Linear Regression

Prof. Surya Prakash

IIT Indore

Linear regression

Domain	Application	Example
Business / Economics	Predicting sales or profit	Estimate sales based on advertising spend
Real Estate	Price prediction	Predict house price using area, location, and rooms
Education	Performance prediction	Predict student marks from study hours
Finance	Risk and return analysis	Predict stock returns based on market indicators
Healthcare	Medical cost estimation	Predict hospital charges based on patient age and condition

Example 1 - Advertisement vs. Sales dataset

Advertisement Spend (XX, ₹)	Sales (YY, ₹)
1000	2000
2000	4000
3000	6000
4000	8000
5000	10000
6000	12000
7000	
8000	
9000	
10000	

$$Y = 2X$$

- X = Advertisement spend (₹)
- Y = Sales revenue (₹)

Example 2 - Travel time vs. Distance dataset

Time (X, hours)	Distance (Y, km)
1	2
2	4
3	6
4	8
5	10
6	
8	
12	

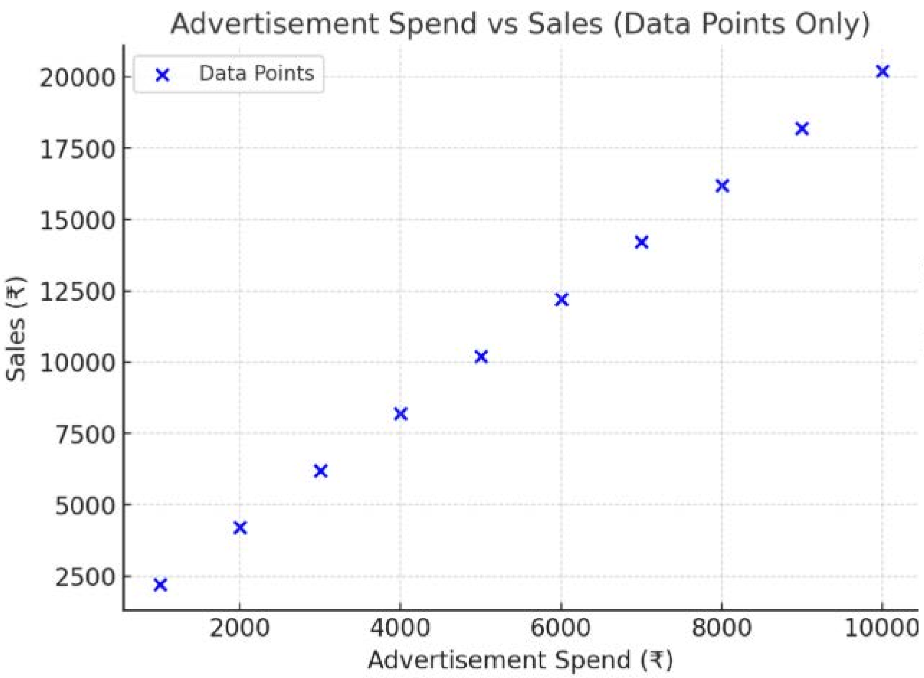
- X = Time (in hours)
- Y = Distance covered (in km)

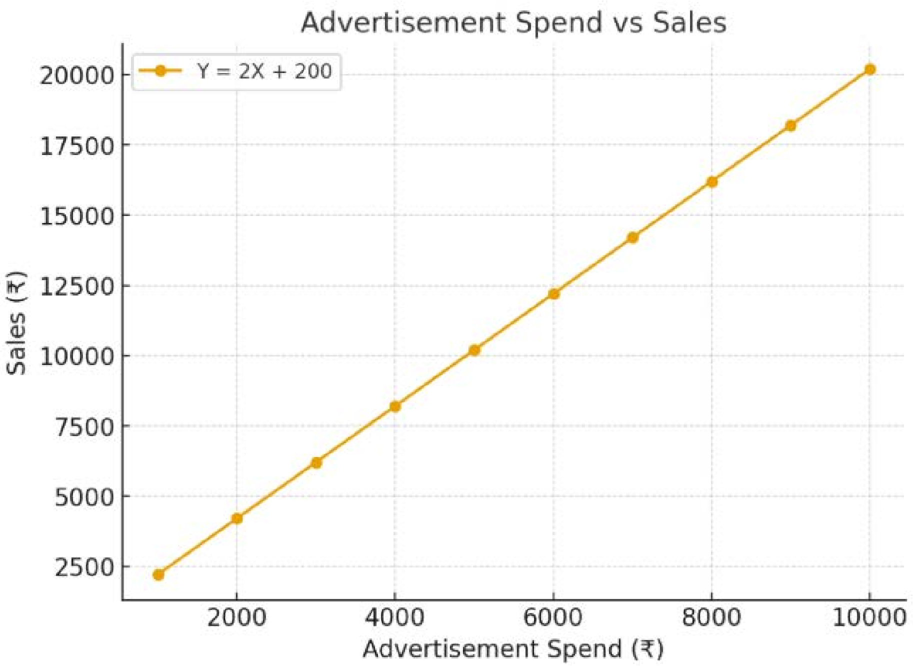
The relationship is:

$$Y = 2X$$

Example 3 - Advertisement vs. Sales dataset

Advertisement Spend (X, ₹)	Sales (Y, ₹)
1000	2200
2000	4200
3000	6200
4000	8200
5000	10200
6000	12200
7000	14200
8000	16200
9000	18200
10000	20200





Let's use the **two-point form of a line equation**:

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

$$(x_1, y_1) = (2000, 4200),$$

$$(x_2, y_2) = (5000, 10200)$$

$$y - 4200 = \frac{10200 - 4200}{5000 - 2000}(x - 2000)$$

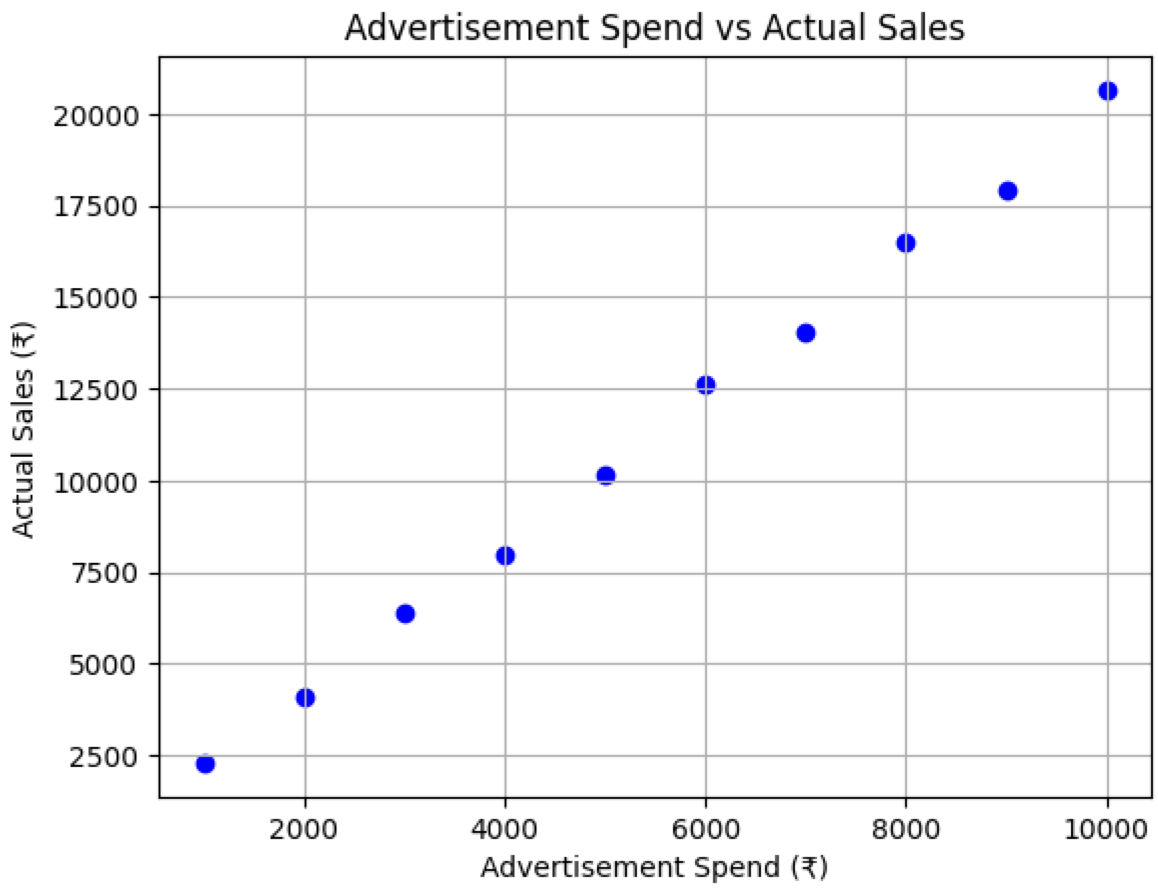
$$y - 4200 = \frac{6000}{3000}(x - 2000)$$

