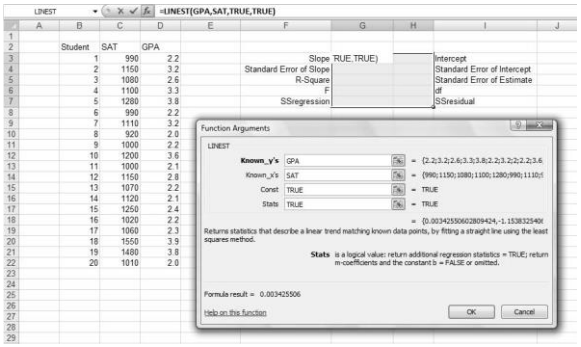


	Student	SAT	HS_Average	GPA	Predicted_GPA
1	990	75	2.2	2.048403376
2	1150	87	3.2	2.967217927
3	1080	88	2.6	2.831485598
4	1100	79	3.3	2.499039035
5	1280	92	3.8	3.511405481
6	990	80	2.2	2.261402606
7	1110	85	3.2	2.780114135
8	920	80	2.0	2.083070431
9	1000	84	2.2	2.457278015
10	1200	91	3.6	3.264997435
11	1000	74	2.1	2.031279555
12	1150	75	2.8	2.456019776
13	1070	78	2.2	2.380011114
14	1120	72	2.1	2.251792163
15	1250	80	2.4	2.923779255
16	1020	78	2.2	2.252630989
17	1060	85	2.3	2.65273401
18	1550	89	3.9	4.071458617
19	1480	90	3.8	3.935726288
20	1010	83	2.0	2.440154194

Рис.34.

LINEST

LINEST поєднує в собі SLOPE, INTERCEPT і STEYX, а також додає декілька додаткових послуг. На рисунку показано діалогове вікно «Аргументи функцій» для LINEST разом із даними та вибраним масивом для відповідей. Зауважте, що це масив з п'яти рядків на два стовпці. Для лінійної регресії таким повинен бути вибраний масив.



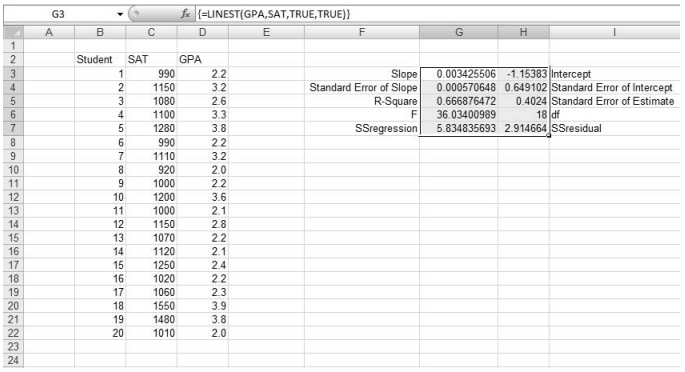


Рис.35.

Регресійна статистика

Регресійна статистика	
Множинний R	0,998364
R-квадрат	0,99673
Нормований R-квадрат	0,996321
Стандартна помилка	0,42405
Спостереження	10

Спочатку розглянемо верхню частину розрахунків, представлену у таблиці -регресійну статистику.

Величина R-квадрат, яка також називається мірою визначеності, характеризує якість отриманої регресійної прямої. Це якість виражається ступенем відповідності між вихідними даними і регресійної моделі (розрахунковими даними). Міра визначеності завжди знаходиться в межах інтервалу [0; 1].

Якщо значення R-квадрат близьке до одиниці, це означає, що побудована модель пояснює майже всю мінливість відповідних змінних. І навпаки, значення R-квадрата, близьке до нуля, означає погану якість побудованої моделі.

