Linear Regression

Prof. Surya Prakash

IIT Indore

Example 1 - 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 2 - Advertisement vs. Sales dataset

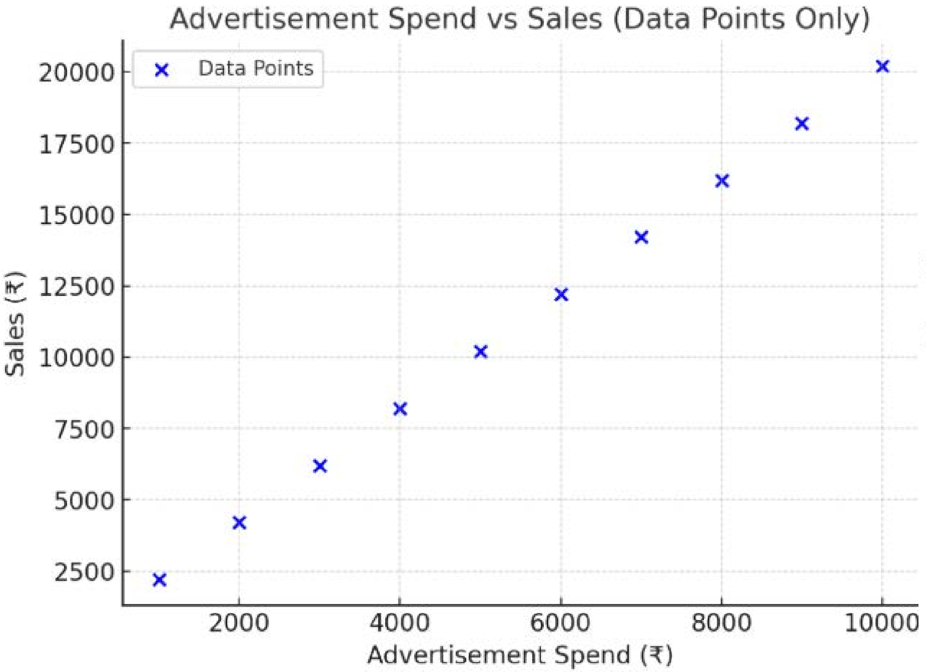
Advertisement Spend (XX, ₹)	Sales (YY, ₹)
1000	2000
2000	4000
3000	6000
4000	8000
5000	10000
6000	12000
7000	
8000	
9000	
10000	

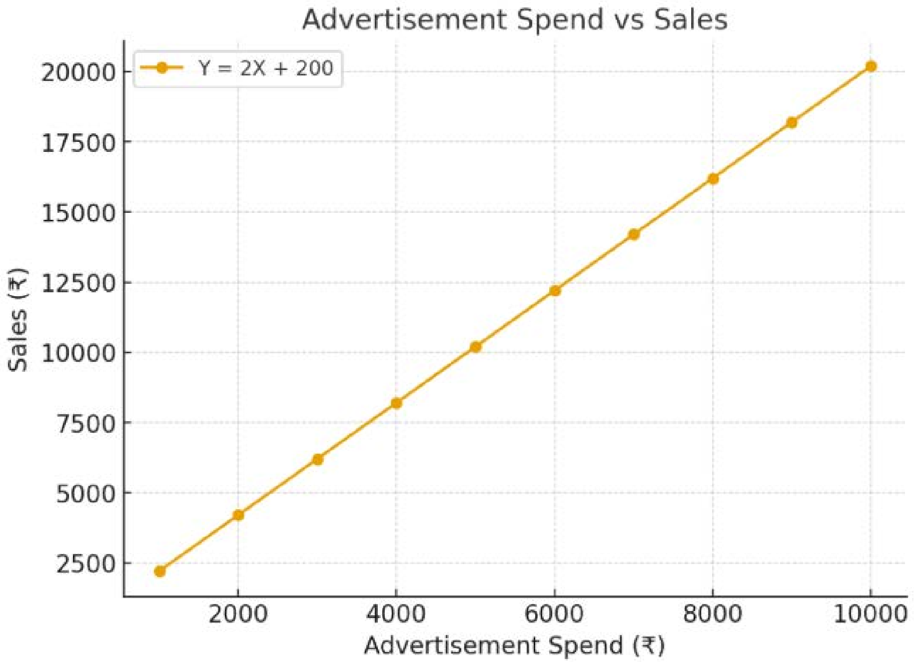
$$Y = 2X$$

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

Example 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**:

