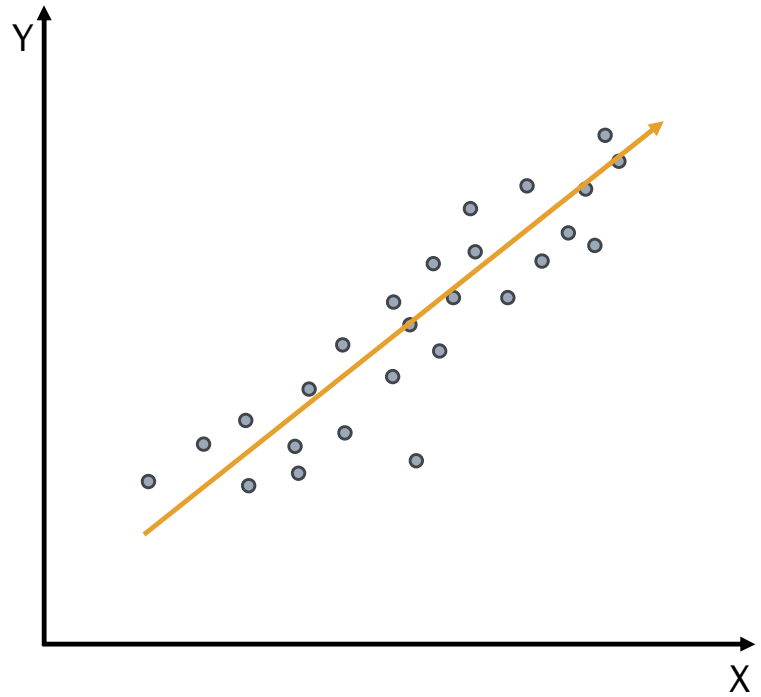
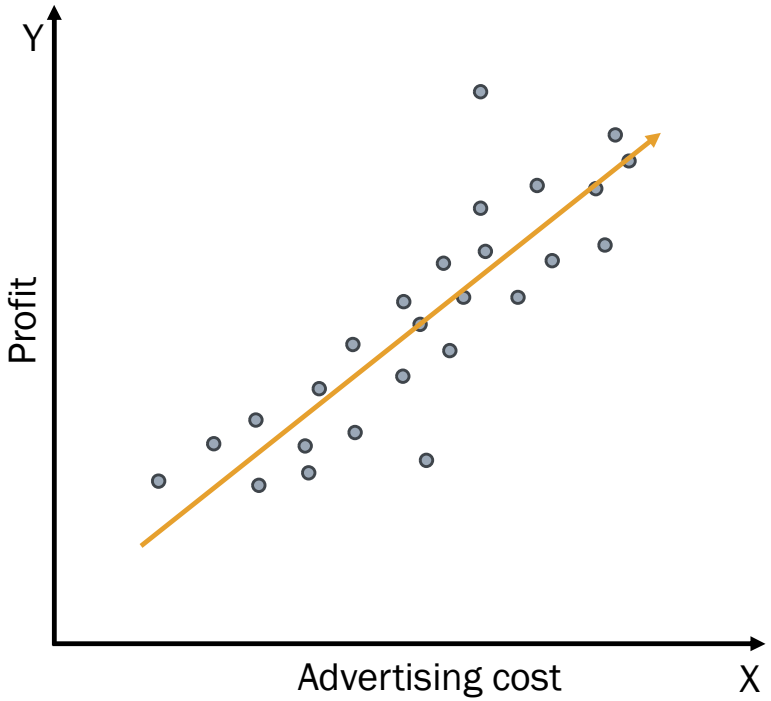
Linear Regression

A Fundamental Approach to Predictive Modeling



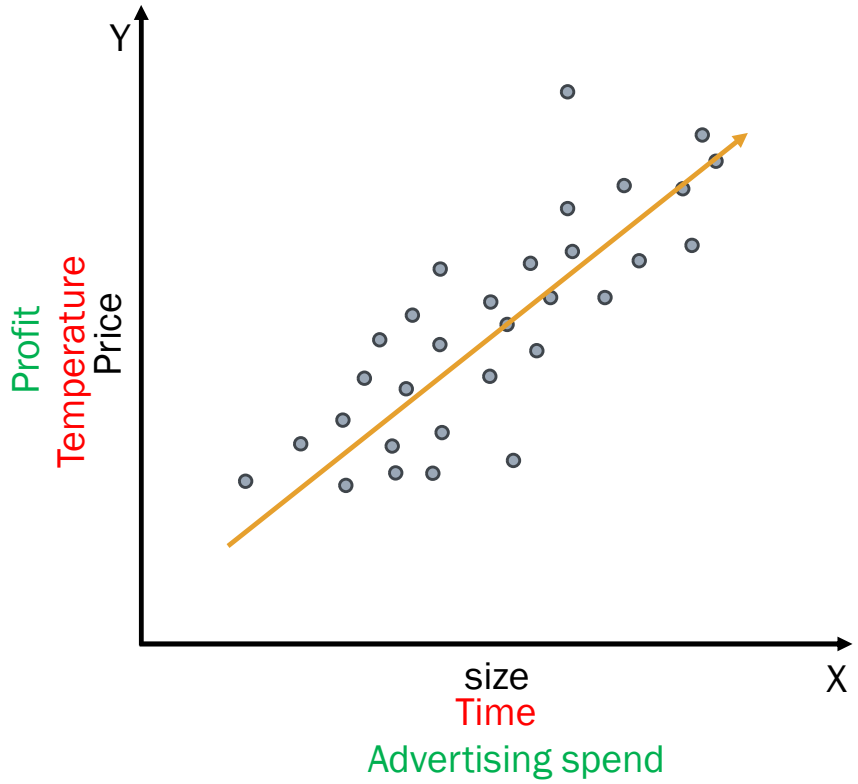
What is Linear Regression?

- **Definition:** Linear regression is a statistical method to model the relationship between a dependent variable (y) and one or more independent variables (x).
- **Goal:** To find a linear relationship between variables, often for prediction or understanding data trends.



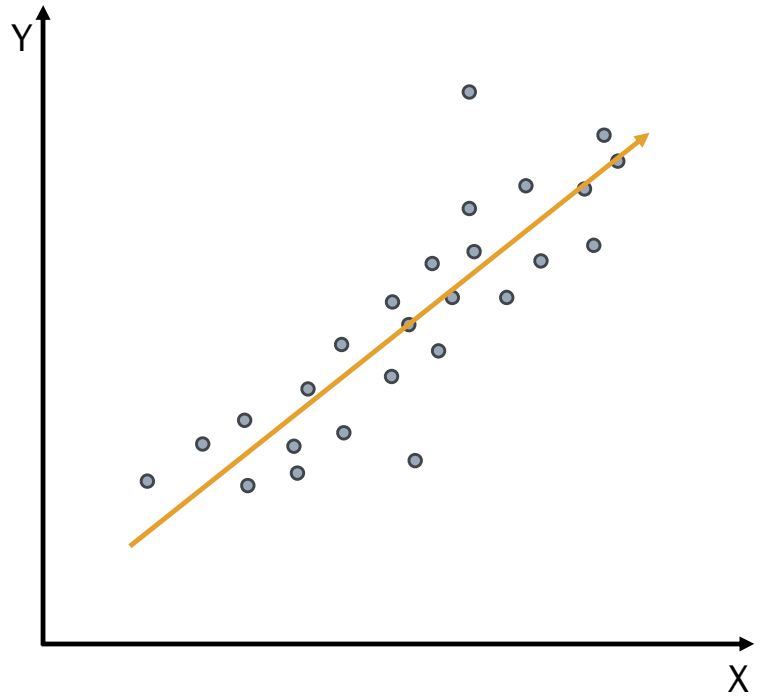
Real-World Examples

- Predicting house prices based on size and location.
- Estimating a company's profit based on advertising spend.
- Forecasting temperature changes over time.