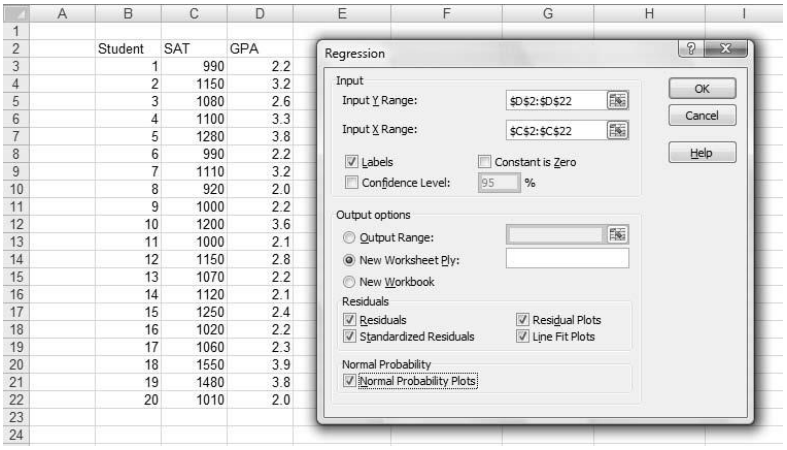
У нашому прикладі міра визначеності дорівнює 0,99673, що говорить про дуже хорошу підгонку регресійної прямої до вихідних даних.

Множинний R - коефіцієнт множинної кореляції R - висловлює ступінь залежності між незалежною змінною (X) і залежною змінною (Y).

Множинний R дорівнює квадратному кореню з коефіцієнта детермінації, ця величина приймає значення в інтервалі від нуля до одиниці.

Інструмент аналізу даних: регресія

Інструмент аналізу даних регресії Excel робить все, що робить LINEST (і більше), і додає заголовки для вихідні даних. На рисунку показано діалогове вікно інструменту регресії разом із даними для прикладу SAT-GPA.



	Student	SAT	GPA
	1	990	2.2
	2	1150	3.2
	3	1080	2.6
	4	1100	3.3
	5	1280	3.8
	6	990	2.2
	7	1110	3.2
	8	920	2.0
	9	1000	2.2
	10	1200	3.6
	11	1000	2.1
	12	1150	2.8
	13	1070	2.2
	14	1120	2.1
	15	1250	2.4
	16	1020	2.2
	17	1060	2.3
	18	1550	3.9
	19	1480	3.8
	20	1010	2.0

Рис.36.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.81662505
R Square	0.666876472
Adjusted R Square	0.648369609
Standard Error	0.402400043
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.834835693	5.834835693	36.03400989	1.12048E-05
Residual	18	2.914664307	0.161925795		
Total	19	8.7495			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.153832541	0.649101962	-1.777582888	0.092372108	-2.517545157	0.209880076	-2.517545157	0.209880076
SAT	0.003425506	0.000570648	6.002833489	1.12048E-05	0.002226619	0.004624393	0.002226619	0.004624393

Рис.37.

RESIDUAL OUTPUT

Observation	*Predicted GPA*	*Residuals*	*Standard Residuals*
1	2.237418427	-0.037418427	-0.095536221
2	2.785499392	0.414500608	1.058297332
3	2.54571397	0.05428603	0.138602356
4	2.61422409	0.68577591	1.750913753
5	3.230815175	0.569184825	1.453234976
6	2.237418427	-0.037418427	-0.095536221
7	2.648479151	0.551520849	1.408135554
8	1.997633005	0.002366995	0.006043379
9	2.271673487	-0.071673487	-0.182995776
10	2.956774693	0.643225307	1.642274131
11	2.271673487	-0.171673487	-0.43831442
12	2.785499392	0.014500608	0.037022757
13	2.511458909	-0.311458909	-0.795212664
14	2.682734211	-0.582734211	-1.487829085
15	3.128049994	-0.728049994	-1.858847373
16	2.340183608	-0.140183608	-0.357914887
17	2.477203849	-0.177203849	-0.452434465
18	4.155701803	-0.255701803	-0.652854376
19	3.915916381	-0.115916381	-0.295956132
20	2.305928548	-0.305928548	-0.78109262

PROBABILITY OUTPUT

Percentile	*GPA*
2.5	2
7.5	2
12.5	2.1
17.5	2.1
22.5	2.2
27.5	2.2
32.5	2.2
37.5	2.2
42.5	2.2
47.5	2.3
52.5	2.4
57.5	2.6
62.5	2.8
67.5	3.2
72.5	3.2
77.5	3.3
82.5	3.6
87.5	3.8
92.5	3.8
97.5	3.9

Рис.38.

Залишки. За допомогою цієї частини звіту ми можемо бачити відхилення кожної точки від побудованої лінії регресії. Найбільше абсолютне значення залишку в нашому випадку - 0,778, найменше - 0,043. Для кращої інтерпретації цих даних скористаємося графіком вихідних даних і побудованої лінією регресії, представленими на рис. 8.3. Як бачимо, лінія регресії досить точно "підігнана" під значення вихідних даних.