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

$$(x_1, y_1) = (2000, 4200),$$
$$(x_2, y_2) = (5000, 10200)$$

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

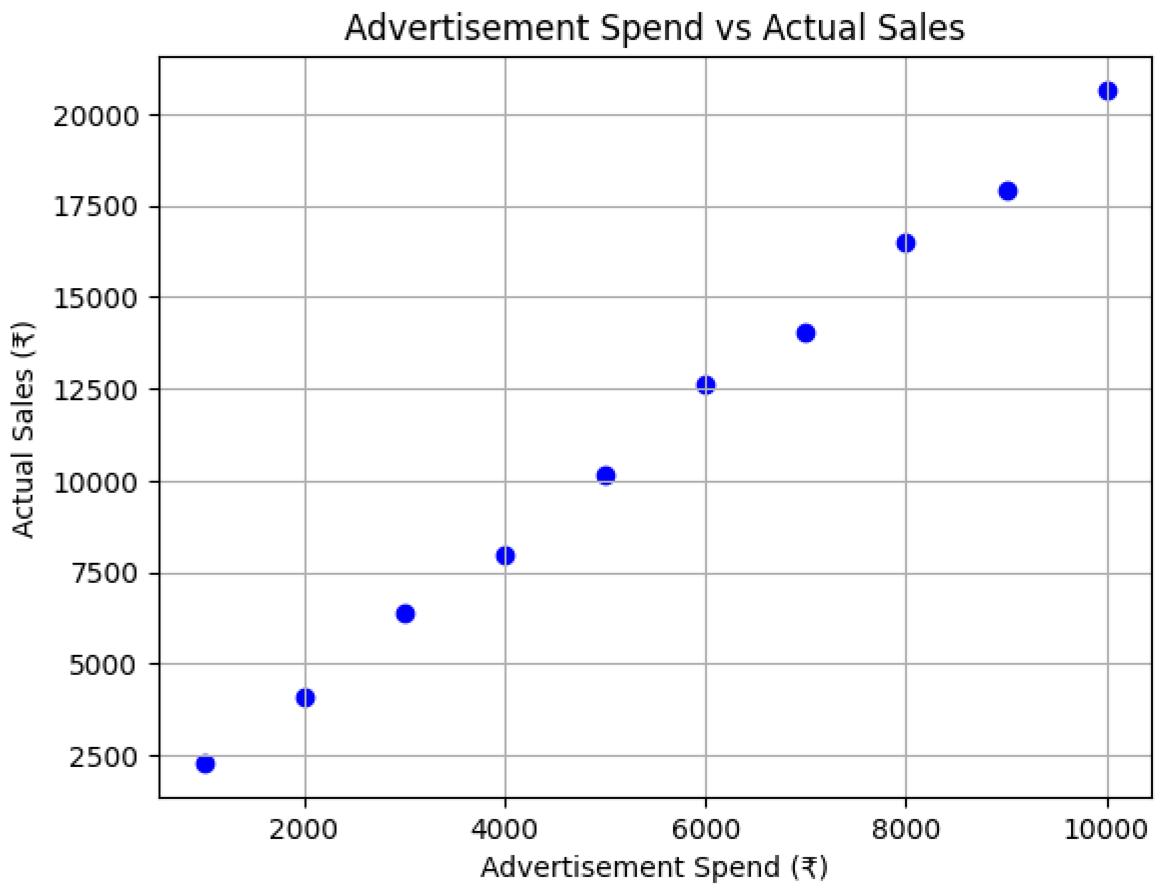
$$y - 4200 = \frac{6000}{3000}(x - 2000)$$