$$y - 4200 = 2(x - 2000)$$

$$y = 2x + 200$$

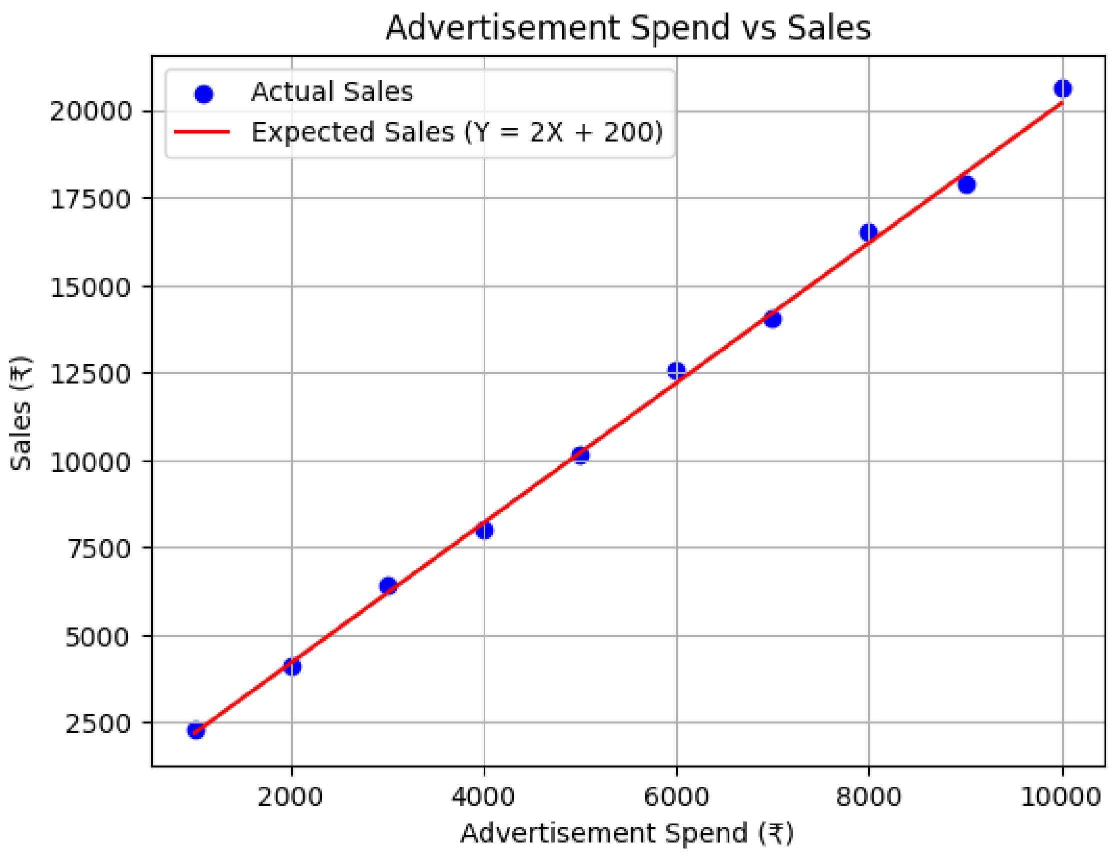
Example 4 - Advertisement vs. Sales dataset

Advertisement Spend (X, ₹)	Actual Sales (Y, ₹)
1000	2300
2000	4100
3000	6400
4000	8000
5000	10150
6000	12600
7000	14050
8000	16500
9000	17900
10000	20650
11000	
120000	



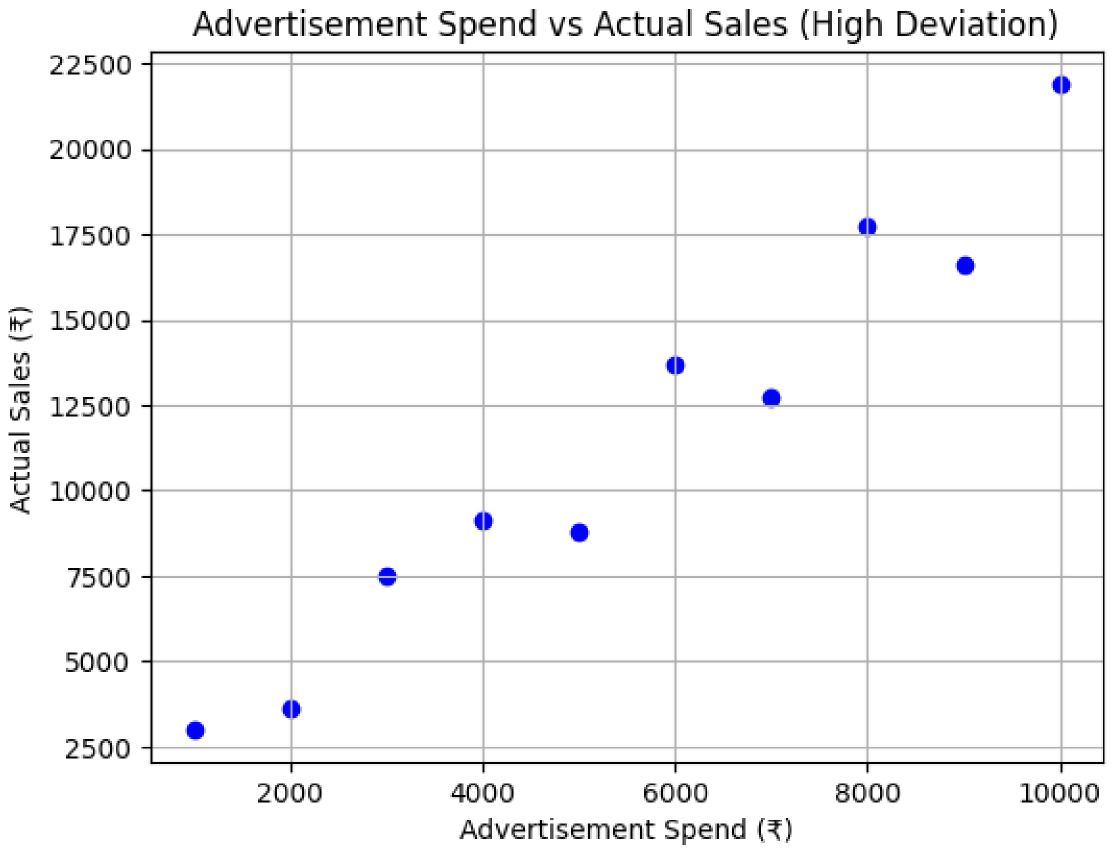
Example 4 - Advertisement vs. Sales dataset

Advertisement Spend (X, ₹)	Actual Sales (Y, ₹)
1000	2300
2000	4100
3000	6400
4000	8000
5000	10150
6000	12600
7000	14050
8000	16500
9000	17900
10000	20650



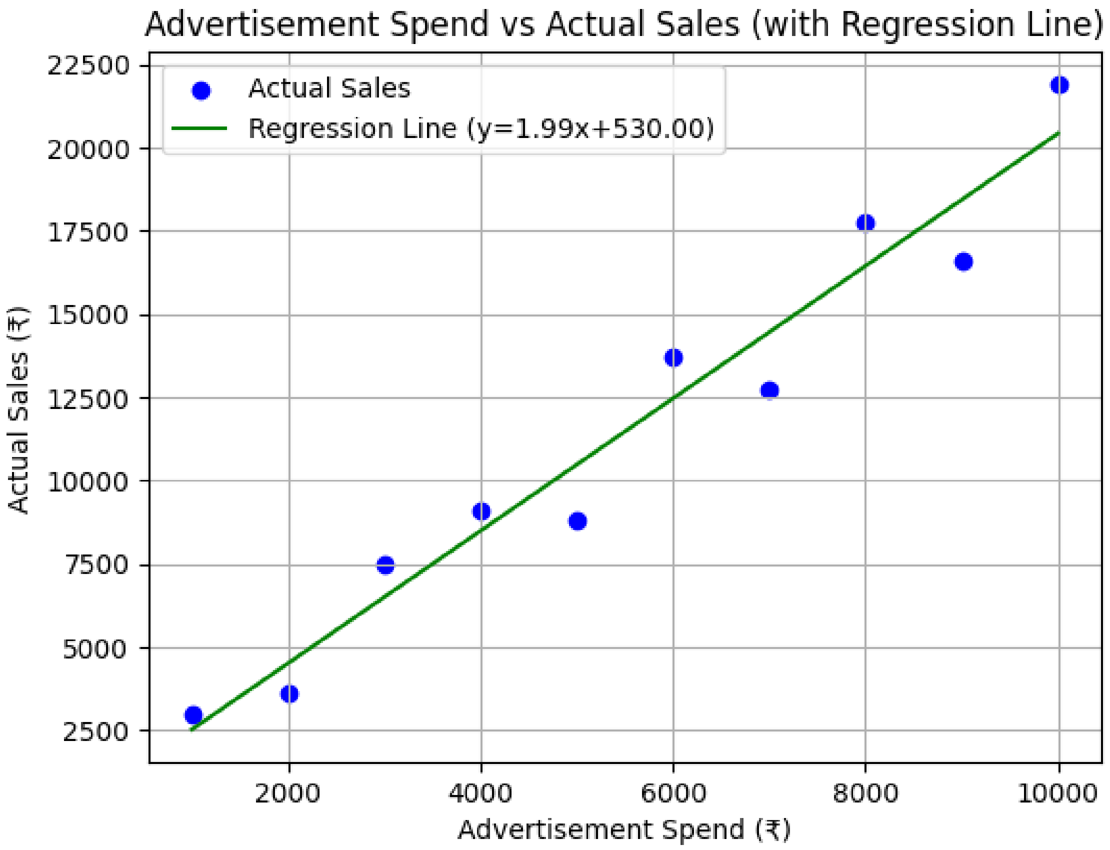
Example 5 - Advertisement vs. Sales dataset