Слід враховувати, що даний приклад є досить простим і далеко не завжди можлива якісна побудова регресійної прямої лінійного виду.

Графічне відображення.

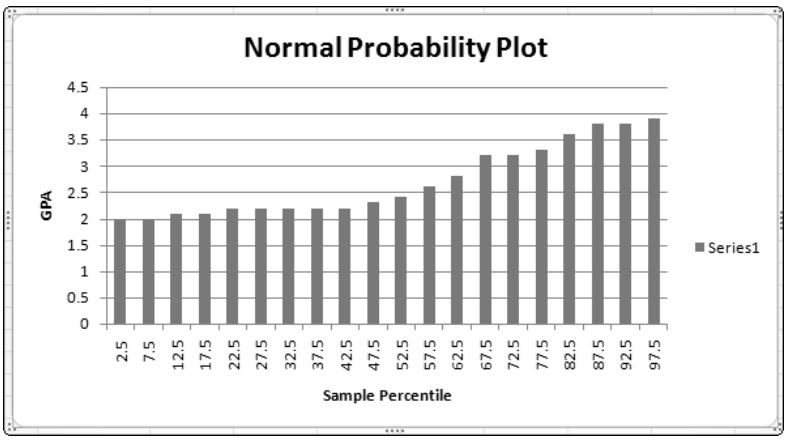


Рис.39.

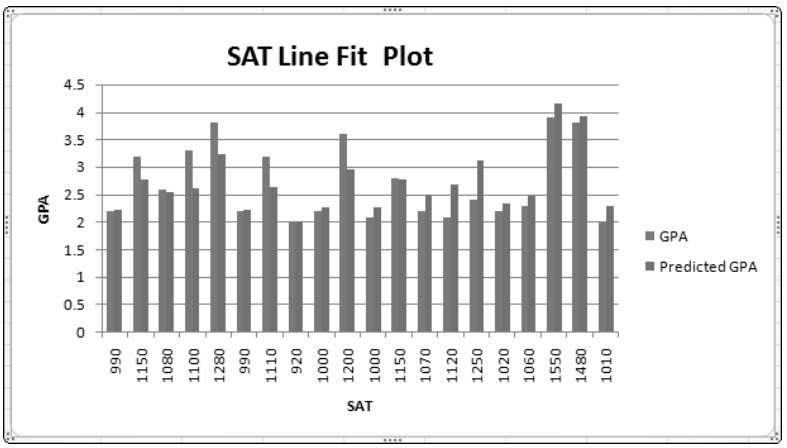


Рис.40.

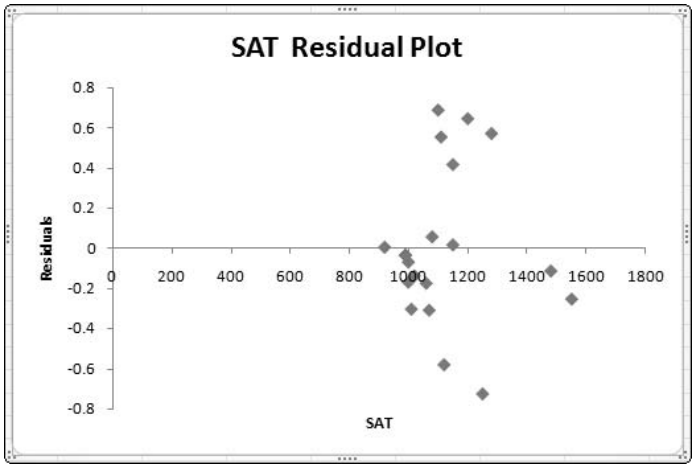


Рис.41.

2.5. Множинна регресія у Excel.

$$y' = a + b_1 x_1 + b_2 x_2$$

Можна перевірити гіпотези щодо загальної придатності та про всі три коефіцієнти регресії.