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

$$y = 2x + 200$$

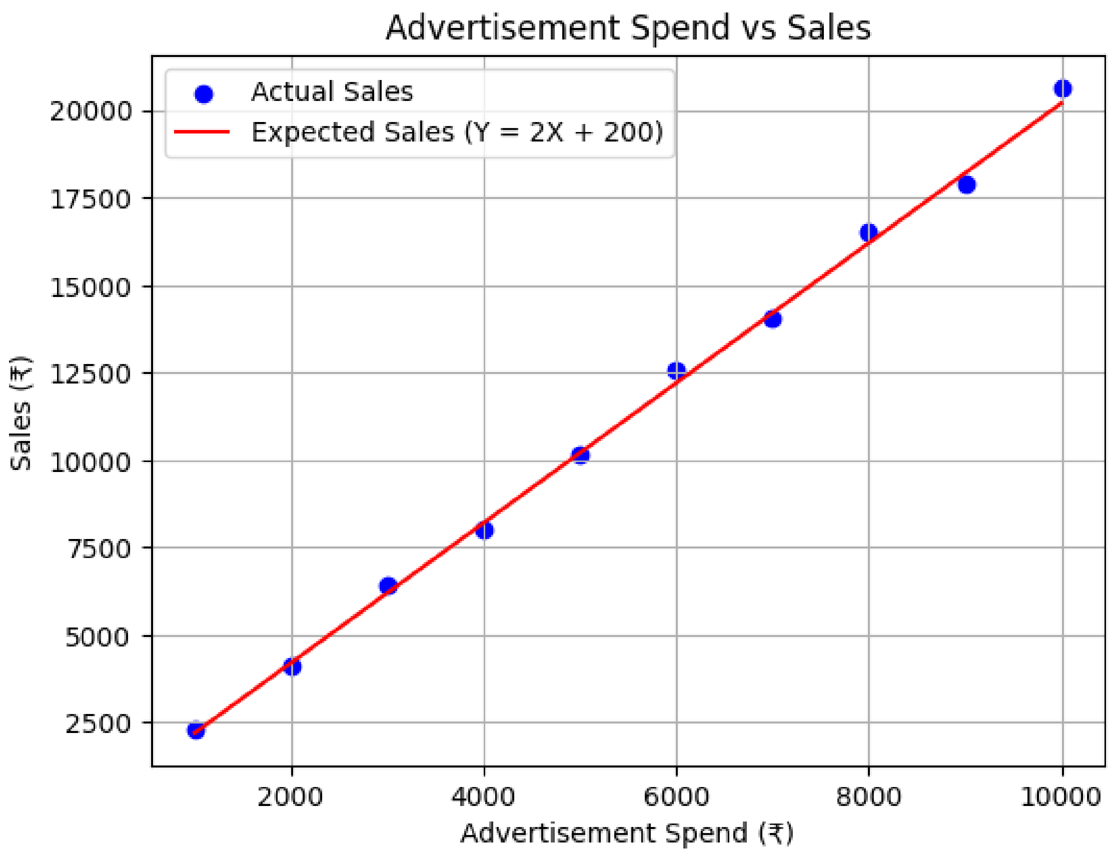
Example 4 - Advertisement vs. Sales dataset

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	



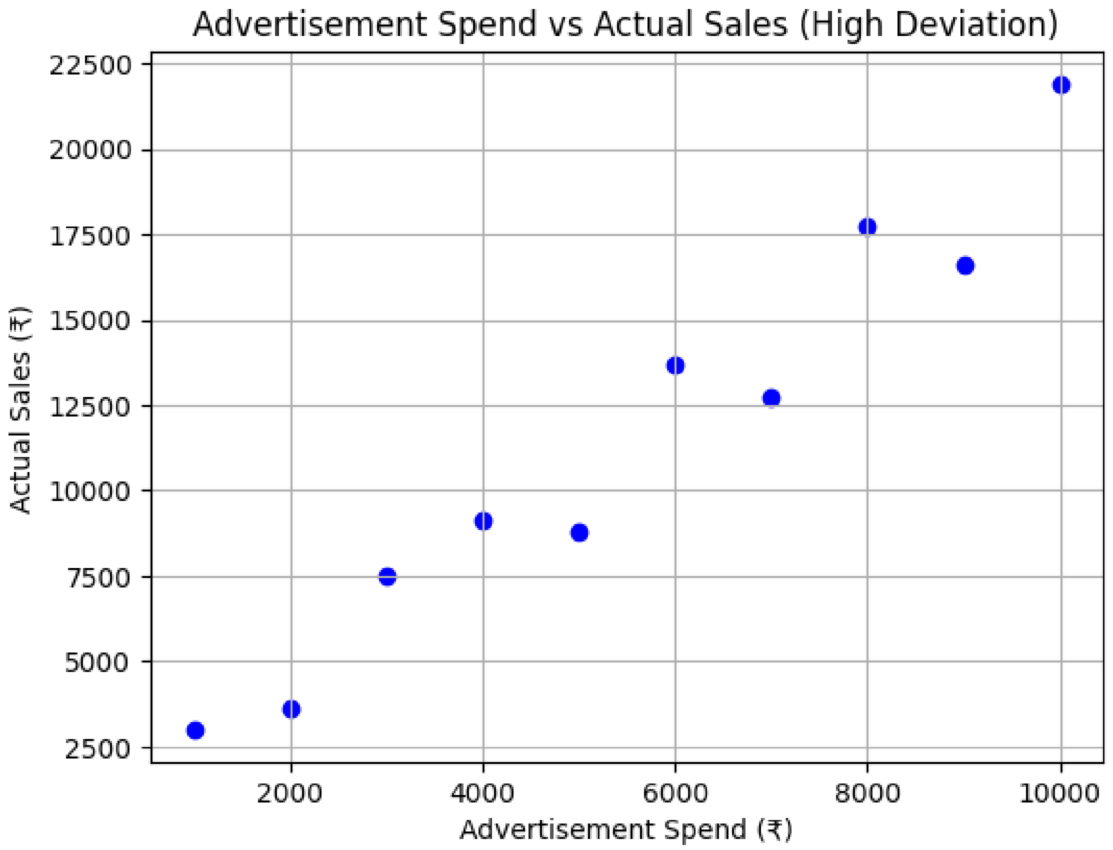
Example 4 - Advertisement vs. Sales dataset