Advertisement Spend (X, ₹)	Actual Sales (Y, ₹)
1000	3000
2000	3600
3000	7500
4000	9100
5000	8800
6000	13700
7000	12700
8000	17750
9000	16600
10000	21900
11000	
12000	



Example 5 - Advertisement vs. Sales dataset

Advertisement Spend (X, ₹)	Actual Sales (Y, ₹)
1000	3000
2000	3600
3000	7500
4000	9100
5000	8800
6000	13700
7000	12700
8000	17750
9000	16600
10000	21900
11000	
12000	

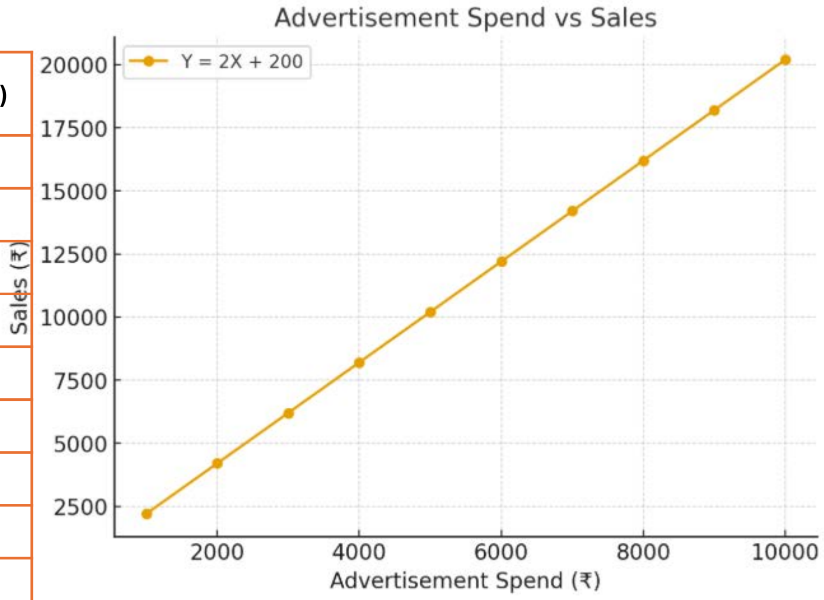


Linear Regression

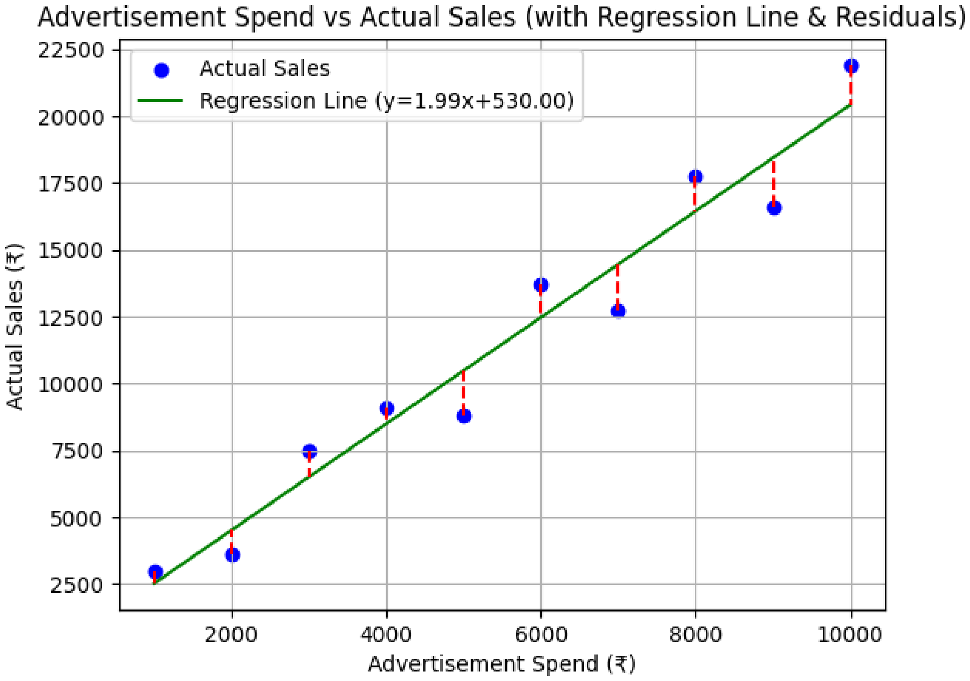
- When your data points don’t lie exactly on a straight line (because of noise, measurement errors, or natural variability), linear regression finds the **best-fit line** that minimizes the error.

Consider these two cases

Advertisement Spend (X, ₹)	Sales (Y, ₹)
1000	2200
2000	4200
3000	6200
4000	8200
5000	10200
6000	12200
7000	14200
8000	16200
9000	18200
10000	20200



Advertisement Spend (X, ₹)	Actual Sales (Y, ₹)
1000	3000
2000	3600
3000	7500
4000	9100
5000	8800
6000	13700
7000	12700
8000	17750
9000	16600
10000	21900



$$y = 2x + 200$$

For $X = 4000$:

$$Y = 2(4000) + 200 = 8000 + 200 = 8200$$

For $X = 6000$:

$$Y = 2(6000) + 200 = 12000 + 200 = 12200$$

$$Y = 1.99X + 530$$

For $X = 4000$:

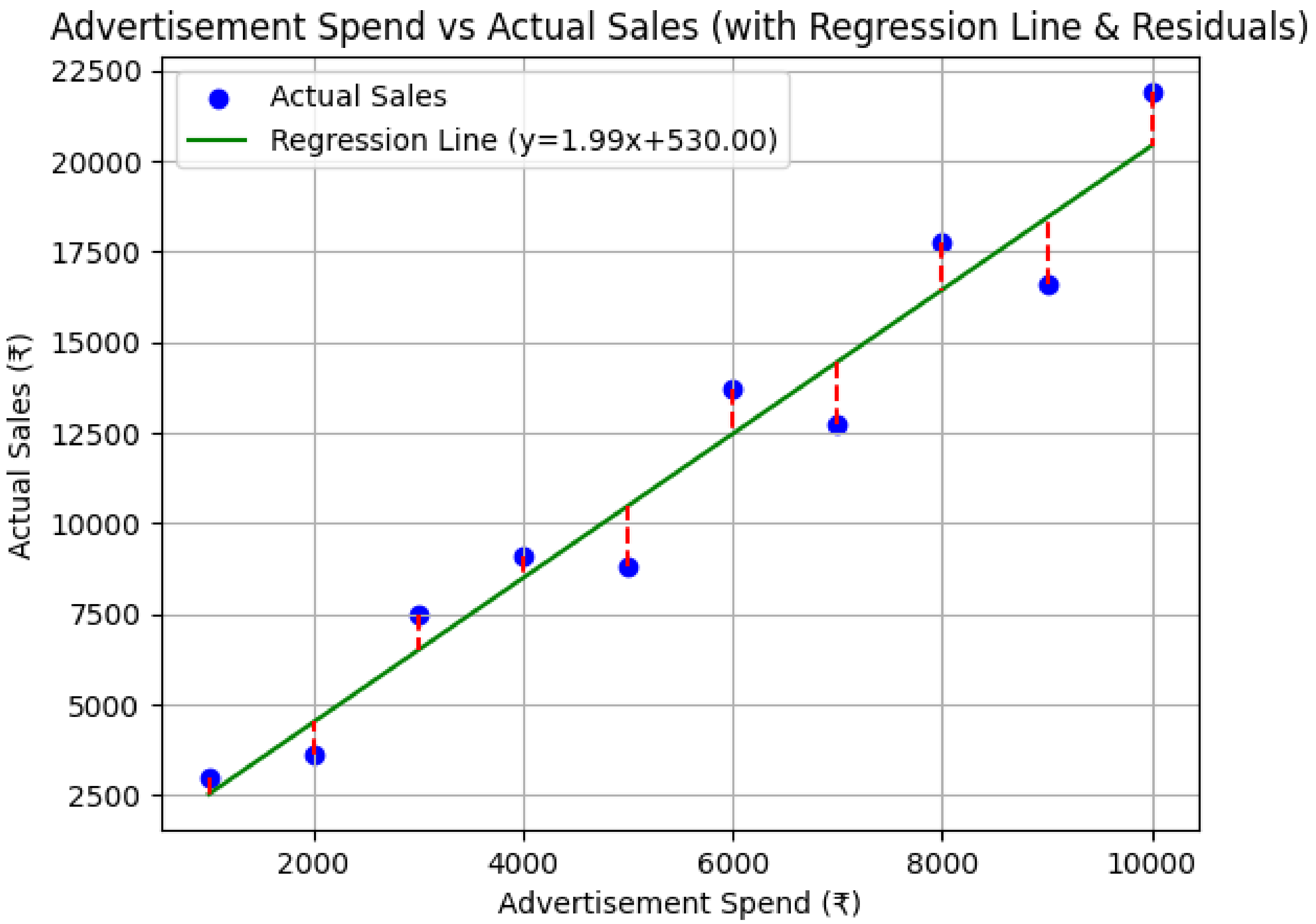
$$Y = 1.99(4000) + 530 = 7960 + 530 = 8490$$