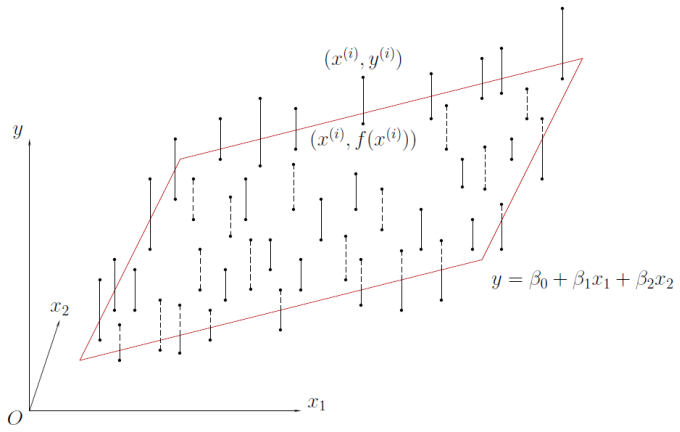


Рис. 25.

$$Predicted\ GPA = a + b_1\left(SAT\right) + b_2\left(High\ School\ Average\right)$$

Множинна регресія: **TREND**

TREND =TREND(GPA,SAT:HS_Average,,TRUE)

	A	B	C	D	E	F
1						
2		Student	SAT	HS_Average	GPA	Predicted_GPA
3		1	990	75	2.2	rage,,TRUE)
4		2	1150	87	3.2	
5		3	1080	88	2.6	
6		4	1100	79	3.3	
7		5	1280	92	3.8	
8		6	990	80	2.2	
9		7	1110	85	3.2	
10		8	920	80	2.0	
11		9	1000	84	2.2	
12		10	1200	91	3.6	
13		11	1000	74	2.1	
14		12	1150	75	2.8	
15		13	1070	78	2.2	
16		14	1120	72	2.1	
17		15	1250	80	2.4	
18		16	1020	78	2.2	
19		17	1060	85	2.3	
20		18	1550	89	3.9	
21		19	1480	90	3.8	
22		20	1010	83	2.0	
23						
24						

Function Arguments

TREND

Known_y's GPA = {2.2;3.2;2.6;3.3;3.8;2.2;3.2;2;2.2;3.6;

Known_x's SAT:HS_Average = {990,75;1150,87;1080,88;1100,79;128

New_x's = reference

Const TRUE = TRUE

= {2.04840337632863;2.9672179268678

Returns numbers in a linear trend matching known data points, using the least squares method.

Const is a logical value: the constant b is calculated normally if Const = TRUE or omitted; b is set equal to 0 if Const = FALSE.

Formula result = 2.048403376

Help on this function

OK Cancel

Рис.26.

F3 {=TREND(GPA,SAT:HS_Average,,TRUE)}

	A	B	C	D	E	F	G
1							
2		Student	SAT	HS_Average	GPA	Predicted_GPA	
3		1	990	75	2.2	2.048403376	
4		2	1150	87	3.2	2.967217927	
5		3	1080	88	2.6	2.831485598	
6		4	1100	79	3.3	2.499039035	
7		5	1280	92	3.8	3.511405481	
8		6	990	80	2.2	2.261402606	
9		7	1110	85	3.2	2.780114135	
10		8	920	80	2.0	2.083070431	
11		9	1000	84	2.2	2.457278015	
12		10	1200	91	3.6	3.264997435	
13		11	1000	74	2.1	2.031279555	
14		12	1150	75	2.8	2.456019776	
15		13	1070	78	2.2	2.380011114	
16		14	1120	72	2.1	2.251792163	
17		15	1250	80	2.4	2.923779255	
18		16	1020	78	2.2	2.252630989	
19		17	1060	85	2.3	2.65273401	
20		18	1550	89	3.9	4.071458617	
21		19	1480	90	3.8	3.935726288	
22		20	1010	83	2.0	2.440154194	
23							

Рис.27.