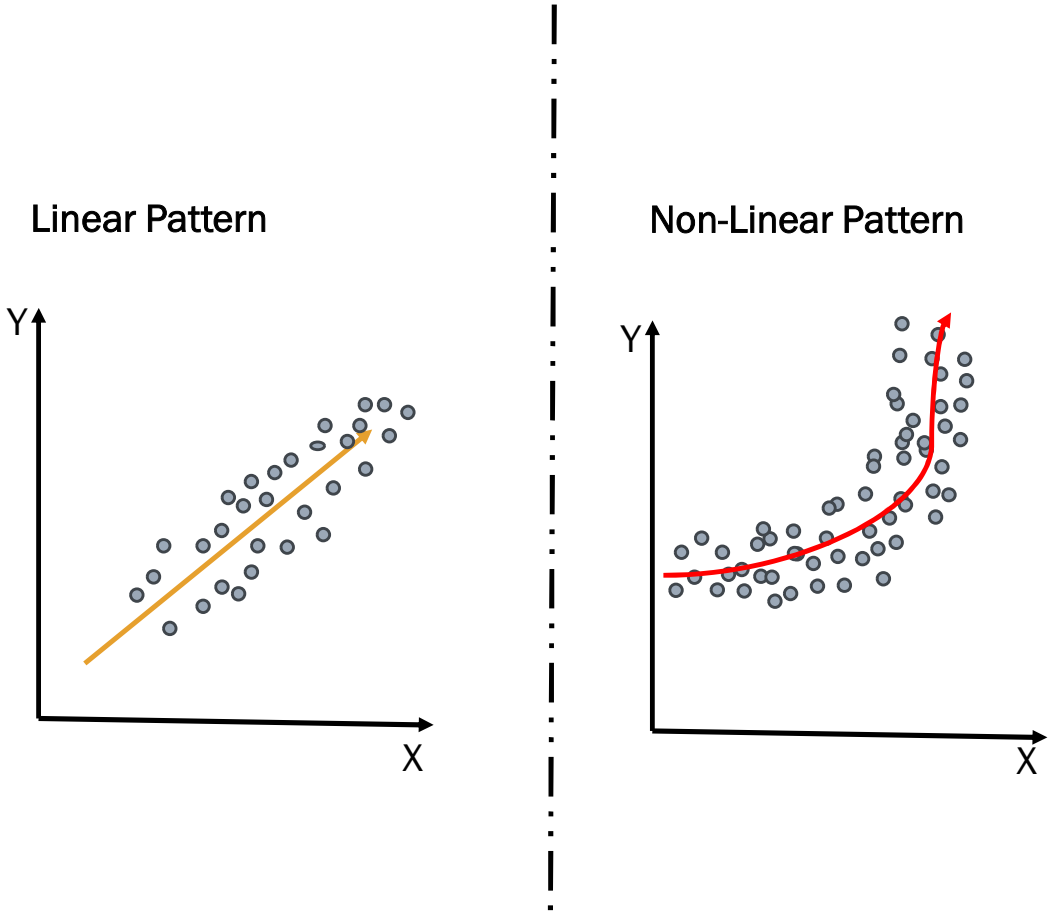
Assumptions of Linear Regression

- **Linearity:** Relationship between x and y is linear.
- **Independence:** Observations are independent.
- **Homoscedasticity:** Constant variance of errors.
- **Normality:** Errors are normally distributed.



Assumptions of Linear Regression

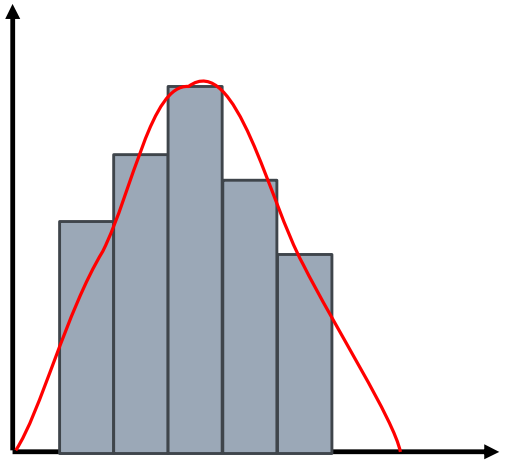
1. **Linearity:** Relationship between x and y is linear.



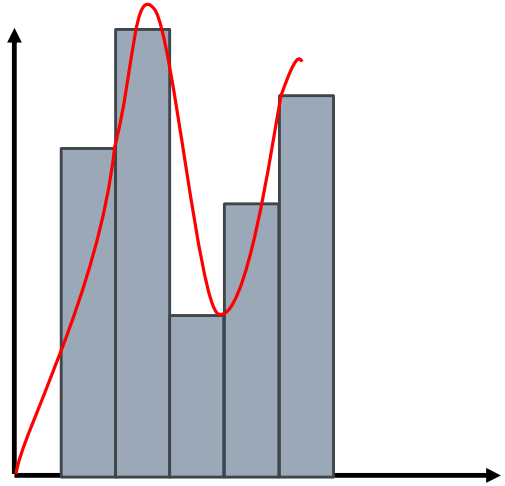
Assumptions of Linear Regression

1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.



Error terms normally distributed



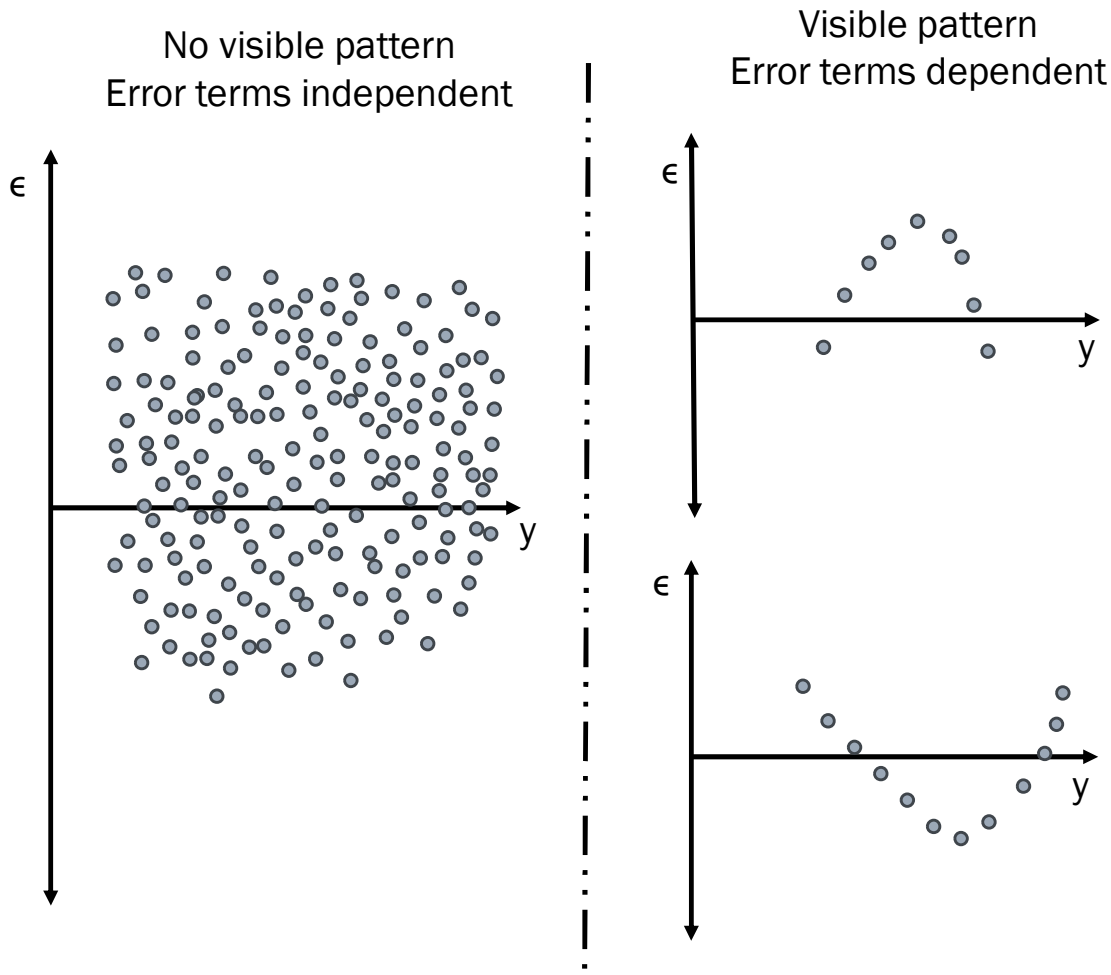
Error terms NOT normally distributed

Assumptions of Linear Regression

1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.

3. **The error terms** are independent of each other.



Assumptions of Linear Regression

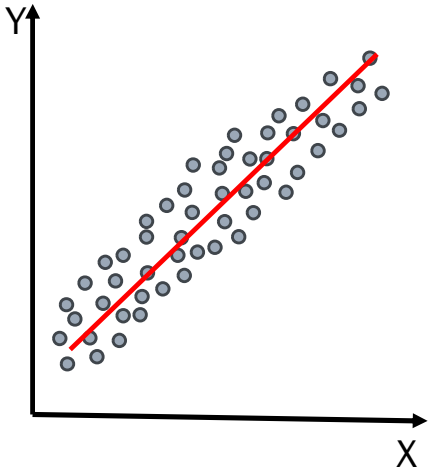
1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.

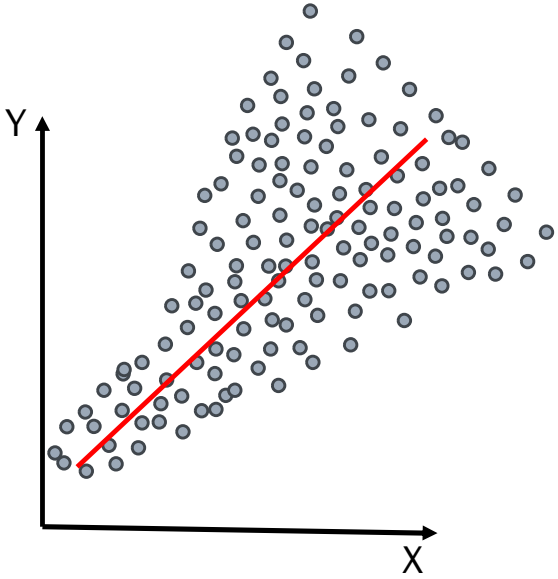
3. **The error terms** are independent of each other.

4. **Homoscedasticity: Errors have** Constant variance.