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



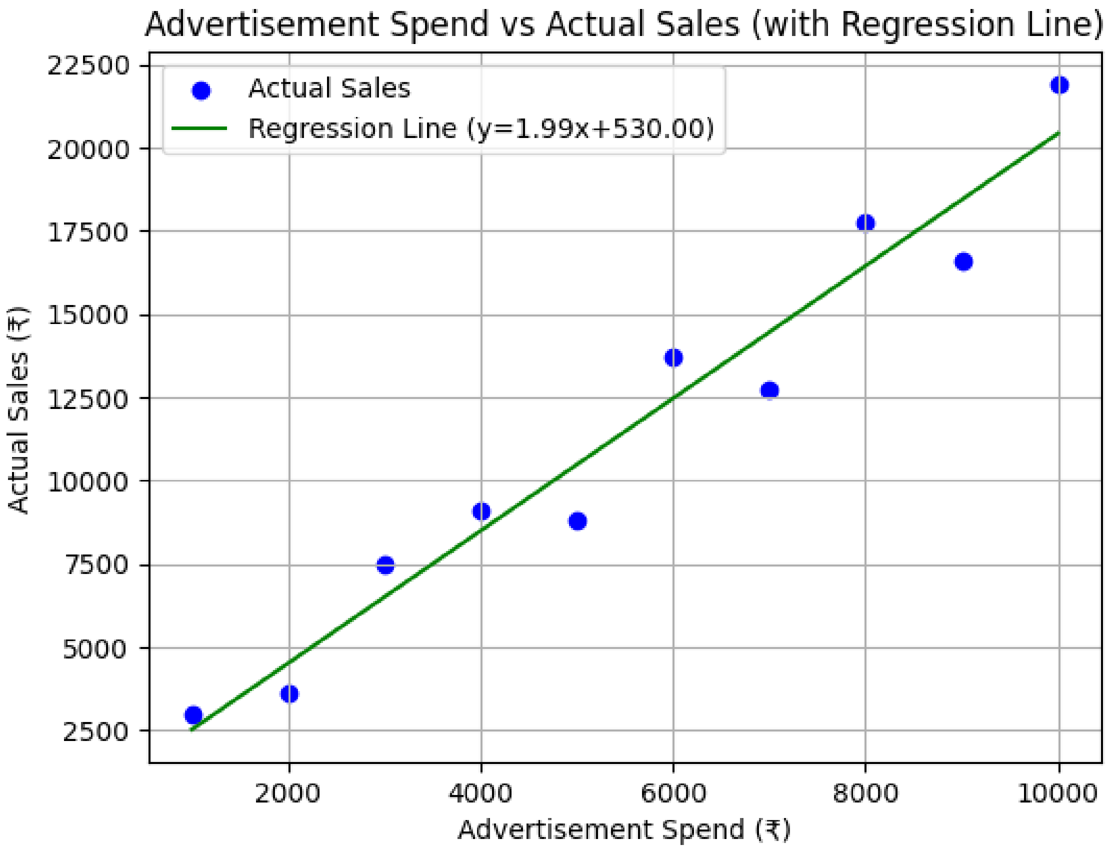
Example 5 - Advertisement vs. Sales dataset