Множинна регресія: **LINEST**

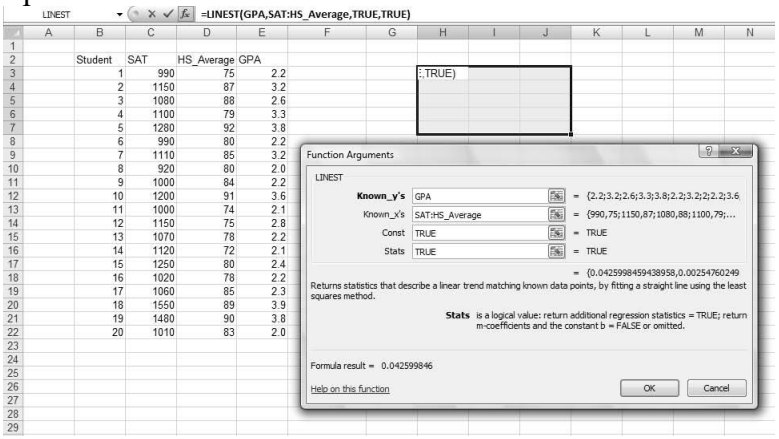
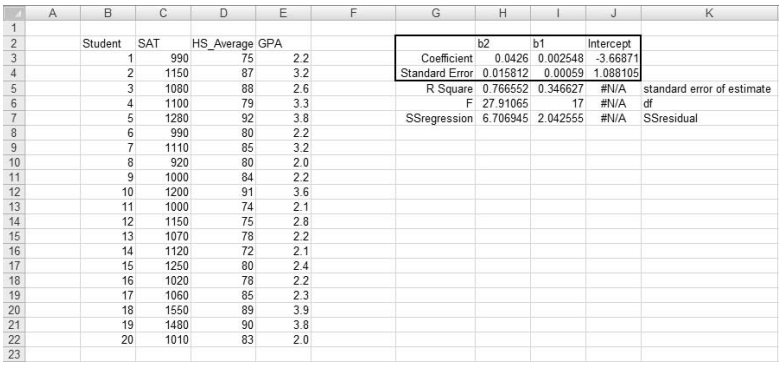


Рис.28.



	Student	SAT	HS_Average	GPA
	1	990	75	2.2
	2	1150	87	3.2
	3	1080	88	2.6
	4	1100	79	3.3
	5	1280	92	3.8
	6	990	80	2.2
	7	1110	85	3.2
	8	920	80	2.0
	9	1000	84	2.2
	10	1200	91	3.6
	11	1000	74	2.1
	12	1150	75	2.8
	13	1070	78	2.2
	14	1120	72	2.1
	15	1250	80	2.4
	16	1020	78	2.2
	17	1060	85	2.3
	18	1550	89	3.9
	19	1480	90	3.8
	20	1010	83	2.0

	b2	b1	Intercept	
Coefficient	0.0426	0.002548	-3.66871	
Standard Error	0.015812	0.00059	1.088105	
R Square	0.766552	0.346627	#N/A	standard error of estimate
F	27.91065	17	#N/A	df
SSregression	6.706945	2.042555	#N/A	SSresidual

Рис.29.

Два верхніх рядки масиву містять значення та стандартні помилки для коефіцієнтів.

Верхній рядок дає інформацію для написання рівняння регресії:

$$y' = -3.67 + .0025x_1 + .043x_2$$

$$Predicted\ GPA = -3.67 + .0025(SAT) + .043(High\ School\ Average)$$

Третій рядок містить R Square (міра міцності взаємозв'язку між GPA та двома іншими змінними та стандартну помилку оцінки. Порівняйте стандартну помилку оцінки для множинної регресії (0,35) зі стандартною помилкою для лінійної регресії (0.40).

У четвертому рядку показано коефіцієнт *F*, який перевіряє гіпотезу про те, чи добре лінія підходить до точкової діаграми, а *df* для знаменника *F*. *df* для чисельника (не показано) - кількість коефіцієнтів мінус 1. Можна використати FINV, щоб переконатися, що *F* з *df* = 2 і 17 є значущими.

4. ЛАБОРАТОРНЕ ЗАВДАННЯ

1. Проведіть однофакторний регресійний аналіз у Weka.

- Візьміть значення Y та X1 зі свого завдання.
- Підготуйте дані у Excel і сформуйте після цього arff файл (теж збережіть csv файл для наступних завдань).
- Вирішіть задачу регресії за допомогою методу Linear regression.
- Встановіть форму залежності і напрямок зв'язку між змінними - позитивна лінійна регресія, яка виражається в рівномірному зростанні функції;
- Встановіть напрямок зв'язку між змінними;
- Оцініть якість отриманої регресійної прямої;

- Визначіть відхилення розрахункових даних від даних вхідного набору;
- Передбачте майбутні значення залежної змінної.
- Які з атрибутів є найбільш значущими для передбачення значень цільового атрибуту, судячи з побудованих моделей? Як зміниться точність передбачення, якщо залишити лише значущі атрибути?
- Графічно передставте отримані результати.