Constant Variance
(Homoscedastic)

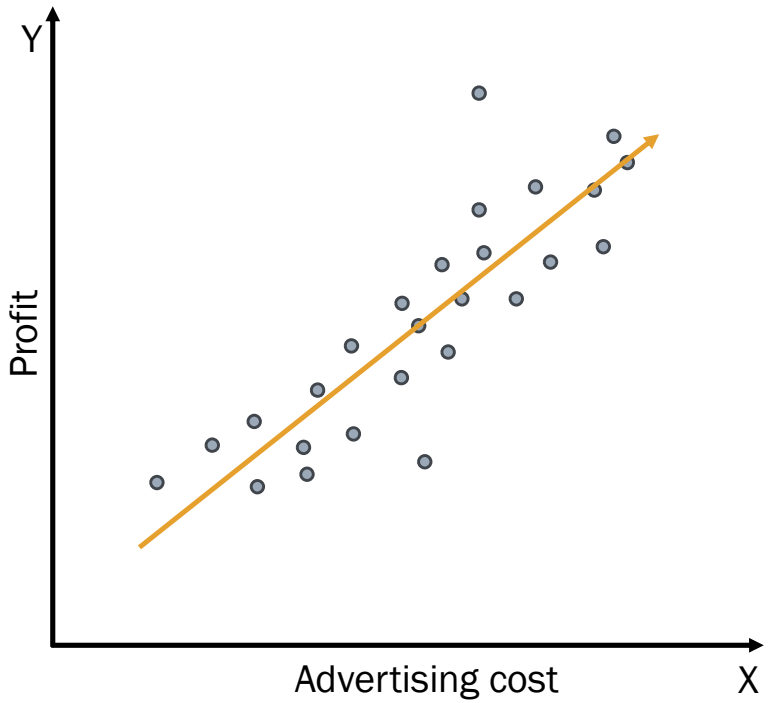


Changing Variance
(Heteroscedastic)



Key Concepts in Linear Regression

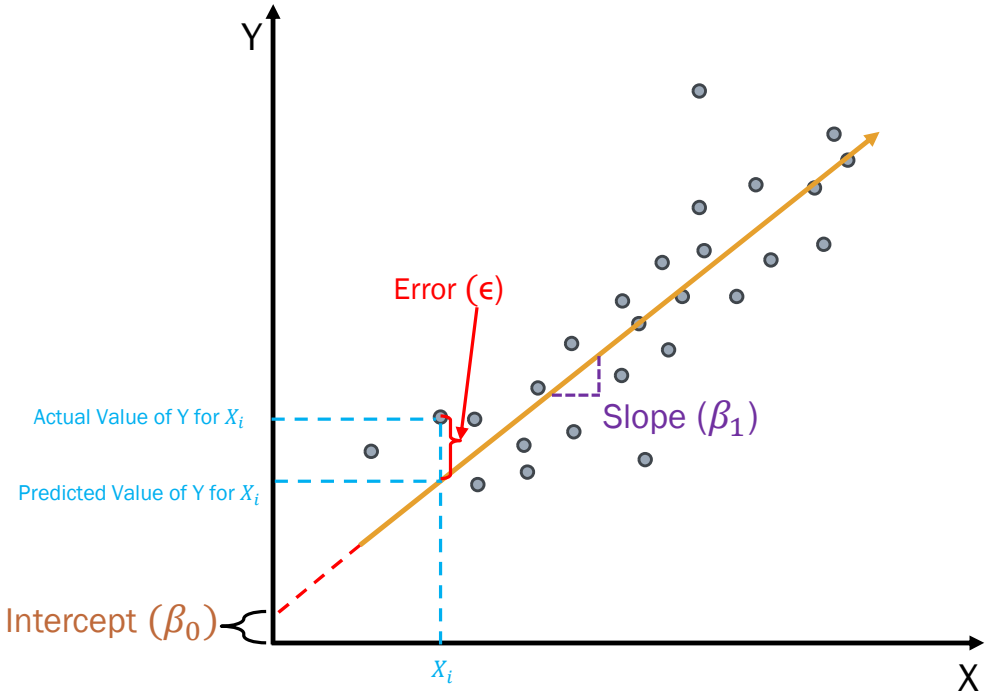
- **Independent Variable (x):** The variable(s) used to make predictions about y.
- **Dependent Variable (y):** The variable we want to predict or understand.
- **Model Assumption:** Linear relationship between x and y.



Understanding the Slope & Intercept

- Intercept (β_0): Intercept (where the line crosses the y-axis).
- Slope (β_1) : Slope (change in y per unit change in x).
- ϵ : Error term (differences between actual and predicted y values).
- The Linear Regression Equation

$$y = \beta_0 + \beta_1 x + \epsilon$$



Understanding the Slope & Intercept

- Slope (β_1) : Slope (change in y per unit change in x).

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

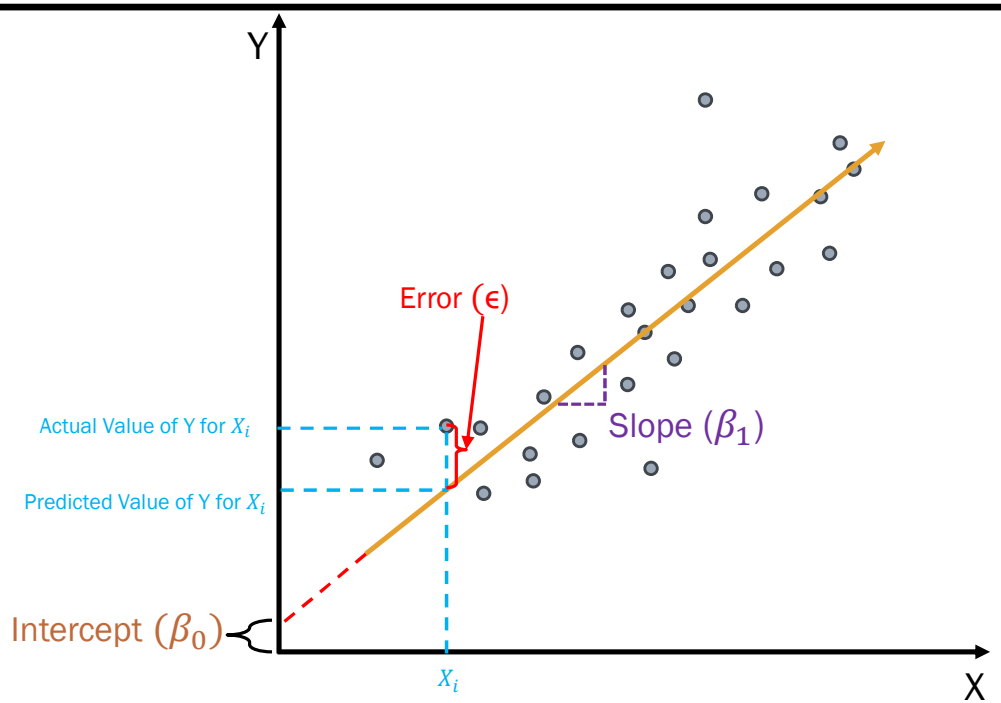
- Intercept (β_0): Intercept (where the line crosses the y-axis).

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where,

$\bar{x}$ = Mean of x
$\bar{y}$ = Mean of y

$$y = \beta_0 + \beta_1 x + \epsilon$$



Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

You have collected the following data from five students:

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

Task:

1. Fit a simple linear regression model to the data.
2. Find the coefficients β_0 (intercept) and β_1 (slope).
3. Interpret the coefficients.
4. Predict the test score for a student who studied for 3.5 hours.

Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

Solution Steps:

1. Calculate the Means:
   - Mean of x = $\bar{x}$ = $\frac{(1+2+3+4+5)}{5} = 3$
   - Mean of y = $\bar{y}$ = $\frac{(55+65+70+75\quad)}{5} = 70$
2. Calculate β_1 (slope):

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{(1-3)(55-70) + (2-3)(65-70) + (3-3)(70-70) + (4-3)(75-70) + (5-3)(85-70)}{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}$$

$$= \frac{(-2)(-15)+(-1)(-5)+(0)(0)+(1)(5)+(2)(15)}{4+1+0+1+4}$$

$$= \frac{30+5+0+5+30}{10} = \frac{70}{10} = 7$$

Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

Solution Steps:

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

3. Calculate β_0 (intercept):

   $\beta_0 = \bar{y} - \beta_1 \bar{x} = 70 - 7 * 3 = 70 - 21 = 49$

4. Final Model:

   $y = 49 + 7x$

5. Interpretation of Coefficients:

   β_0 = 49: When no hours are studied, the predicted test score is 49.

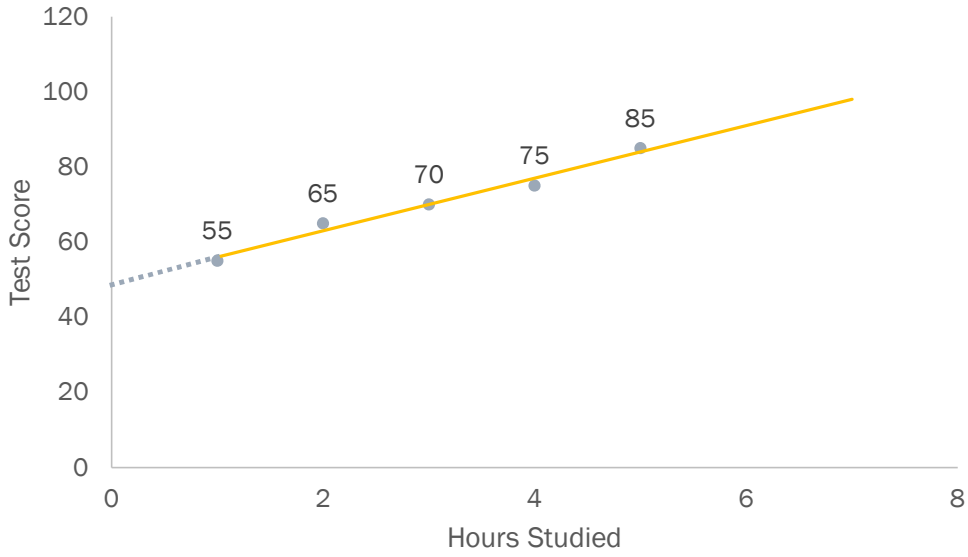
   β_1 =7: For each additional hour studied, the test score is expected to increase by 7 points.