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	

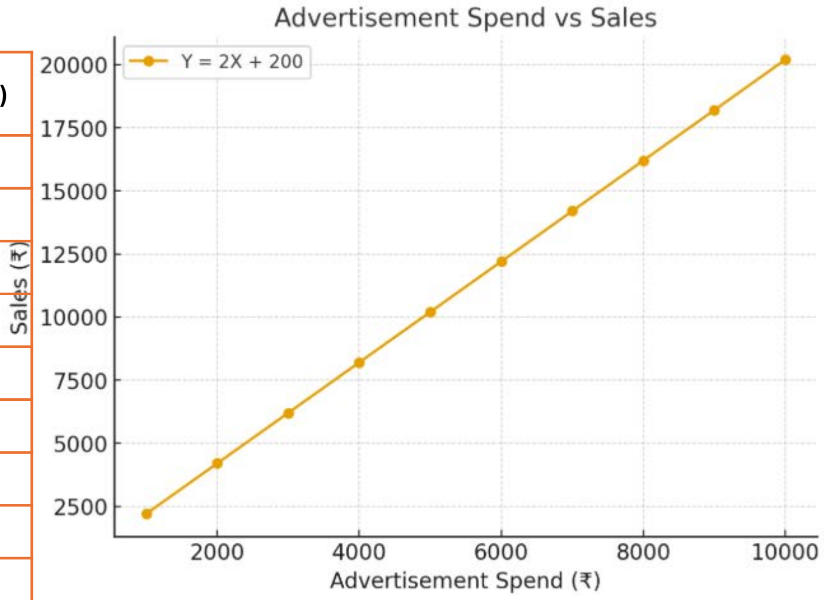


Linear Regression

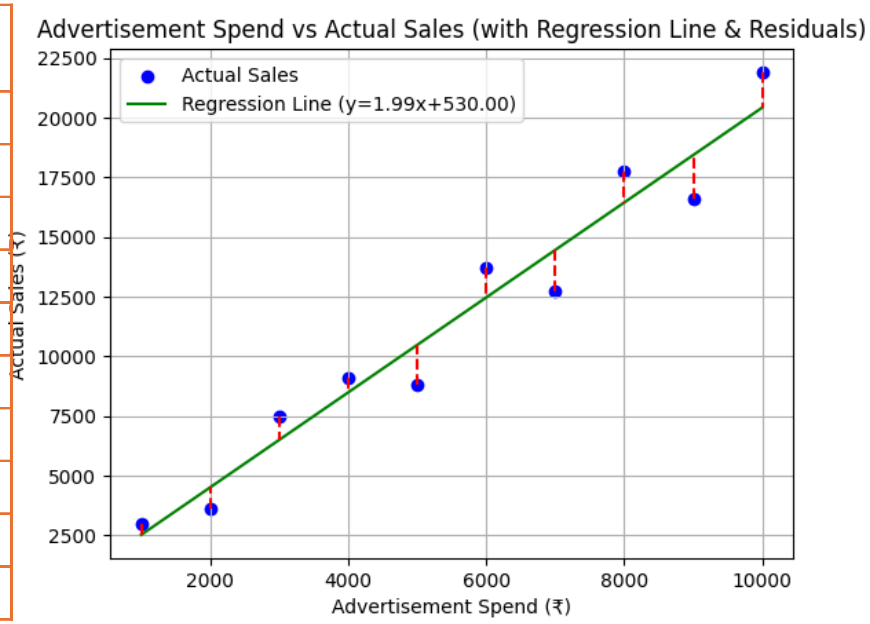
- When your data points don’t lie exactly on a straight line (because of noise, measurement errors, or natural variability), linear regression finds the **best-fit line** that minimizes the error.

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 $X = 4000$:

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

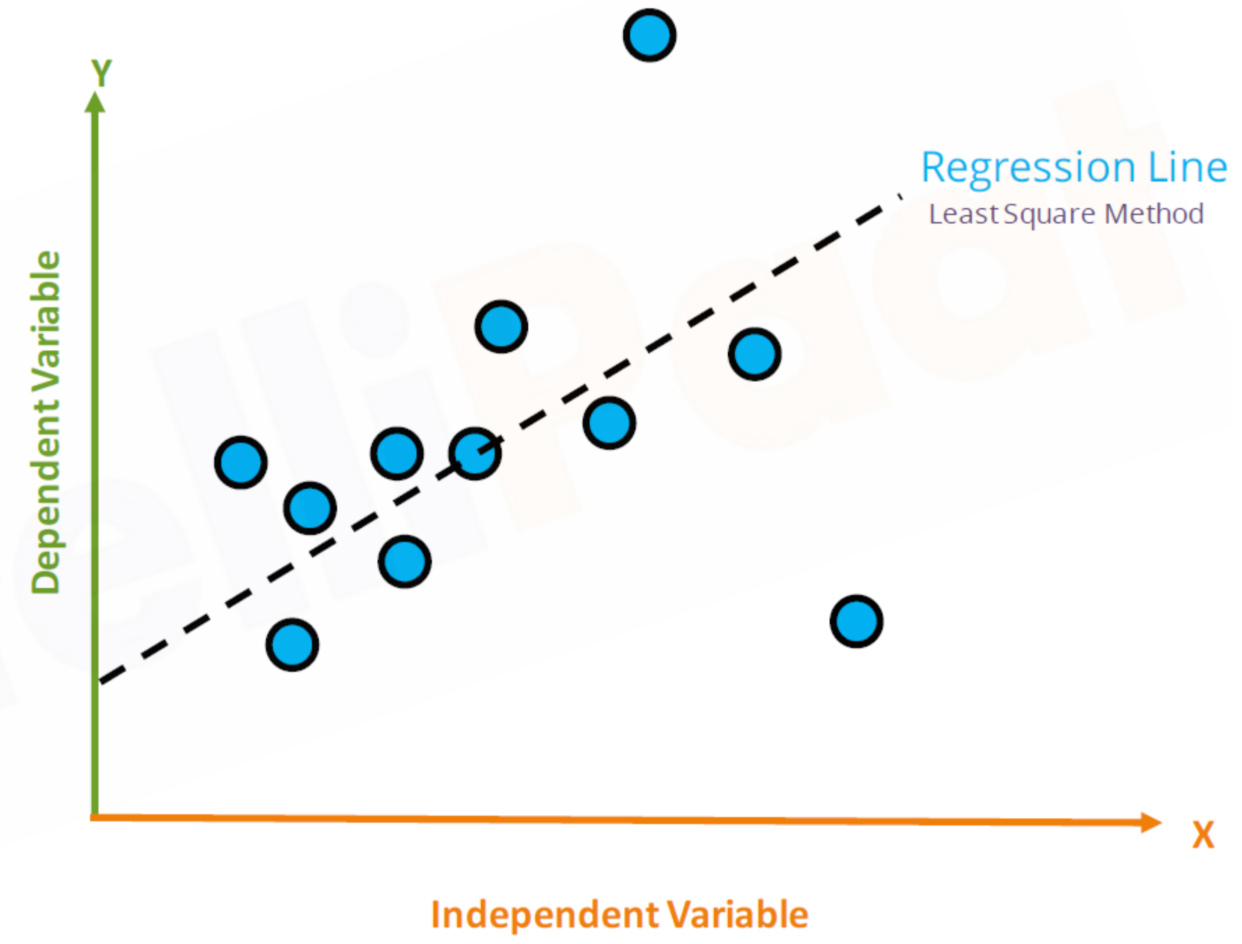
For $X = 6000$:

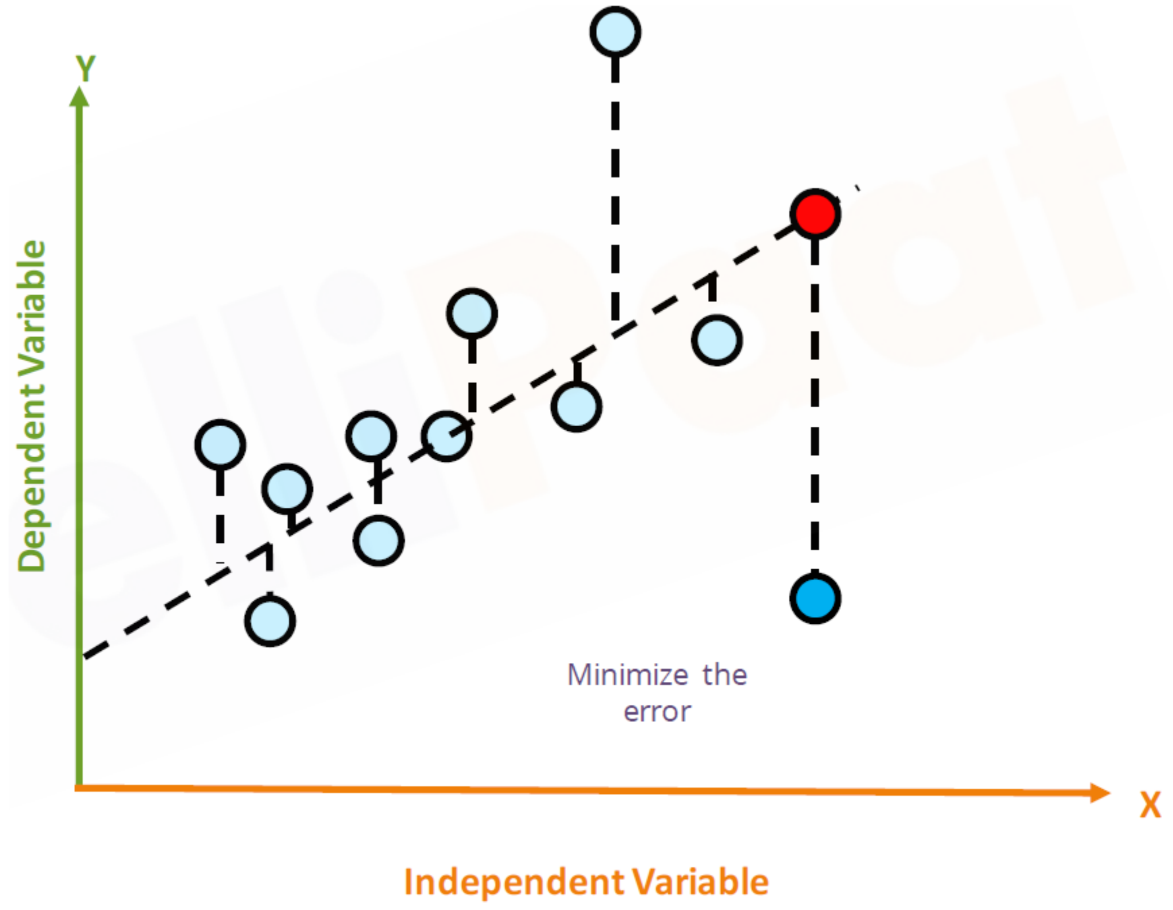
$$Y = 1.99(6000) + 530 = 11940 + 530 = 12470$$