2. Проведіть однофакторний регресійний аналіз в Excel

- Візьміть підготовані дані із завдання 1.
- Побудуйте лінію регресії.
- Сформуйте гіпотези щодо ваших даних.
- Розрахуйте регресійну статистику за допомогою інструменту регресії (1) Data Analysis/Regression та (2) статистичних функції.
- Інтерпретуйте дисперсійний аналіз.
- Оцініть параметрі і статистику.
- Проаналізуйте залишки та прогнозовані значення.
- Перевірте регресійну модель.
- Перевірте прямолінійне припущення.

3. Проведіть багатофакторний регресійний аналіз у Weka та Excel.

- Візьміть значення Y та X1, X2 зі свого завдання.
- Підготуйте дані у Excel. Сформуйте arff файл для аналізу у Weka.
- Побудуйте рівняння регресії.
- Опишіть отримані моделі і порівняйте їхню ефективність (точність передбачення).
- Які з атрибутів є найбільш значущими для передбачення значень цільового атрибуту, судячи з побудованих моделей? Чому? Як зміниться точність передбачення, якщо залишити лише значущі атрибути?

Варіанти завдань:

Варіант 1

№	y	x_1	x_2
1	82	18,9	43
2	70	22,8	54
3	66	23,1	48,8
4	60	22,8	57,2
5	50	27,3	44,2
6	38	32,4	65
7	37	31,5	65
8	24	37	73,6
9	20	39	77
10	19	39	83
11	17,5	42,7	62,2
12	16	42	64

Варіант 2

№	y	x_1	x_2
1	182	18,9	40
2	162	24,8	43
3	140	29	60
4	125	37	64
5	110	46,8	69,2
6	98	44,4	74,8
7	78	37,8	91
8	66	49	80
9	46	48,8	93,6
10	34	58	93,4
11	14	58,8	107
12	12	58	110

Варіант 3

№	y	x_1	x_2
1	24	25	12,5
2	26,8	31	19,2
3	28,2	34	21,2
4	29,6	57	37,4
5	31	40	48
6	32,4	43	55,6
7	33,8	46	56,8
8	35,2	69	73,8
9	36,6	52	83,6
10	35,9	55	92
11	37	60	80
12	38	64	85

Варіант 4

№	y	x_1	x_2
1	11	30	12,5
2	16	32	20
3	19	36	19,5
4	30	41	40
5	33	42	44,5
6	40	55	57
7	47	48	58
8	65	52	80
9	66	54	79
10	70	55	80
11	75	42,7	81
12	80	42	85

Варіант 5

№	y	x_1	x_2
1	24	25	12,5
2	26,8	31	19,2
3	34	9	25,5
4	33,2	7,6	17,1
5	40,4	10,8	29,8
6	39,6	9,2	37,3
7	46,8	12,6	54,6
8	47	10,4	55,5
9	53,2	14,4	57,4
10	52,4	12	55,7
11	59,6	16,2	68,2
12	58,8	13,2	79,9

Варіант 6

№	y	x_1	x_2
1	21	30	12,5
2	24	35	25,5
3	26,8	51	17,2
4	28,2	44	29,7
5	29,6	54,5	37,4
6	31	50	54,5
7	32,4	45	55,6
8	33,8	56	57,3
9	35,2	61	73,8
10	36,6	62	68,1
11	35,9	60	80
12	38	64	85

Варіант 7

№	y	x_1	x_2
1	8	3,6	25,5
2	8,8	36,2	17,2
3	9,2	4,5	29,7
4	9,6	1,3	37,4
5	10	5,1	54,5
6	10,4	45	55,6
7	10,8	5,7	57,3
8	9,9	6	80
9	10	6,3	68,1
10	11,2	61	73,8
11	11,2	61	73,8
12	11,6	6,3	68,1

Варіант 8

№	y	x_1	x_2
1	34	21,4	25,5
2	36,4	20,6	17,2
3	46,8	29,8	29,6
4	49,2	35	37,6
5	59,6	38,2	54,2
6	63	32,6	56
7	72,4	46,6	56,8
8	74,8	50,2	56,4
9	85,2	55	67,4
10	87,6	47,6	80,8
11	90	61	73,8
12	92	63	68,1

Варіант 9

№	y	x_1	x_2
1	145	21,4	25,5
2	163	20,6	17,2
3	202	31,4	19
4	212	32,6	9,2
5	282	43,8	14,1
6	330	51	21,6
7	344	56,2	37,7
8	399	52,6	32
9	442	68,6	25,3
10	478	74,2	24,4
11	524	81	34,9
12	525	75,6	40,8