Trendline for our Equation: $y = 49 + 7x$

Prediction for 3.5 hours:

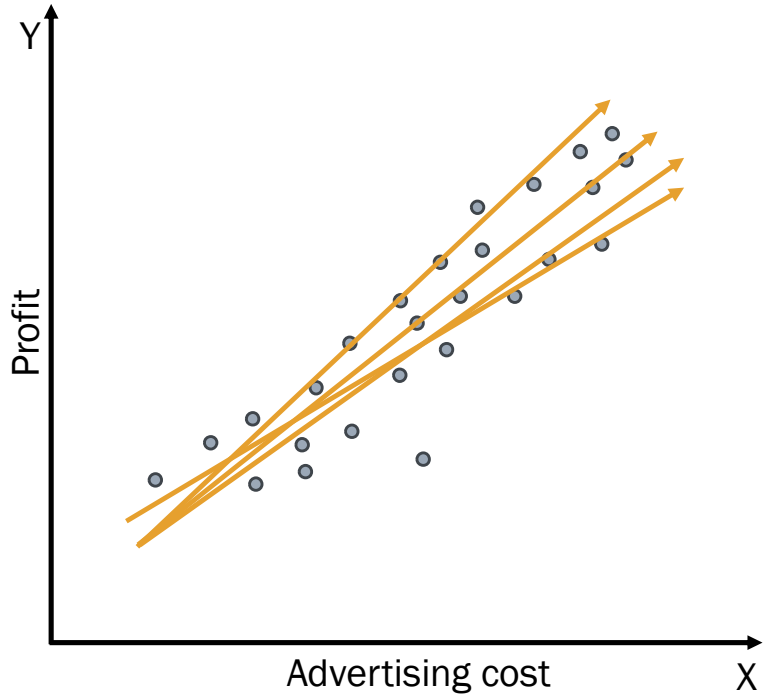
$$y = 49 + 7(3.5) = 49 + 24.5 = 73.5$$

Thus, a student who studies for **3.5 hours** is predicted to score approximately **73.5 on the test.**



The Concept of "Line of Best Fit"

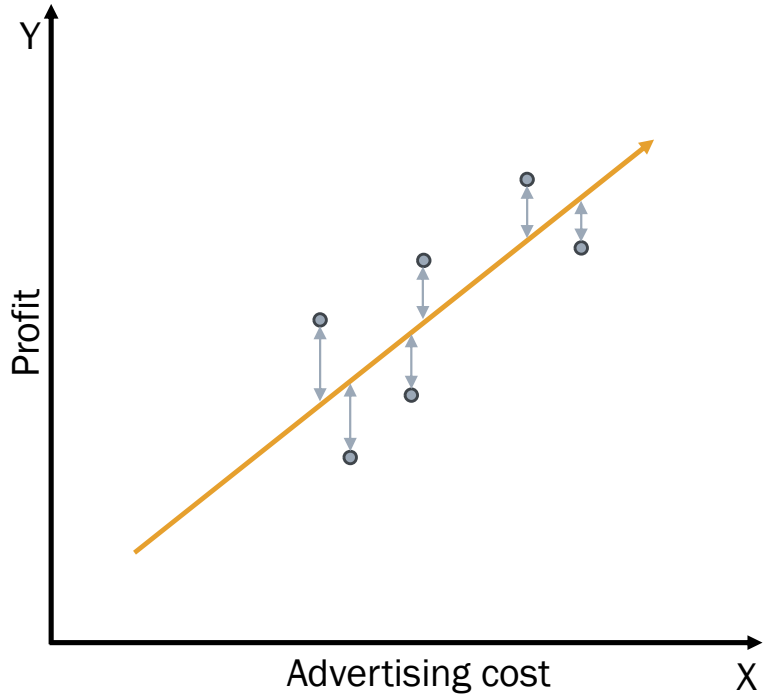
$$y = \beta_0 + \beta_1 x + \epsilon$$



The Concept of "Line of Best Fit"

- **Definition:** The line that best minimizes the difference between predicted and actual values (minimizes errors).
- **Goal:** Minimizing the sum of squared differences between actual and predicted y values.
- Residual

$$y = \beta_0 + \beta_1 x + \epsilon$$



The Concept of "Line of Best Fit"

- **Definition:** The line that best minimizes the difference between predicted and actual values (minimizes errors).
- **Goal:** Minimizing the sum of squared differences between actual and predicted y values.
- Residual Sum of Squares (RSS)

$$y = \beta_0 + \beta_1 x + \epsilon$$

The formula for the residual for each point is:

$$Residual = y_{actual} - y_{predicted}$$

The sum of squared residuals (RSS) is given by:

$$RSS = \sum (y_{actual} - y_{predicted})^2$$

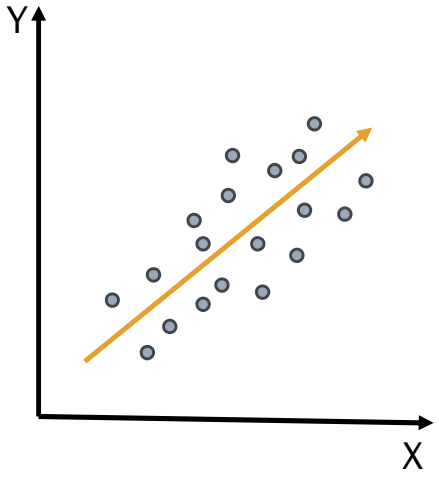
Where:
y_{actual} is the actual value.
$y_{predicted} = \beta_0 + \beta_1 x$,is the value predicted by the linear model, with β_1 as the slope and β_0 as the intercept

Evaluating Model Fit – R-squared

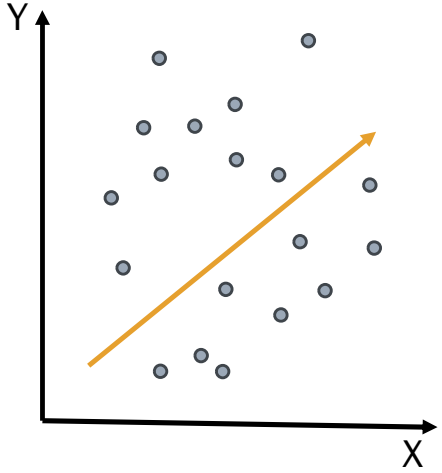
- **R-squared (R^2):** A metric to assess how well the model fits the data, representing the proportion of variance in y explained by x.
- **Values:** Ranges from 0 to 1, where closer to 1 means a better fit.

An R^2 value closer to 1 suggests a better fit.

R^2 = 0.81

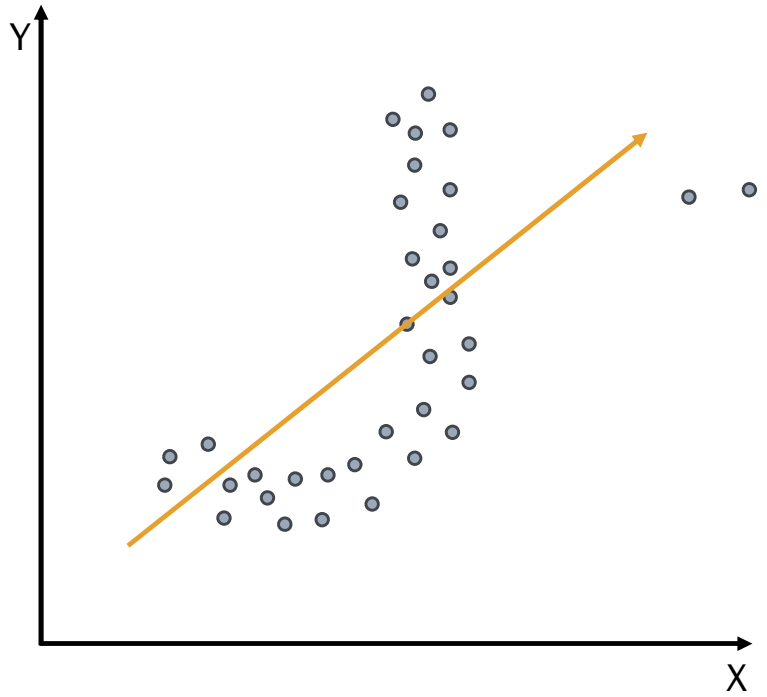


R^2 = 0.24



Limitations of Linear Regression

- **Limited to Linear Relationships:** Non-linear data requires more advanced techniques.
- **Sensitivity to Outliers:** Can skew the line.