Lease Square Fitting

Best Fit Line Equation:

$$\hat{y} = mx + c$$





Error Function (Mean Squared Error):

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

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

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

Gradients for Gradient Descent:

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

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

Gradient Descent Update Rules:

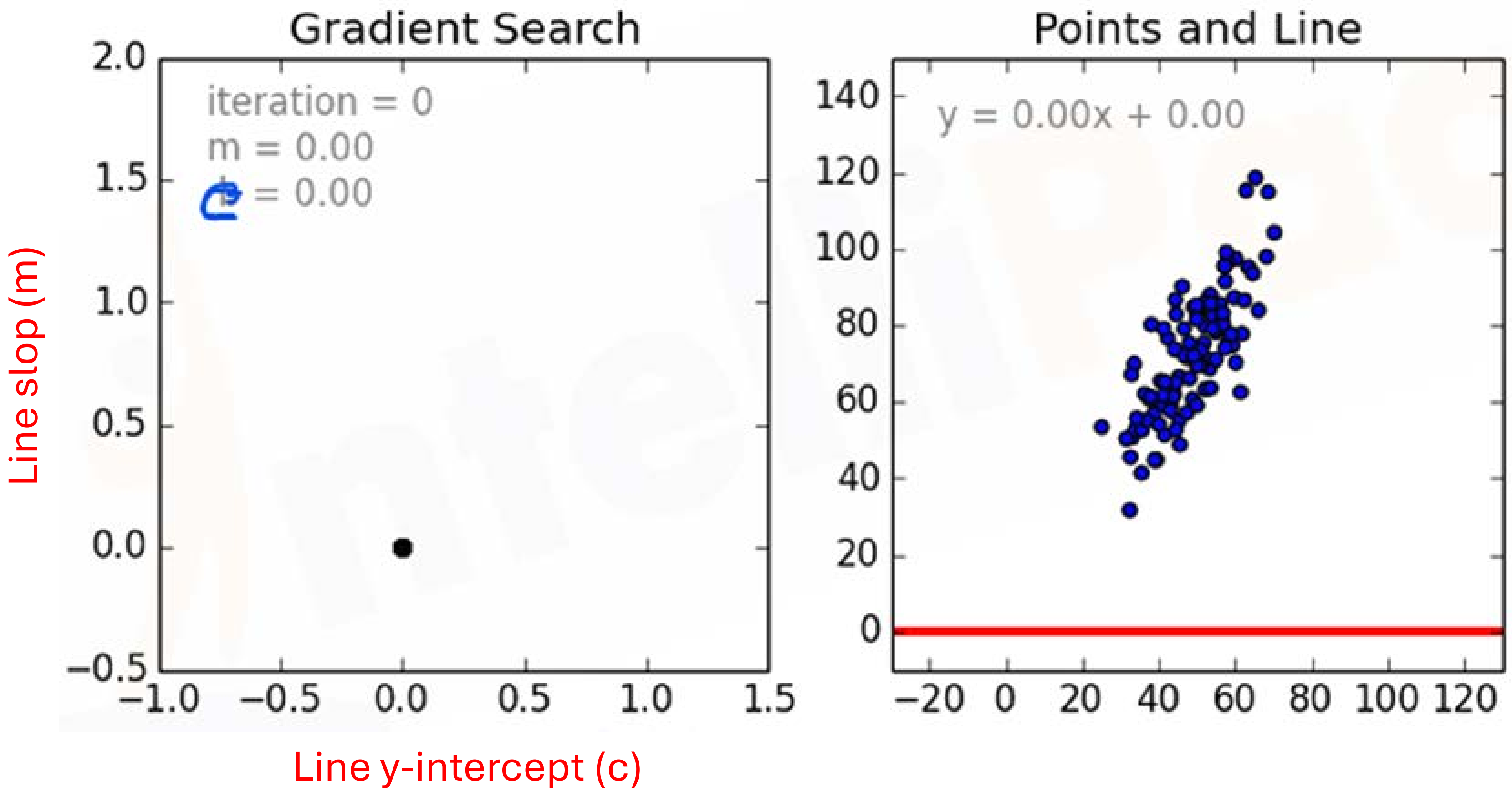
$$m \leftarrow m - \alpha \cdot \frac{\partial \text{Error}}{\partial m}$$