Варіант 10

№	y	x_1	x_2
1	202	31,2	106
2	218	29,3	102,2
3	242	37,4	98,6
4	258	42,5	105,2
5	282	43,6	112
6	299	38,3	109
7	322	49,8	106,2
8	338	54,1	113,6
9	362	56,1	121,2
10	378	49,5	119
11	380	100,7	90
12	400	102	95

Варіант 11

№	y	x_1	x_2
1	95	44	80
2	101	47	100
3	109	58,3	97,8
4	121	56,4	93,4
5	129	68,9	94,8
6	141	65,8	93
7	150	79,5	91
8	161	75,2	84,8
9	169	90,1	86,4
10	181	94,6	85,8
11	189	100,7	90
12	190	102	91

Варіант 12

№	y	x_1	x_2
1	150	15	106
2	182	17	89,8
3	221	25,3	87,8
4	218	20,4	83,6
5	250	29,9	84,8
6	254	23,8	82,8
7	270	34,5	81
8	290	27,2	75
9	305	39,1	76,4
10	326	40,6	75,6
11	341	43,7	80
12	400	102	95

Варіант 13

№	y	x_1	x_2
1	126	21	93
2	182	54	89,8
3	162	51,2	104
4	174	46,3	97,8
5	194	61,4	91,4
6	206	73,5	94,8
7	226	71,6	98
8	239	63,3	91
9	258	81,8	83,8
10	270	93,1	86,4
11	290	92,1	88,8
12	302	82,5	85

Варіант 14

№	y	x_1	x_2
1	24	17	69,4
2	26	25,3	67,8
3	29	20,4	64
4	31	29,9	64,8
5	34	23,8	62,4
6	36	34,5	61
7	39	27,2	55,4
8	41	39,1	56,4
9	43	40,6	55,2
10	45	43,7	60
11	51	69	88,8
12	58	82,5	85

Варіант 15

№	y	x_1	x_2
1	49	17	31
2	48,5	16	37
3	48	16	46
4	47,6	15	52
5	47,3	12	71
6	46,9	11	91
7	46,5	11,5	92
8	46,3	11	145
9	46,1	10	190
10	45,9	10	204
11	45,7	9,5	222
12	45,6	8	271

Варіант 16

№	y	x_1	x_2
1	50	43	15
2	44	41	14,8
3	47	43,8	10,6
4	41	36,4	6,6
5	45	43,4	14,8
6	36	31,8	23,2
7	38	38	19,8
8	31	27,2	16,6
9	36	36,6	25,6
10	28	28	34,8
11	26	30,2	37
12	28	30	35

Варіант 17

№	y	x_1	x_2
1	249	17	91,9
2	237	15,3	92,2
3	219	20,4	92,7
4	207	39,9	95,2
5	189	23,8	97,9
6	177	24,5	99
7	159	27,2	100,3
8	147	49,1	103,6
9	129	30,6	107,1
10	117	33,7	109
11	108	41	109
12	100	30	112

Варіант 18

№	y	x_1	x_2
1	47	44	5,6
2	50	43,3	5
3	41	36,4	6,6
4	44	43,9	8,4
5	35	31,8	10,4
6	38	36,5	12,6
7	29	27,2	13,8
8	32	37,1	19,6
9	23	22,6	22,4
10	26	29,7	23,4
11	27	33	26
12	30	30	32

Варіант 19

№	y	x_1	x_2
1	27	44	8,2
2	36	43,8	7,8
3	33	36,4	7,6
4	42	43,4	11,6
5	39	31,8	15,8
6	48	37	16,2
7	45	27,2	16,8
8	54	36,6	21,6
9	51	22,6	26,6
10	60	30,2	27,8
11	62	33	29
12	65	34	32

Варіант 20

№	y	x_1	x_2
1	27	16,8	164,1
2	35	15,5	197
3	30	15,2	187,5
4	40	14,9	218,6
5	35	11,6	230,1
6	46	12,3	240,6
7	39	12	241
8	52	11,7	272,8
9	48	10,4	278
10	57	10	305
11	55	9	300
12	51	9	304