What We Covered

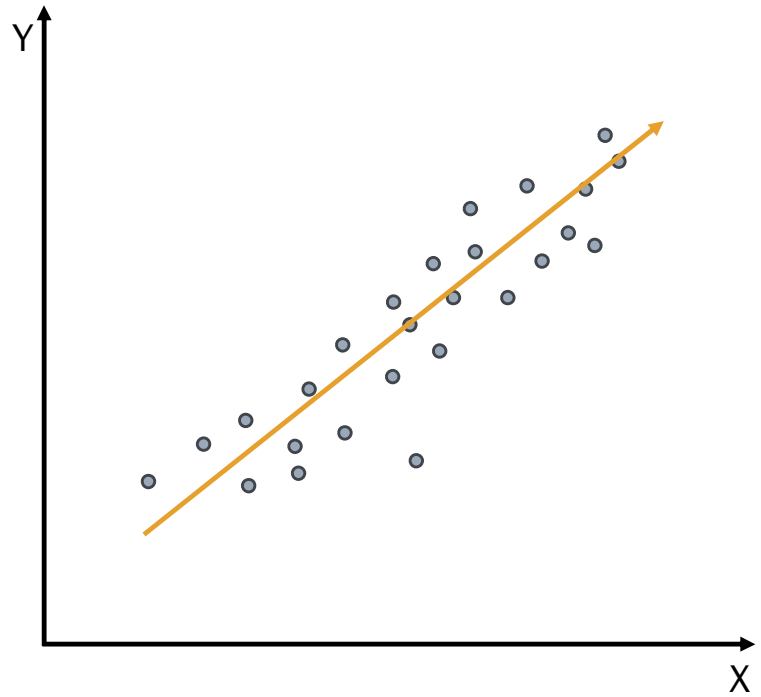
- ❑ Basic understanding of Linear Regression.
- ❑ Assumptions and Example of Linear Regression.
- ❑ The linear equation, line of best fit.
- ❑ Limitations of Linear Regression.

Thank You

- @DataByteSun

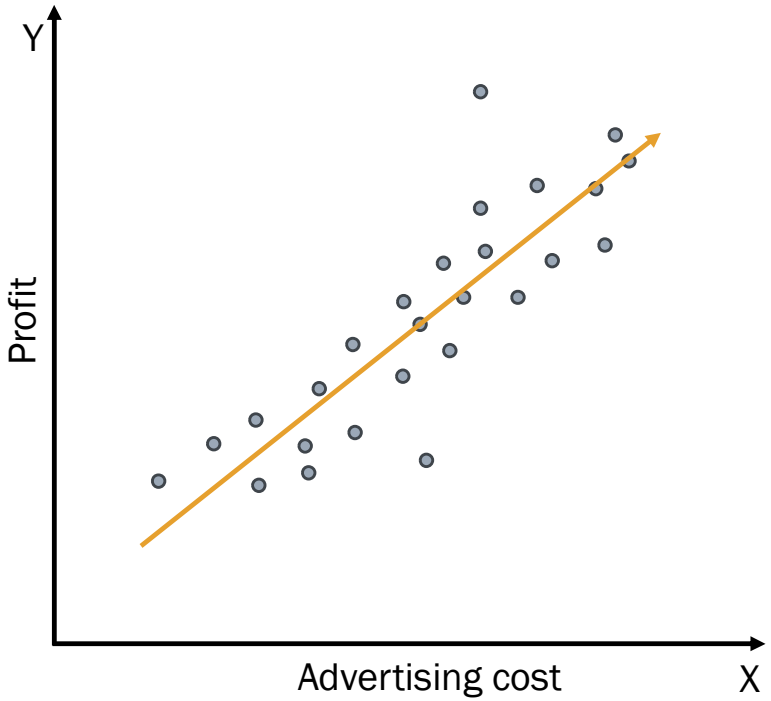
Linear Regression

A Fundamental Approach to Predictive Modeling



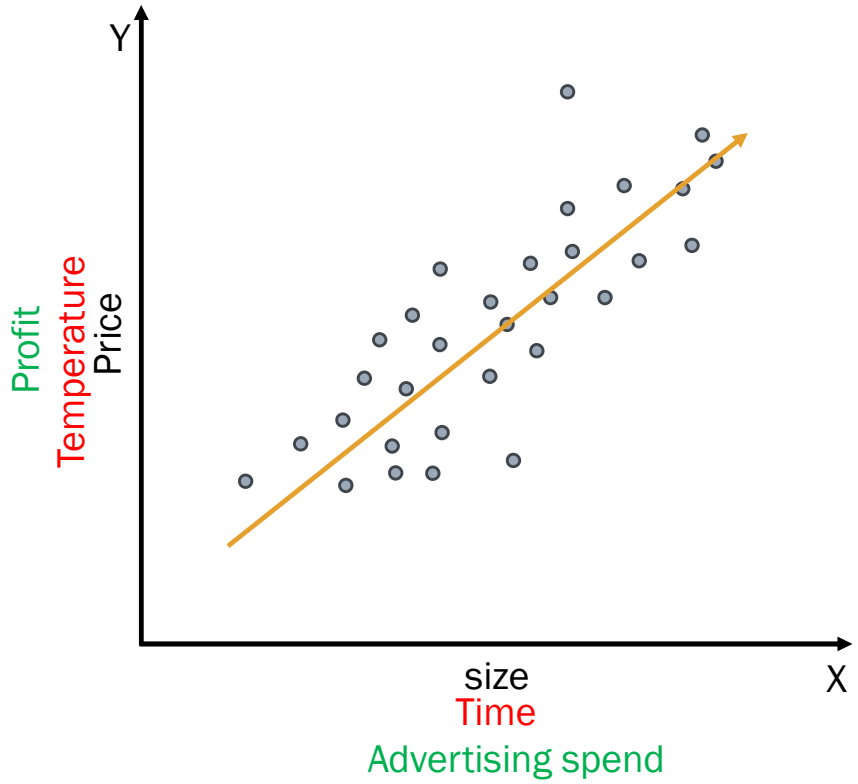
What is Linear Regression?

- **Definition:** Linear regression is a statistical method to model the relationship between a dependent variable (y) and one or more independent variables (x).
- **Goal:** To find a linear relationship between variables, often for prediction or understanding data trends.



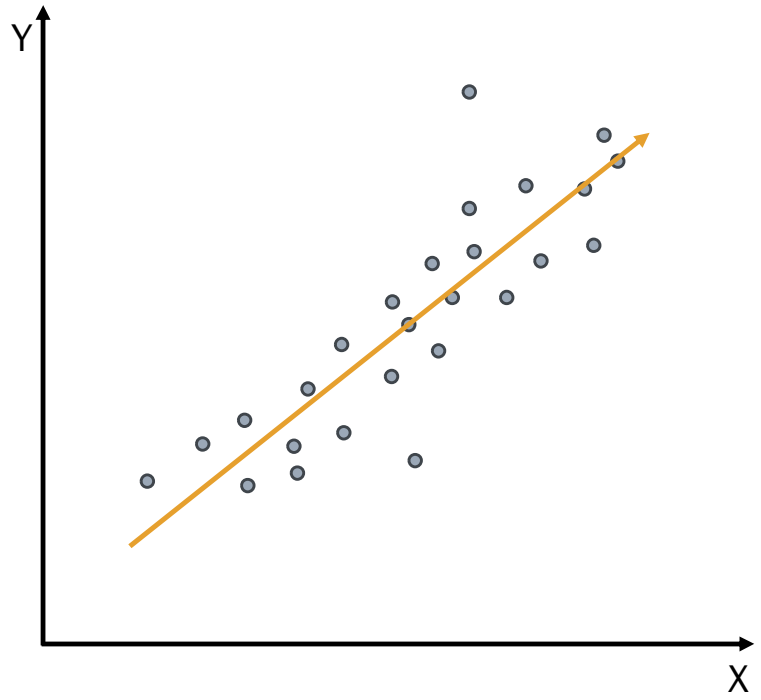
Real-World Examples

- Predicting house prices based on size and location.
- Estimating a company's profit based on advertising spend.
- Forecasting temperature changes over time.