$$c \leftarrow c - \alpha \cdot \frac{\partial \text{Error}}{\partial c}$$

Type	Learning Rate α	Comment
Small (safe)	0.001 to 0.01	Slow but stable convergence
Medium (common)	0.01 to 0.1	Often used for standard problems; good balance
Large (risky)	0.1 to 1.0	Faster but may overshoot or diverge if error surface is steep or irregular

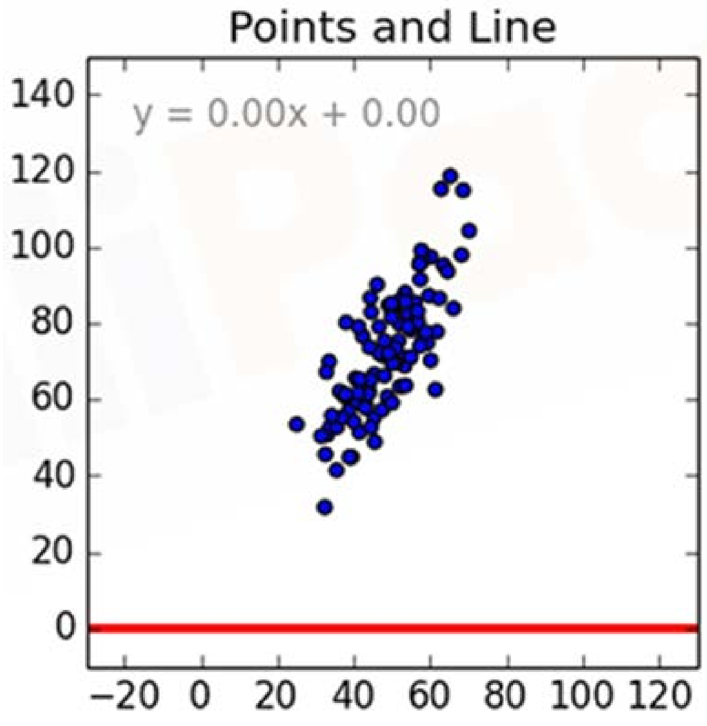
$$m \leftarrow m - \alpha \cdot \frac{\partial \text{Error}}{\partial m}$$

$$c \leftarrow c - \alpha \cdot \frac{\partial \text{Error}}{\partial c}$$

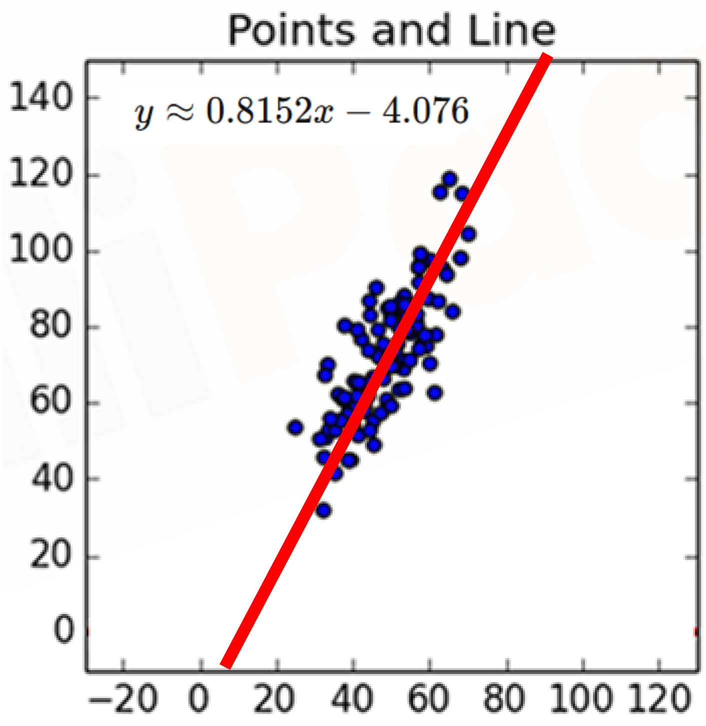


$$m \leftarrow m - \alpha \cdot \frac{\partial \text{Error}}{\partial m}$$

$$c \leftarrow c - \alpha \cdot \frac{\partial \text{Error}}{\partial c}$$



Initial



Final