Варіант 21

№	y	x_1	x_2
1	270	16,8	152
2	350	15,5	197
3	414	2,1	520
4	468	1,9	450
5	498	1,8	610
6	552	1,9	720
7	582	1,4	750
8	636	1,5	770
9	666	1,4	850
10	720	1,6	830
11	750	1,3	990
12	804	1,1	980

Варіант 22

№	y	x_1	x_2
1	78	43	57,4
2	90	61,1	63,2
3	94	59,8	73,8
4	106	70,7	72,8
5	110	51	76
6	122	87,7	80,4
7	126	52,6	90,6
8	138	107,5	88
9	142	72,6	92,6
10	154	85	95,6
11	159	89	99
12	204	145	98

Варіант 23

№	y	x_1	x_2
1	7,2	31	15
2	7,9	45	16
3	9,2	61	19
4	9,4	72	18,2
5	11	75	17
6	11,3	77	18
7	11,5	85	13,9
8	12,3	83	15,4
9	13	99	11,7
10	13,8	98	13,2
11	14	95	15
12	17	98	22

Варіант 24

№	y	x_1	x_2
1	50	17	93
2	47	15,3	105
3	41	20,4	91,7
4	34	29,9	91,6
5	33	23,8	91,4
6	32	24,5	90
7	30	27,2	86
8	26	39,1	91
9	25	30,6	83,8
10	24	33,7	86,4
11	21	41	86,2
12	20	30	74

Варіант 25

№	y	x_1	x_2
1	59	43	28,2
2	67	61	29,6
3	71	59,8	35,4
4	79	70,8	34,4
5	83	51	38
6	91	87,6	37,2
7	95	52,6	41,8
8	103	107,6	40
9	107	72,6	45,8
10	115	112,6	42,8
11	141	95	52
12	172	98	54,1

Варіант 26

№	y	x_1	x_2
1	59	61	12
2	67	57,8	9,2
3	77	66	7,8
4	86	77,2	8,8
5	89	78	10
6	91	75	8,4
7	100	85	6,6
8	108	92,5	8
9	114	102	10
10	115	97	7
11	116	95	6
12	117	98	5

Варіант 27

№	y	x_1	x_2
1	59	12,6	9,4
2	67	12,9	9,4
3	71	15,4	9,6
4	79	18,1	10
5	83	18,2	10,6
6	91	18,1	11,4
7	95	21	12,4
8	103	24,1	13,6
9	107	23,8	15
10	115	23,3	16,6
11	116	25	18
12	59	12,6	9,4

Варіант 28

№	y	x_1	x_2
1	57	18,9	42
2	50	16,6	50,2
3	42	18,1	59,6
4	39	21,4	57,2
5	37	27,3	54
6	30	29,6	65
7	19	31,5	77,2
8	17	38,4	73,6
9	16	35,7	69,2
10	15	30,2	83
11	9	25	99
12	8	33,6	88

Варіант 29

№	y	x_1	x_2
1	62	18,9	43
2	56	22,8	50,2
3	56	23,1	48,8
4	54	22,8	57,2
5	54	27,3	44,2
6	42	32,4	65
7	34	41	65
8	29,8	46,8	73,6
9	28	48,7	62,2
10	21	52	83
11	20	51	89
12	15	64	102

Варіант 30

№	y	x_1	x_2
1	101	18,9	43
2	89	18,4	50,2
3	72	23,1	58,4
4	69	28	57,2
5	68	27,3	55,4
6	59	26,4	65
7	44	31,5	75,6
8	29	43,8	73,6
9	28	42,7	71
10	19	44,4	83
11	9	48	86,2
12	8	52	78

5. КОНТРОЛЬНІ ЗАПИТАННЯ

1. У чому полягає задача регресії? Наведіть практичний приклад?
2. Чим задача регресії схожа і чим відрізняється від задачі класифікації?
3. Що таке навчання з учителем і без учителя? До якого типу належить завдання регресії?
4. Задача регресії є описовою або прогнозуючою і чому?
5. Опишіть один з розглянутих методів, що вирішують завдання регресії.
6. Як оцінити якість побудованої моделі для завдання регресії?
7. Метод найменших квадратів.
8. Закон нормального розподілу.

6. ЗМІСТ ЗВІТУ

1. Тема і мета роботи.
2. Завдання до роботи.
3. Результати виконання завдань.
4. Відповіді на контрольні запитання.
5. Висновки, що відображують результати виконання роботи та їх критичний аналіз.