For $X = 6000$:

$$Y = 1.99(6000) + 530 = 11940 + 530 = 12470$$

Least Square Fitting

Best Fit Line

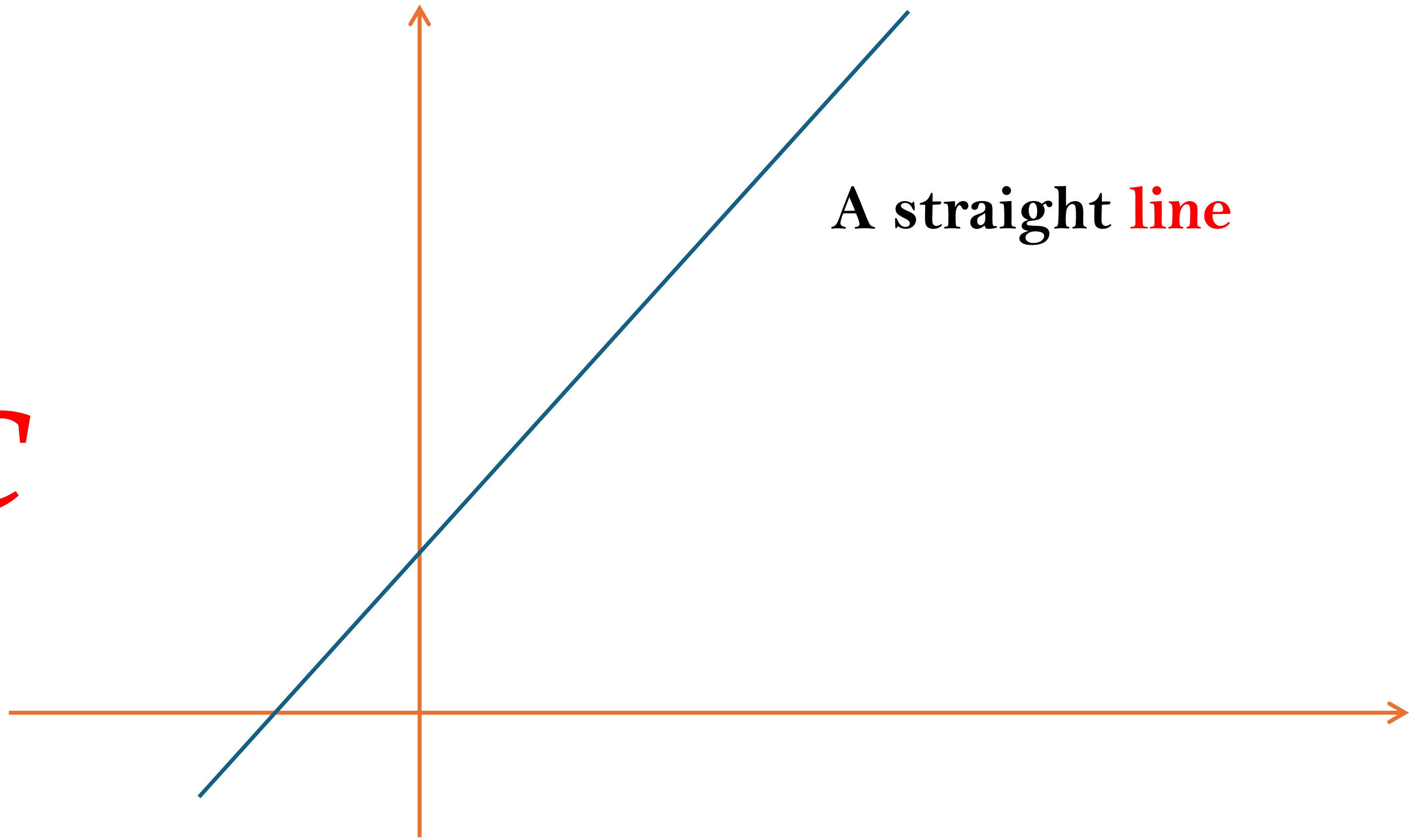


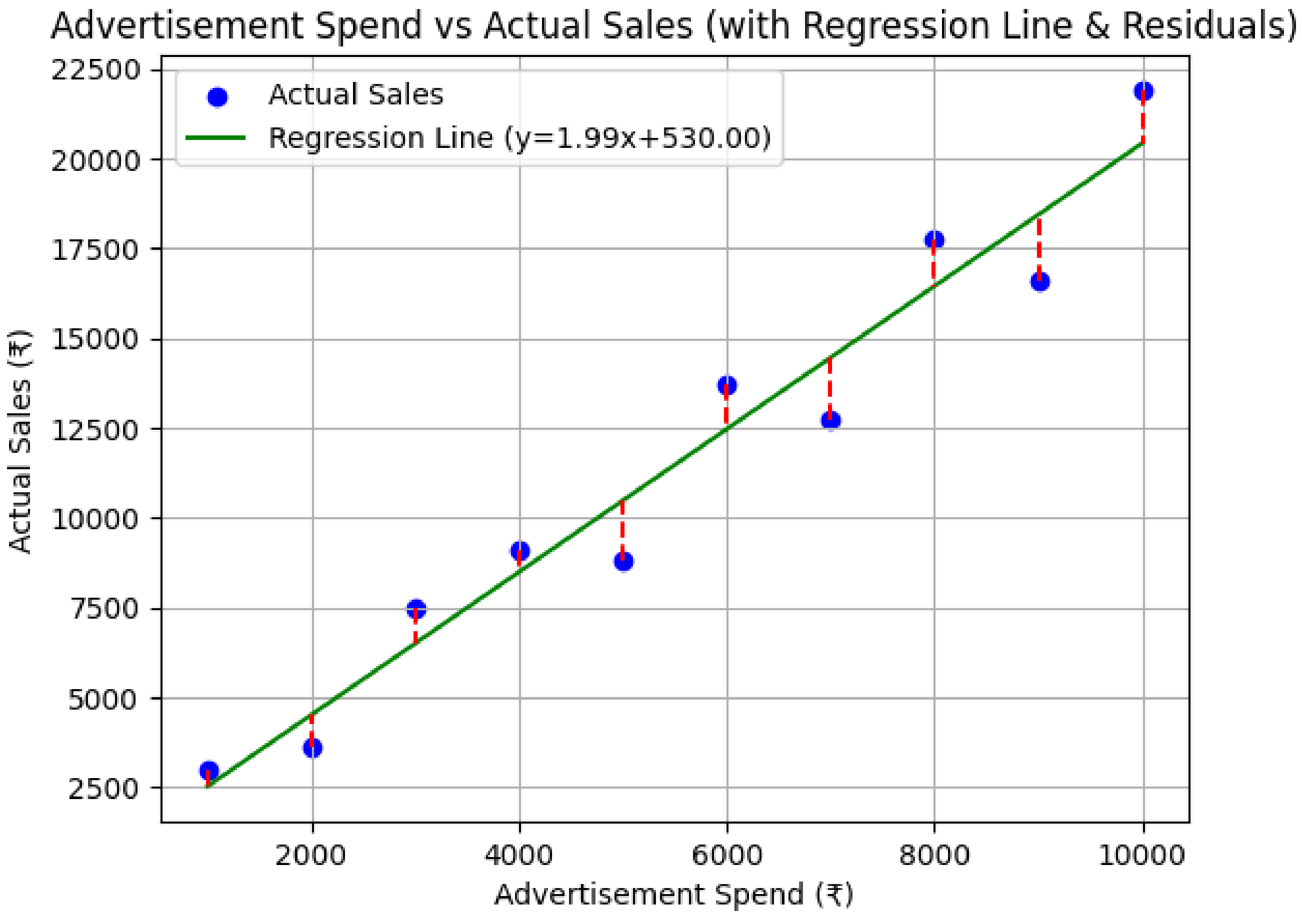
Equation of a Line

$$y = {\color{red}m}\,x + {\color{red}c}$$

slop

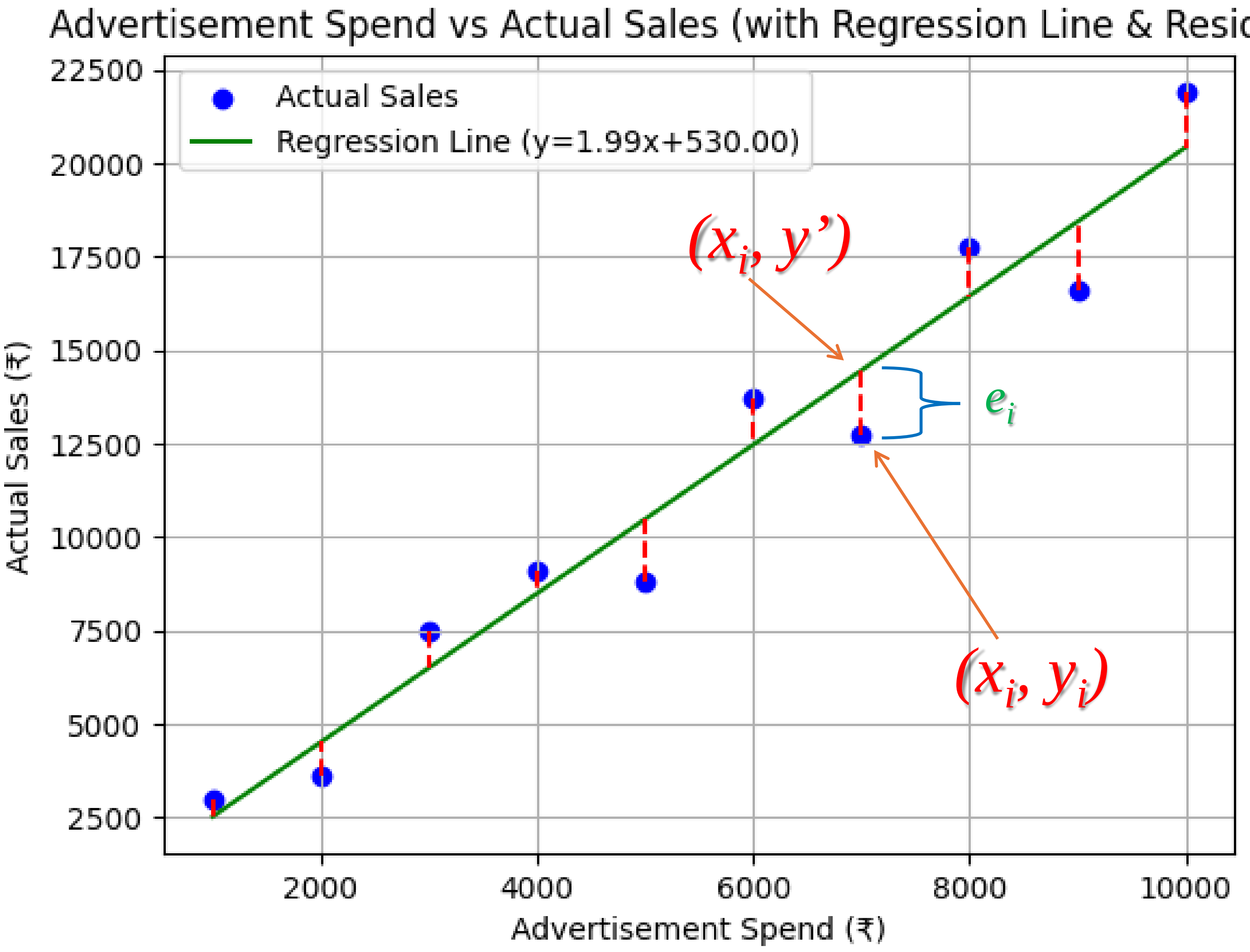
y-intercept





Error Function (Mean Squared Error):

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}(mx_i + c - y_i)^2$$



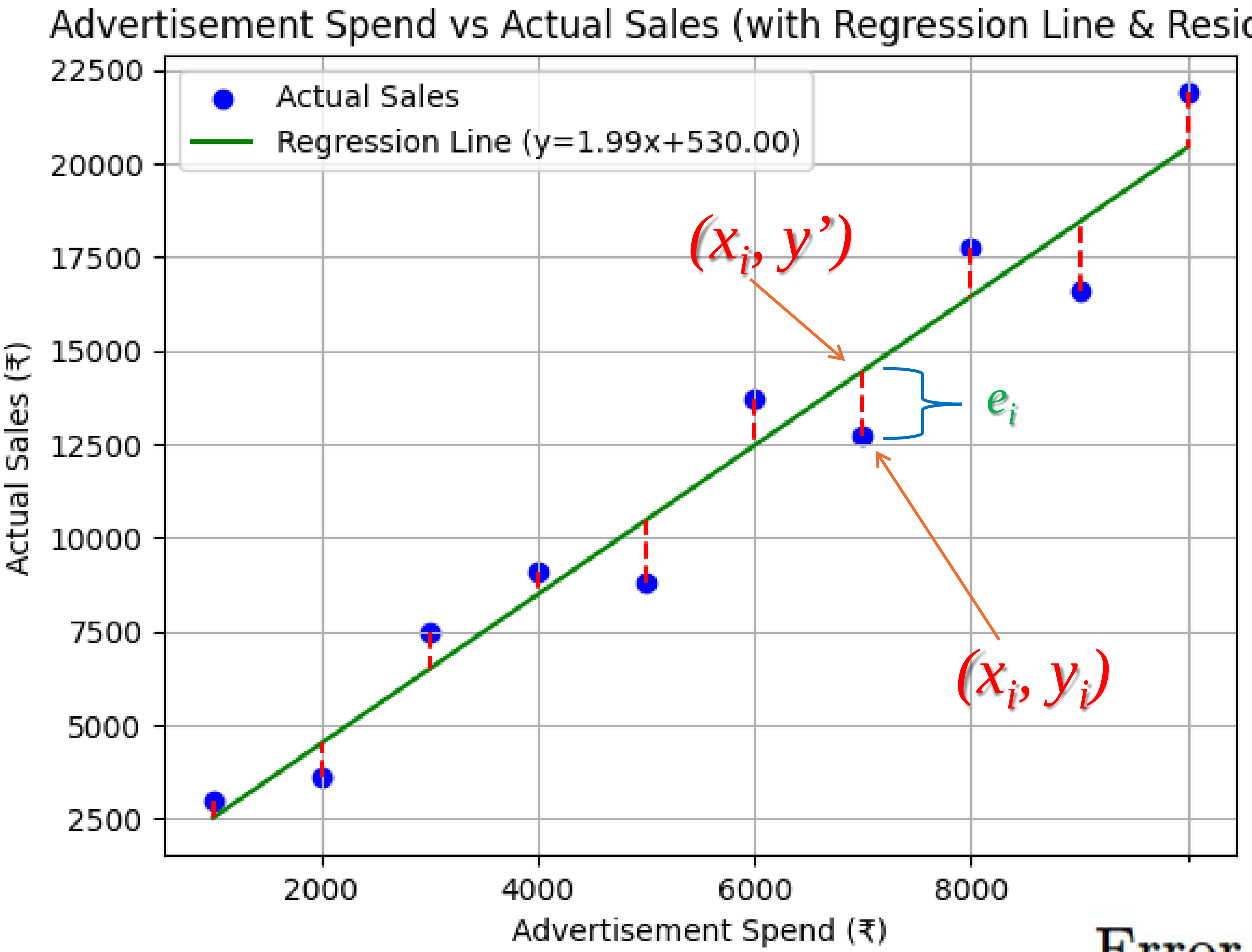
Equation of the line

$y = m\,x + c$

Vertical distance of the data point from the fitted line

$$e_i = y' - y_i$$

$$e_i = \underline{m\,x_i + c} - y_i$$



Equation of the line

$y = m\,x + c$

Vertical distance of the data point from the fitted line

$$e_i = y' - y_i$$

$$e_i = m\,x_i + c - y_i$$

Mean Squared Error

$$\mathrm{Error} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

$$\mathrm{Error} = \frac{1}{n}\sum_{i=1}^{n}(mx_i + c - y_i)^2$$

Closed form solution for 2-D case

Error Function (Mean Squared Error):

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}(mx_i + c - y_i)^2$$

Set partial derivatives to zero

$$\frac{\partial E}{\partial m} = \frac{2}{n}\sum_{i=1}^{n} x_i(mx_i + c - y_i) = 0, \qquad \frac{\partial E}{\partial c} = \frac{2}{n}\sum_{i=1}^{n}(mx_i + c - y_i) = 0$$

This gives the normal equations:

$$m\sum x_i^2 + c\sum x_i = \sum x_i y_i$$

$$m\sum x_i + nc = \sum y_i.$$

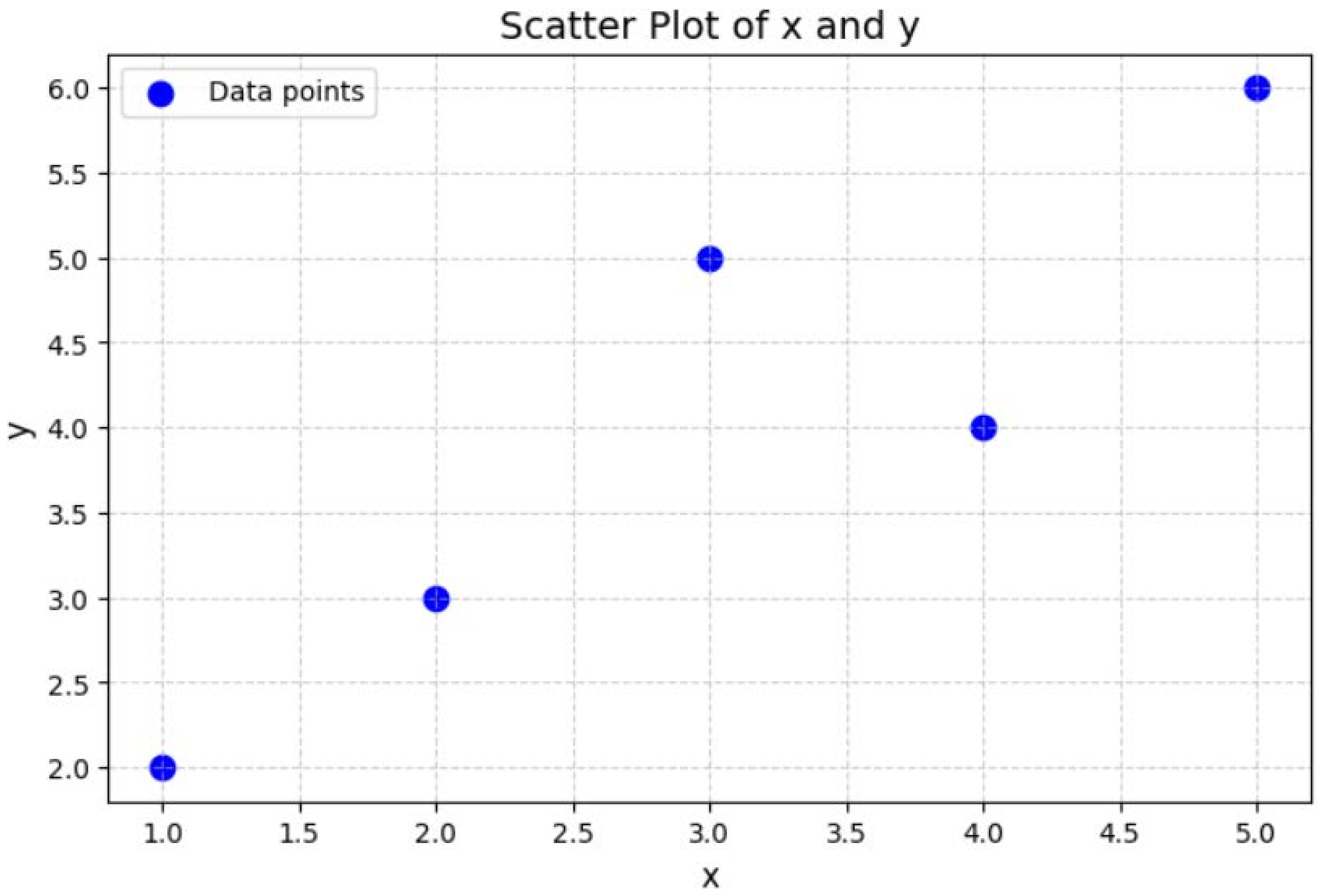
Solve the 2×2 linear system

Slope ***m*** and intercept ***c*** that minimize the Mean Squared Error

$$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad c = \bar{y} - m\bar{x}$$

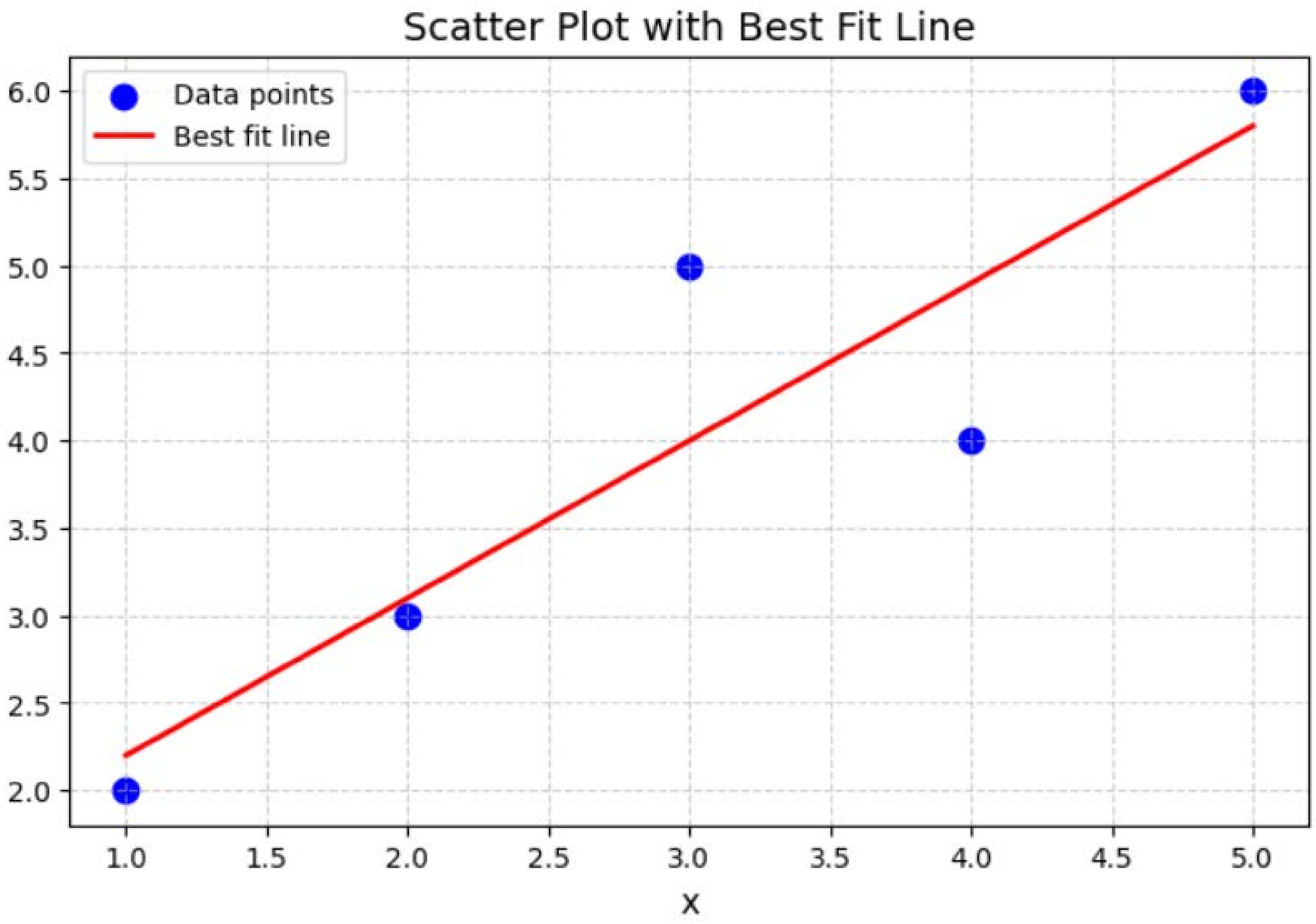
Example

x_i	y_i
1	2
2	3
3	5
4	4
5	6



Example

x_i	y_i
1	2
2	3
3	5
4	4
5	6



→From above given data, find the best fit line

→Compute m and c values of the best fit line

Example

i	x_i	y_i	$(x_i)^2$	$x_i y_i$
1	1	2	1	2
2	2	3	4	6
3	3	5	9	15
4	4	4	16	16
5	5	6	25	30
Σ	**15**	**20**	**55**	**69**

Slope m:

$$m = \frac{n \times \sum x_i y_i - (\sum x_i) \times (\sum y_i)}{n \times \sum x_i^2 - (\sum x_i)^2}$$

Intercept c:

$$c = \frac{\sum y_i - m \times \sum x_i}{n}$$

Example

i	x_i	y_i	$(x_i)^2$	$x_i y_i$
1	1	2	1	2
2	2	3	4	6
3	3	5	9	15
4	4	4	16	16
5	5	6	25	30
Σ	**15**	**20**	**55**	**69**

Slope m:

$$m = \frac{n \times \sum x_i y_i - (\sum x_i) \times (\sum y_i)}{n \times \sum x_i^2 - (\sum x_i)^2}$$

Numeric substitution:

$$m = \frac{5 \times 69 - 15 \times 20}{5 \times 55 - 15^2} = \frac{345 - 300}{275 - 225} = \frac{45}{50} = 0.9$$

Intercept c:

$$c = \frac{\sum y_i - m \times \sum x_i}{n}$$

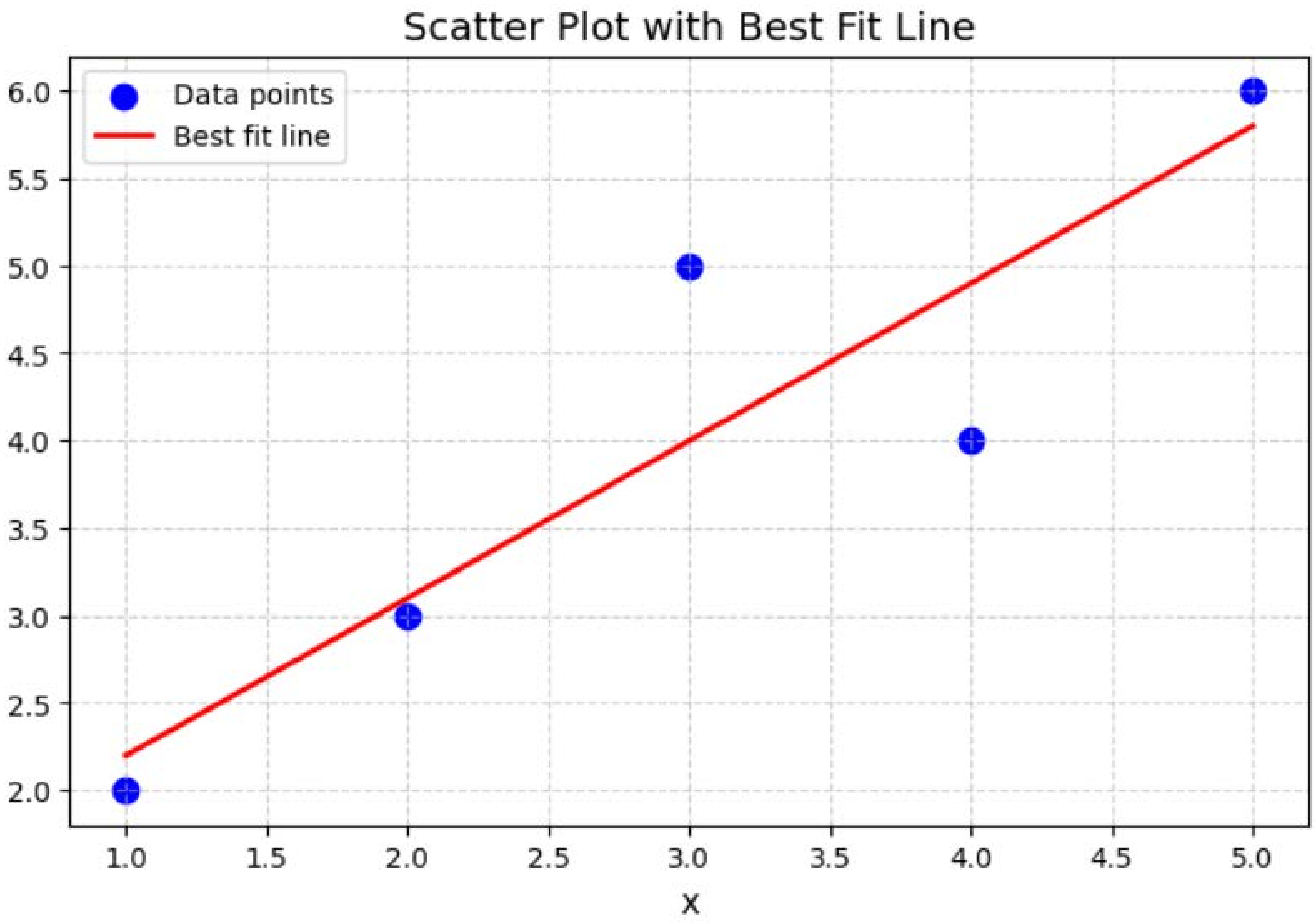
Numeric substitution:

$$c = \frac{20 - 0.9 \times 15}{5} = \frac{6.5}{5} = 1.3$$

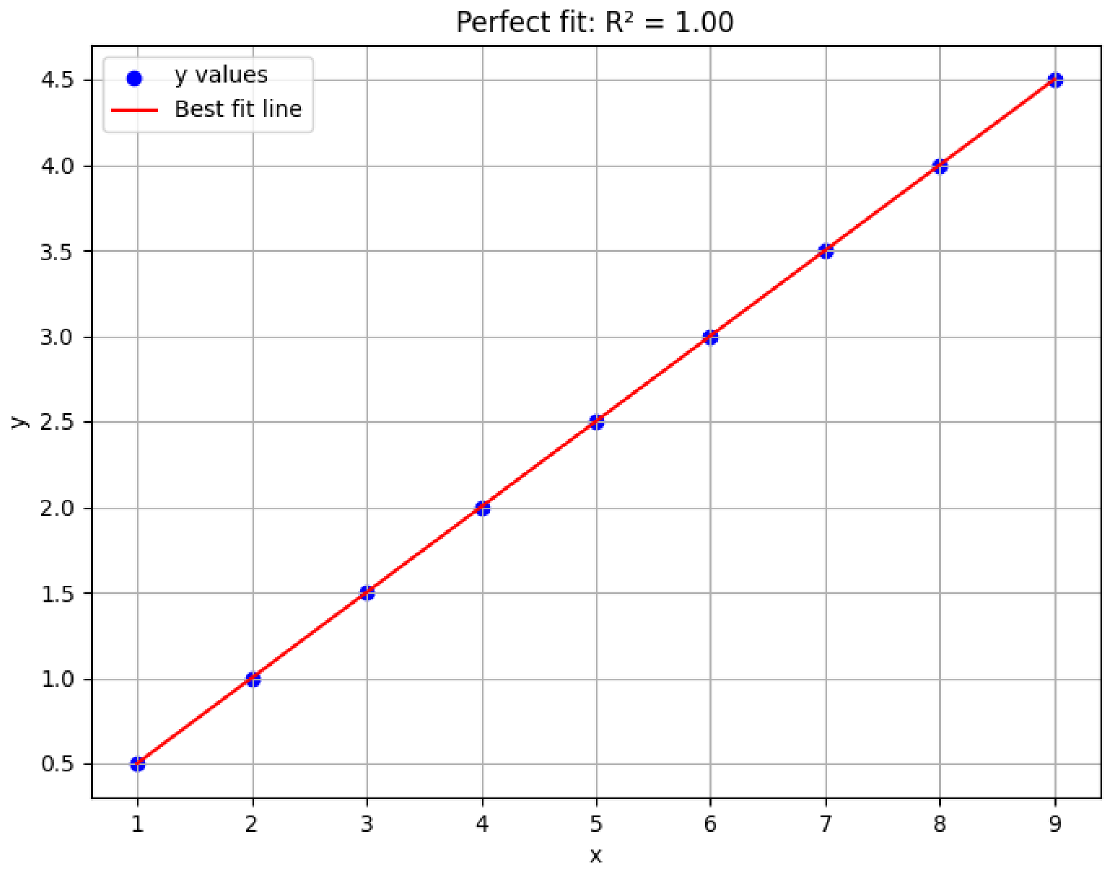
Example

i	x_i	y_i	$(x_i)^2$	$x_i y_i$
1	1	2	1	2
2	2	3	4	6
3	3	5	9	15
4	4	4	16	16
5	5	6	25	30
Σ	**15**	**20**	**55**	**69**