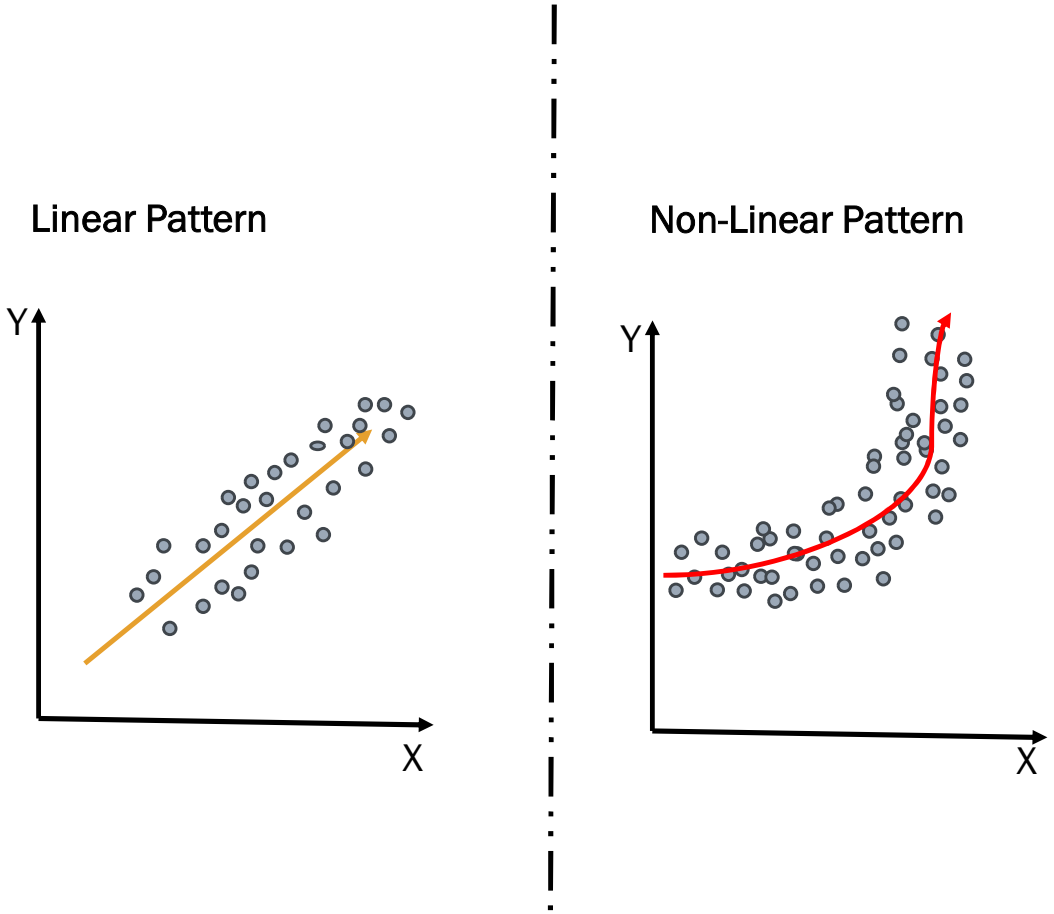
Assumptions of Linear Regression

- **Linearity:** Relationship between x and y is linear.
- **Independence:** Observations are independent.
- **Homoscedasticity:** Constant variance of errors.
- **Normality:** Errors are normally distributed.



Assumptions of Linear Regression

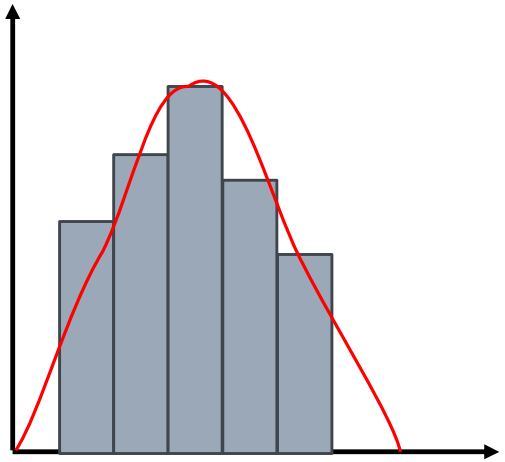
1. **Linearity:** Relationship between x and y is linear.



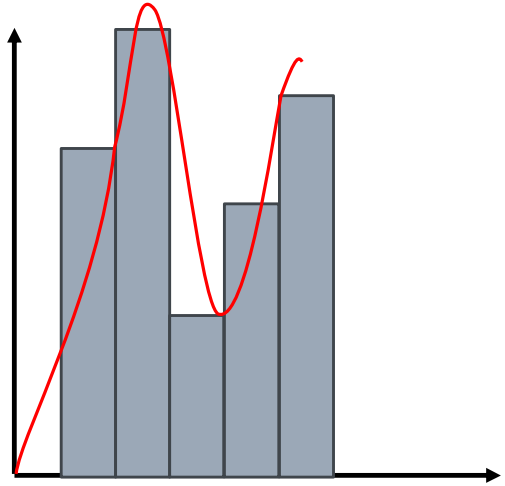
Assumptions of Linear Regression

1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.



Error terms normally distributed



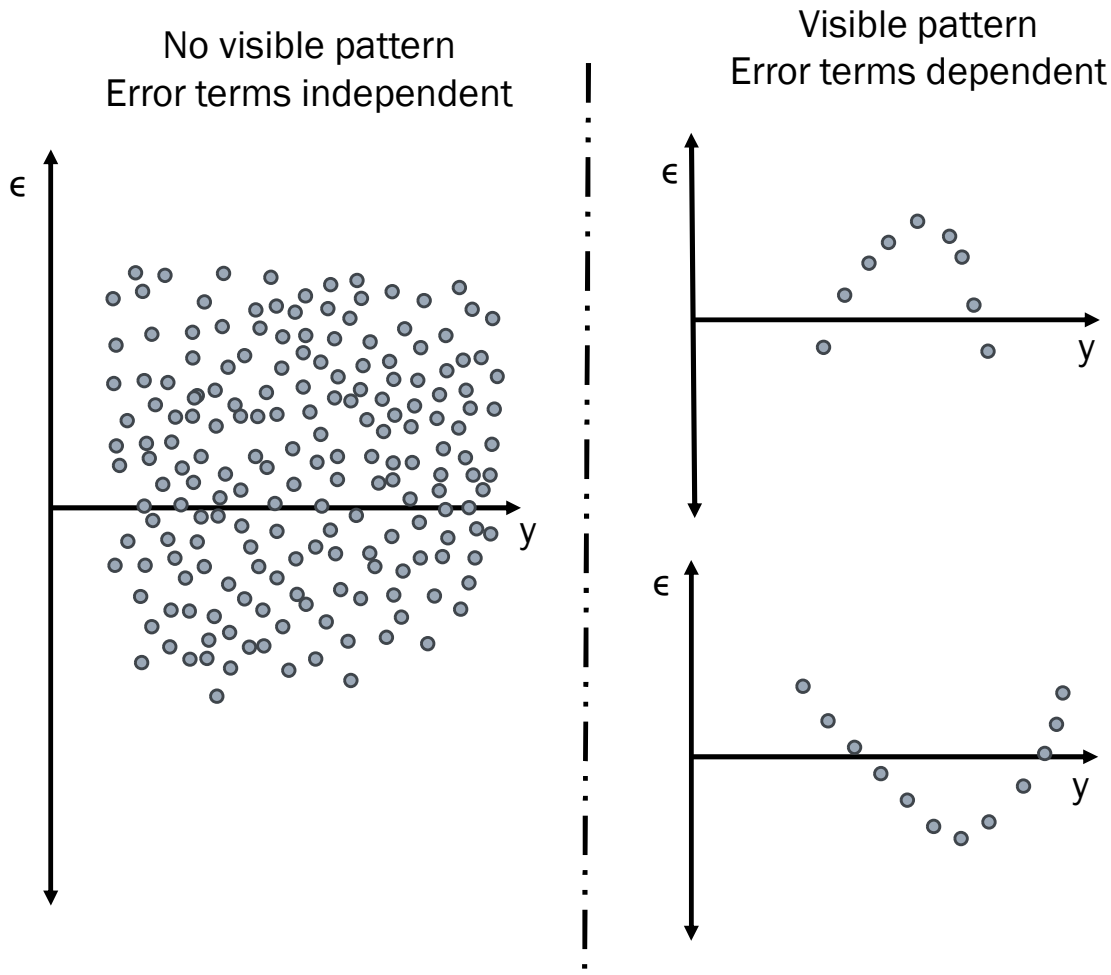
Error terms NOT normally distributed

Assumptions of Linear Regression

1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.

3. **The error terms** are independent of each other.



Assumptions of Linear Regression

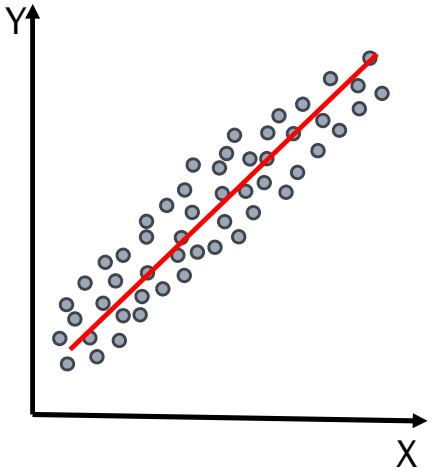
1. **Linearity:** Relationship between x and y is linear.

2. Errors are normally distributed with mean equal to zero.

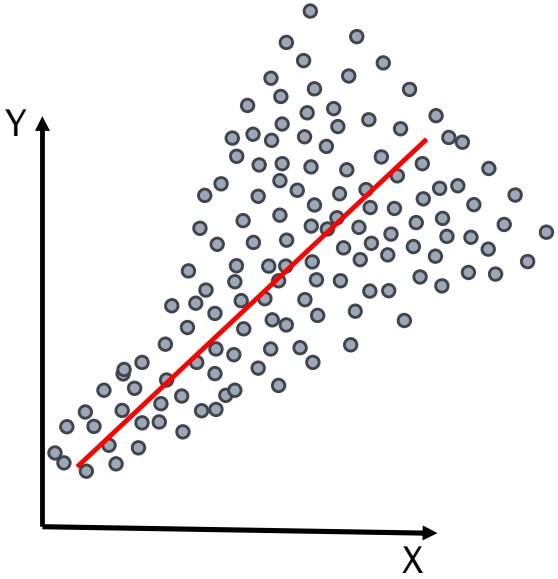
3. **The error terms** are independent of each other.

4. **Homoscedasticity: Errors have** Constant variance.

Constant Variance
(Homoscedastic)

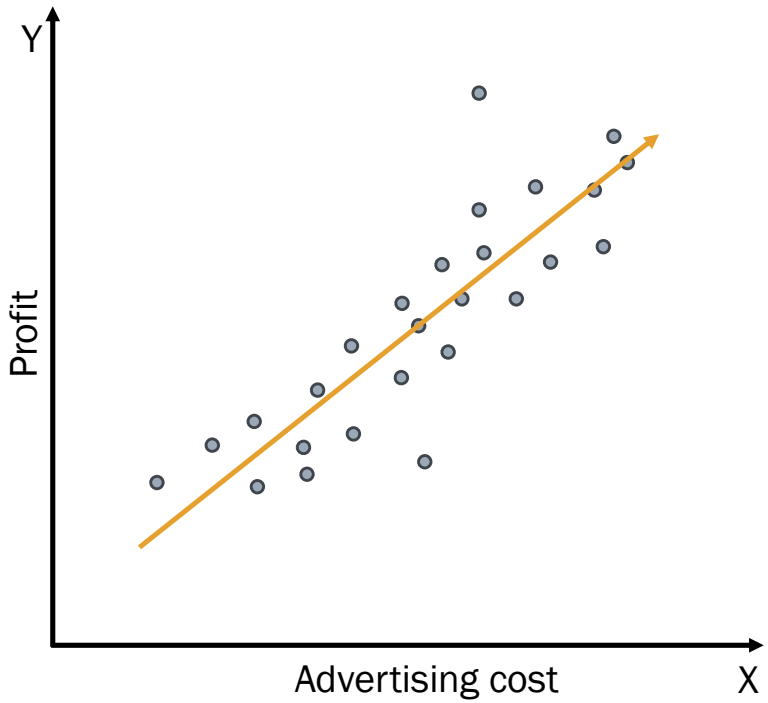


Changing Variance
(Heteroscedastic)



Key Concepts in Linear Regression

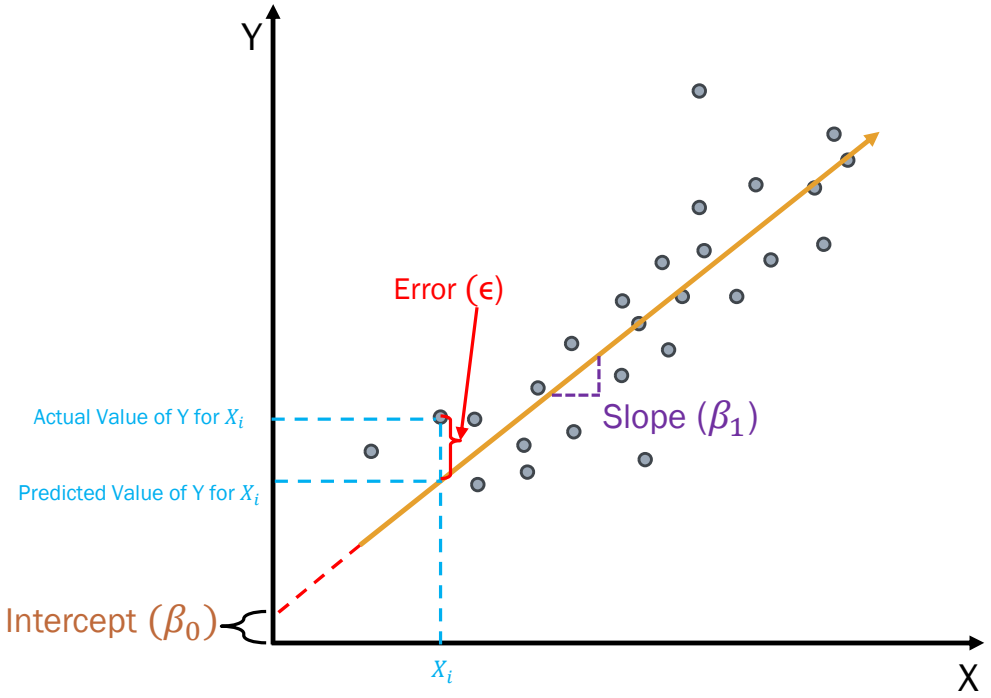
- **Independent Variable (x):** The variable(s) used to make predictions about y.
- **Dependent Variable (y):** The variable we want to predict or understand.
- **Model Assumption:** Linear relationship between x and y.



Understanding the Slope & Intercept

- Intercept (β_0): Intercept (where the line crosses the y-axis).
- Slope (β_1) : Slope (change in y per unit change in x).
- ϵ : Error term (differences between actual and predicted y values).
- The Linear Regression Equation

$$y = \beta_0 + \beta_1 x + \epsilon$$



Understanding the Slope & Intercept

- Slope (β_1) : Slope (change in y per unit change in x).

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

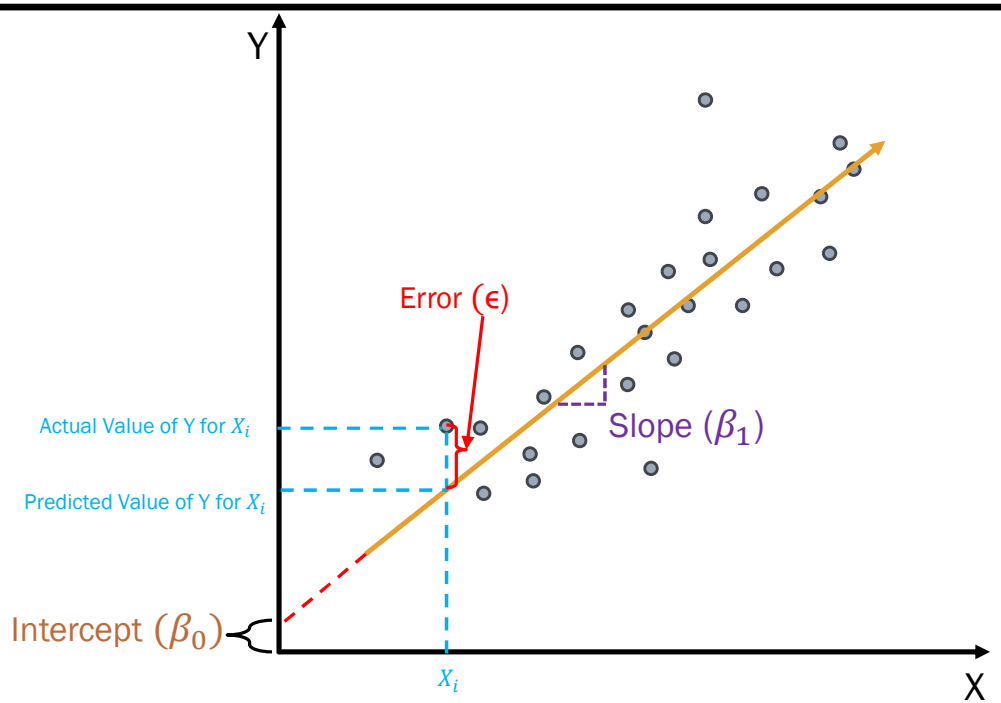
- Intercept (β_0): Intercept (where the line crosses the y-axis).

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where,

$\bar{x}$ = Mean of x
$\bar{y}$ = Mean of y

$$y = \beta_0 + \beta_1 x + \epsilon$$



Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

You have collected the following data from five students:

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

Task:

1. Fit a simple linear regression model to the data.
2. Find the coefficients β_0 (intercept) and β_1 (slope).
3. Interpret the coefficients.
4. Predict the test score for a student who studied for 3.5 hours.

Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

Solution Steps:

1. Calculate the Means:
   - Mean of x = $\bar{x}$ = $\frac{(1+2+3+4+5)}{5} = 3$
   - Mean of y = $\bar{y}$ = $\frac{(55+65+70+75+8\ )}{5} = 70$
2. Calculate β_1 (slope):

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{(1-3)(55-70) + (2-3)(65-70) + (3-3)(70-70) + (4-3)(75-70) + (5-3)(85-70)}{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}$$

$$= \frac{(-2)(-15)+(-1)(-5)+(0)(0)+(1)(5)+(2)(15)}{4+1+0+1+4}$$

$$= \frac{30+5+0+5+30}{10} = \frac{70}{10} = 7$$

Problem: You are studying the relationship between the number of hours studied and the score achieved on a test.

Solution Steps:

Hours Studied (x)	Test Score (y)
1	55
2	65
3	70
4	75
5	85

3. Calculate β_0 (intercept):

   $\beta_0 = \bar{y} - \beta_1 \bar{x} = 70 - 7 * 3 = 70 - 21 = 49$

4. Final Model:

   $y = 49 + 7x$

5. Interpretation of Coefficients:

   β_0 = 49: When no hours are studied, the predicted test score is 49.

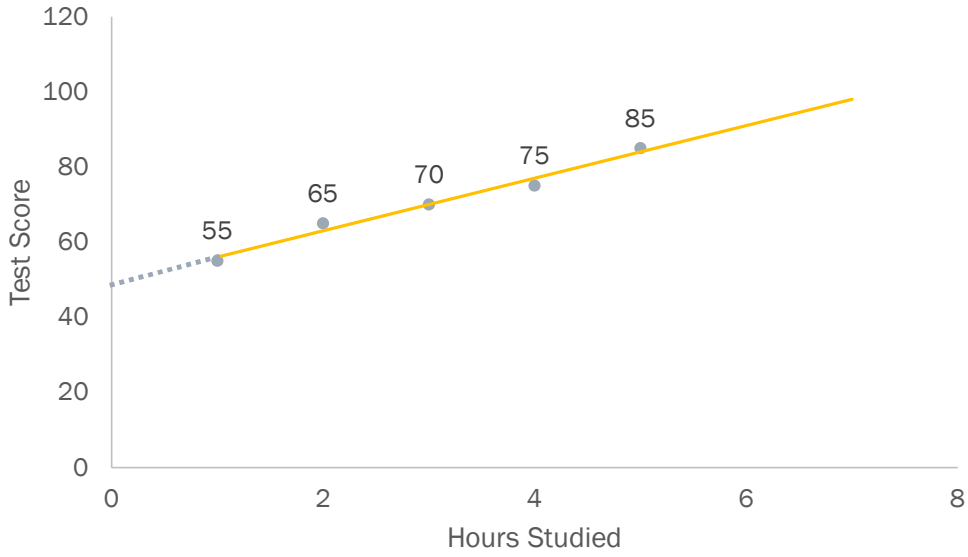
   β_1 =7: For each additional hour studied, the test score is expected to increase by 7 points.