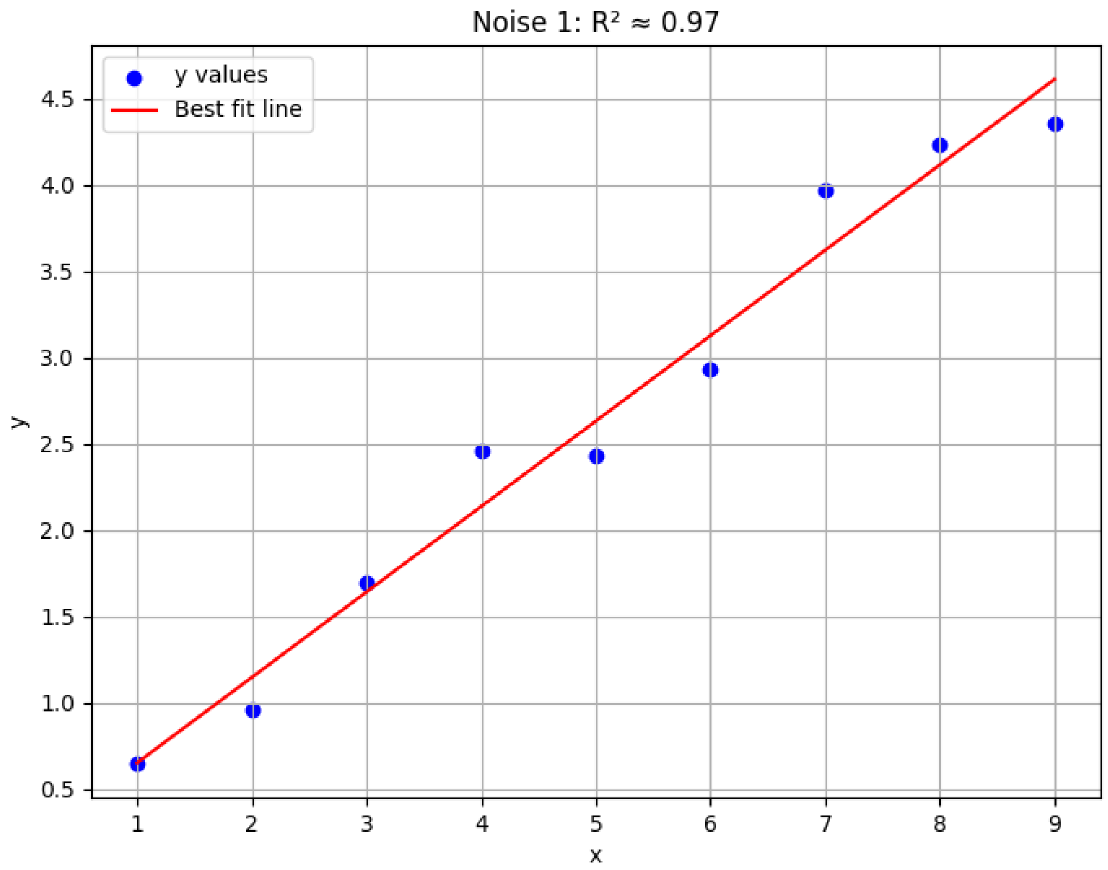
$$\dot{y} = 1.3 + 0.9x$$



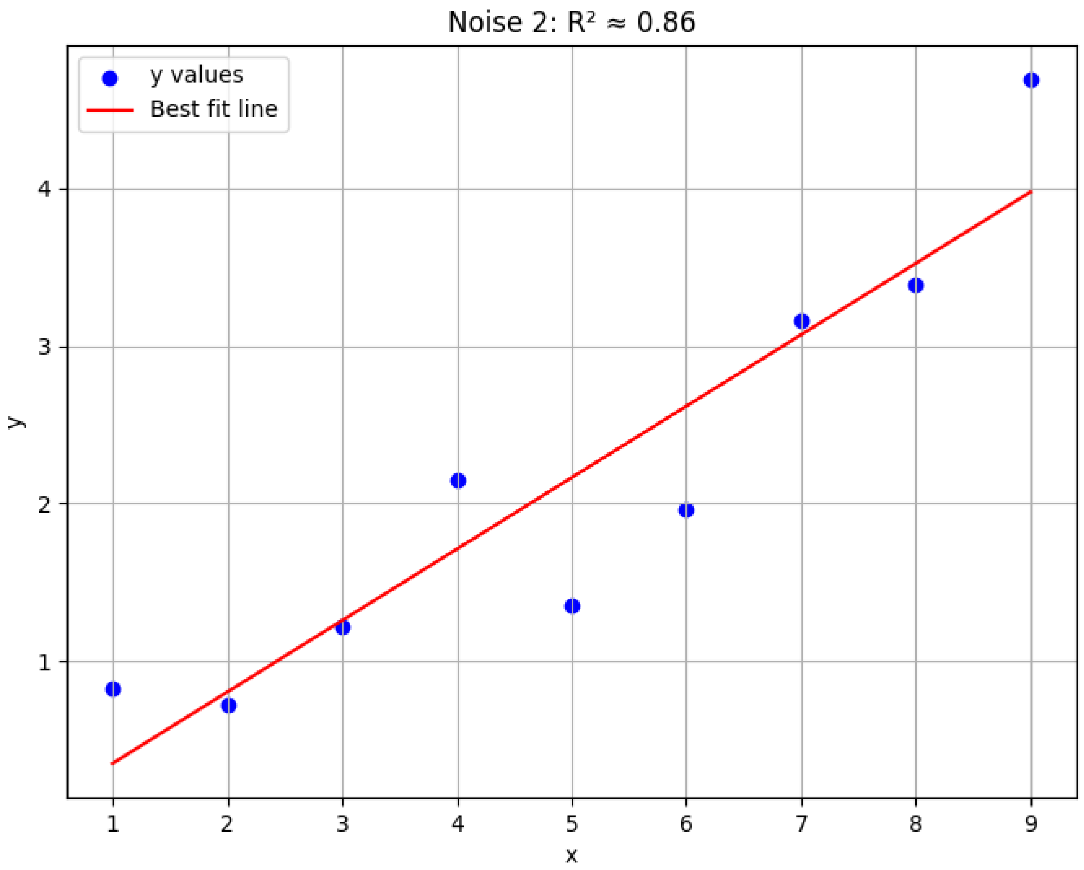
Goodness of fit in linear regression



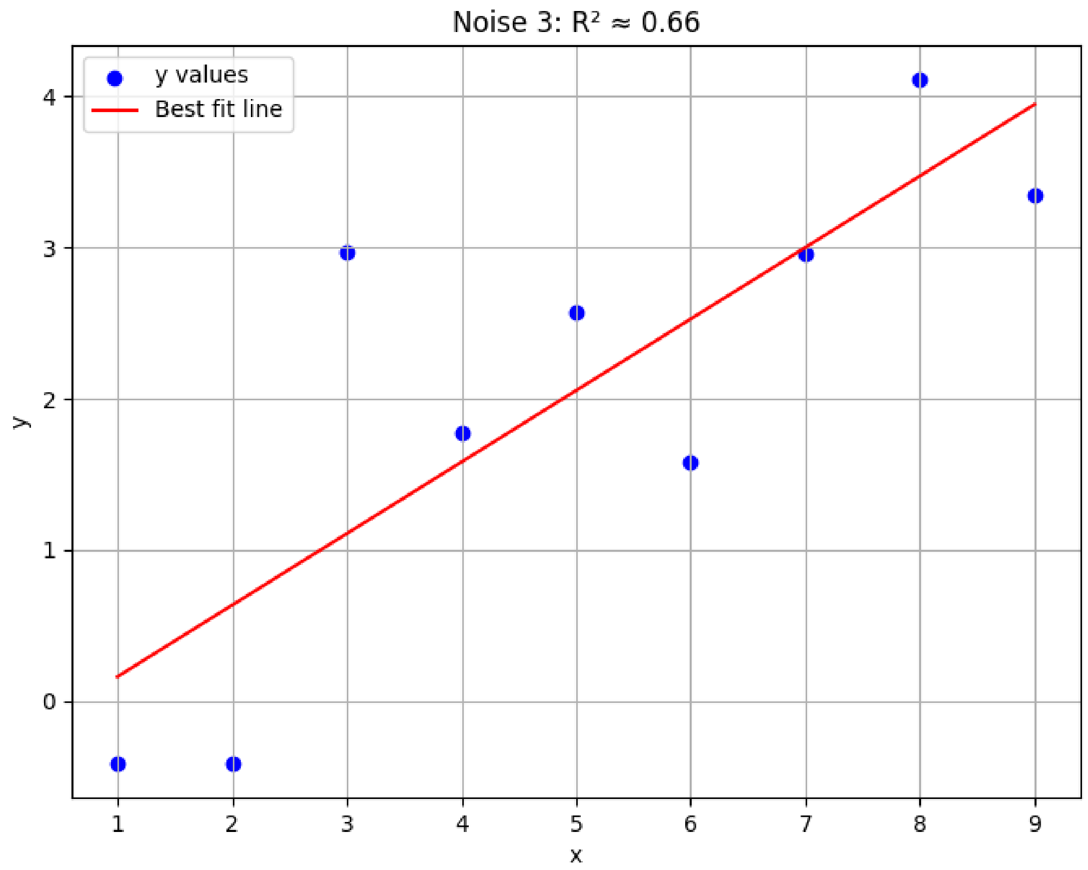
Goodness of fit in linear regression



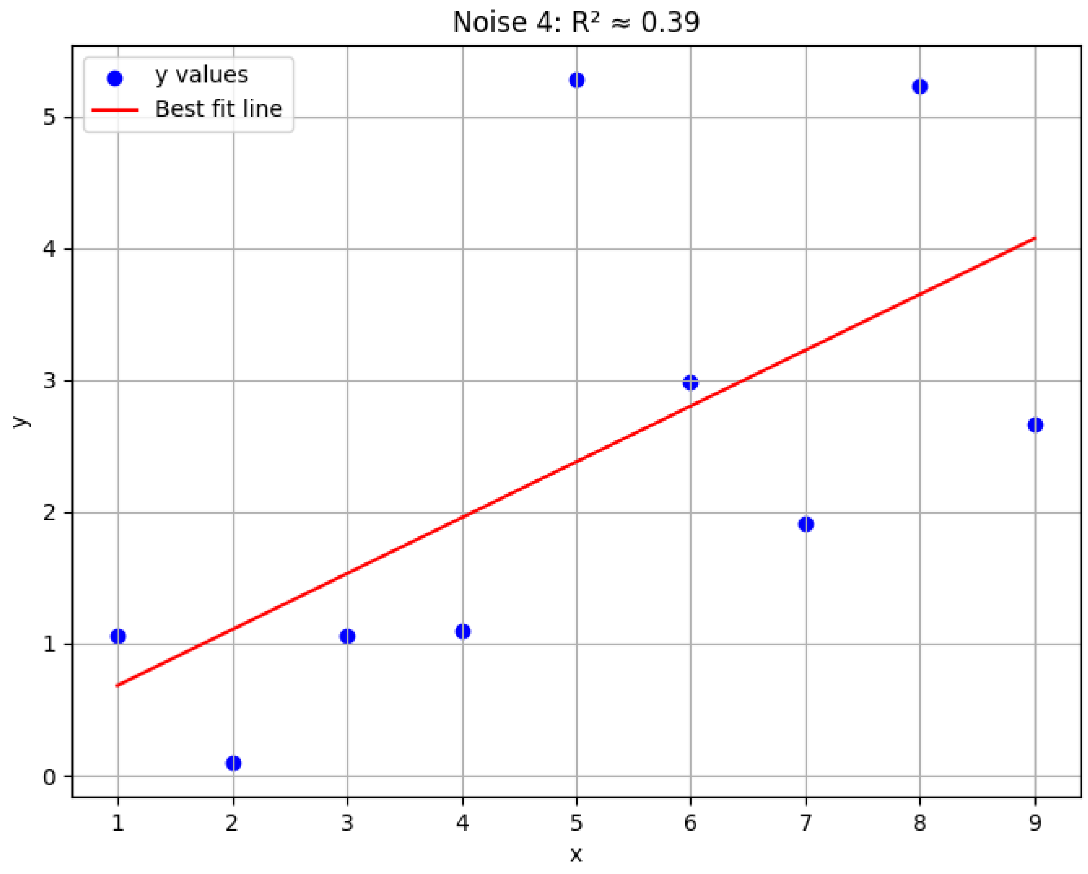
Goodness of fit in linear regression



Goodness of fit in linear regression



Goodness of fit in linear regression



Goodness of fit in linear regression

- In linear regression, **goodness of fit** tells us **how well the regression line explains the variation in the dependent variable**. Essentially:
  - A *good fit* means the predicted values are close to the actual values.
  - A *poor fit* means the regression line doesn't explain the data well.

- Formula

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

y_p = predicted value

y = actual value

$\bar{y}$ = mean of actual y

Goodness of fit in linear regression

- **Good Fit**
  - If the regression predicts the data well, y_p will be close to the actual y.
  - This means the **explained sum of squares** will be **large**, close to the total sum of squares.

$$R^2 \approx 0.8 \text{ to } 1.0$$

- Example:
  - $R^2 = 0.85 \rightarrow$ only 85% of variation in y is explained by the model.

Goodness of fit in linear regression

- **Poor Fit**
  - If the regression predicts poorly, y_p will be far from actual y.
  - The **explained sum of squares** will be **small** compared to total variation.

$$R^2 \approx 0 \; to \; 0.3$$

- Example:
  - $R^2 = 0.15$ →only 15% of variation is explained; most variation is unexplained.

Thanks