Trendline for our Equation: $y = 49 + 7x$

Prediction for 3.5 hours:

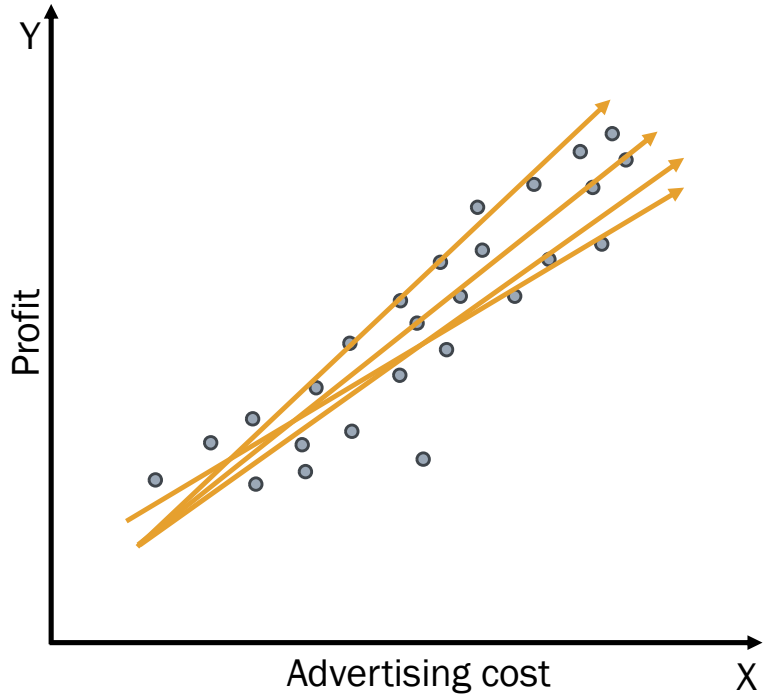
$$y = 49 + 7(3.5) = 49 + 24.5 = 73.5$$

Thus, a student who studies for **3.5 hours** is predicted to score approximately **73.5 on the test.**



The Concept of "Line of Best Fit"

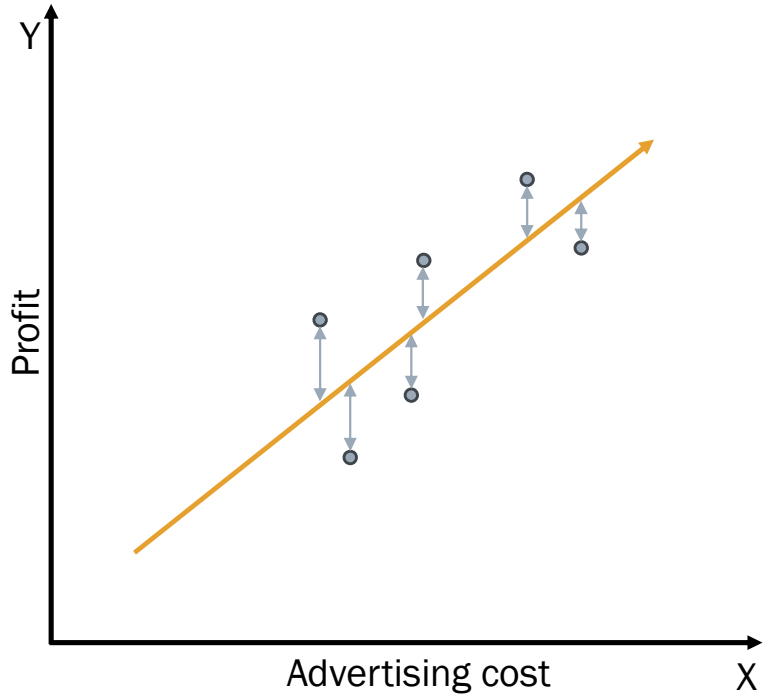
$$y = \beta_0 + \beta_1 x + \epsilon$$



The Concept of "Line of Best Fit"

- **Definition:** The line that best minimizes the difference between predicted and actual values (minimizes errors).
- **Goal:** Minimizing the sum of squared differences between actual and predicted y values.
- Residual

$$y = \beta_0 + \beta_1 x + \epsilon$$



The Concept of "Line of Best Fit"

- **Definition:** The line that best minimizes the difference between predicted and actual values (minimizes errors).
- **Goal:** Minimizing the sum of squared differences between actual and predicted y values.
- Residual Sum of Squares (RSS)

$$y = \beta_0 + \beta_1 x + \epsilon$$

The formula for the residual for each point is:

$$Residual = y_{actual} - y_{predicted}$$

The sum of squared residuals (RSS) is given by:

$$RSS = \sum (y_{actual} - y_{predicted})^2$$

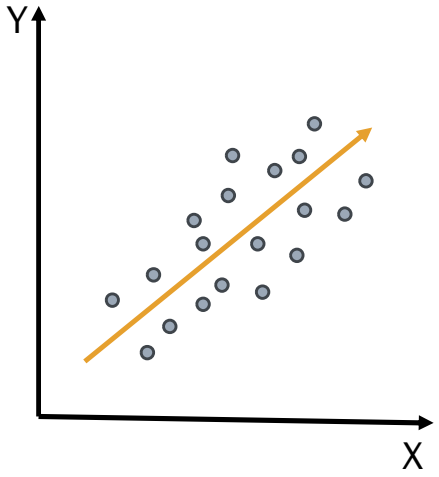
Where:
y_{actual} is the actual value.
$y_{predicted} = \beta_0 + \beta_1 x$,is the value predicted by the linear model, with β_1 as the slope and β_0 as the intercept

Evaluating Model Fit – R-squared

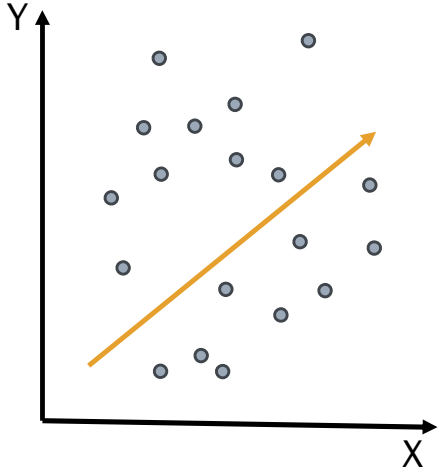
- **R-squared (R^2):** A metric to assess how well the model fits the data, representing the proportion of variance in y explained by x.
- **Values:** Ranges from 0 to 1, where closer to 1 means a better fit.

An R^2 value closer to 1 suggests a better fit.

R^2 = 0.81

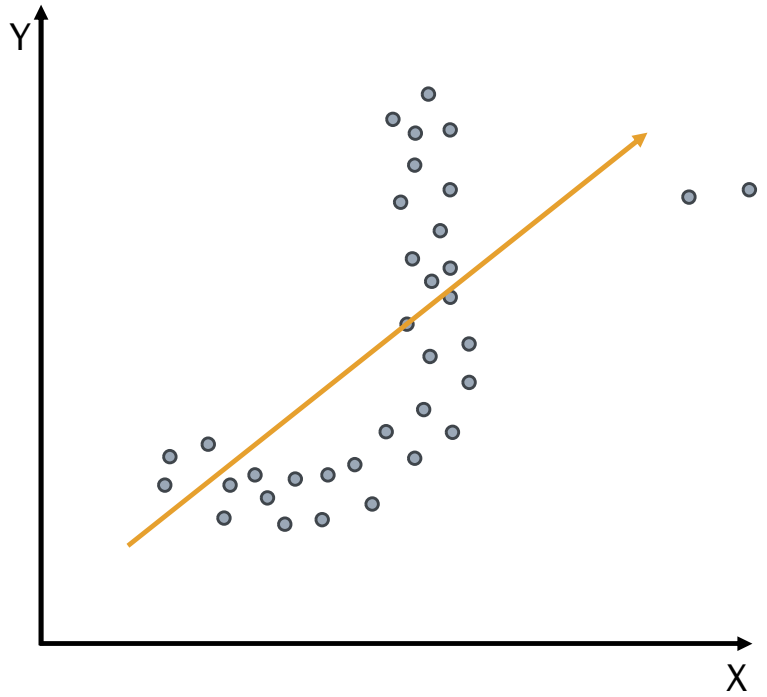


R^2 = 0.24



Limitations of Linear Regression

- **Limited to Linear Relationships:** Non-linear data requires more advanced techniques.
- **Sensitivity to Outliers:** Can skew the line.



What We Covered

- ❑ Basic understanding of Linear Regression.
- ❑ Assumptions and Example of Linear Regression.
- ❑ The linear equation, line of best fit.
- ❑ Limitations of Linear Regression.

Thank You

- @DataByteSun