Tishk International University
IT Department
Course Code: IT-344/A

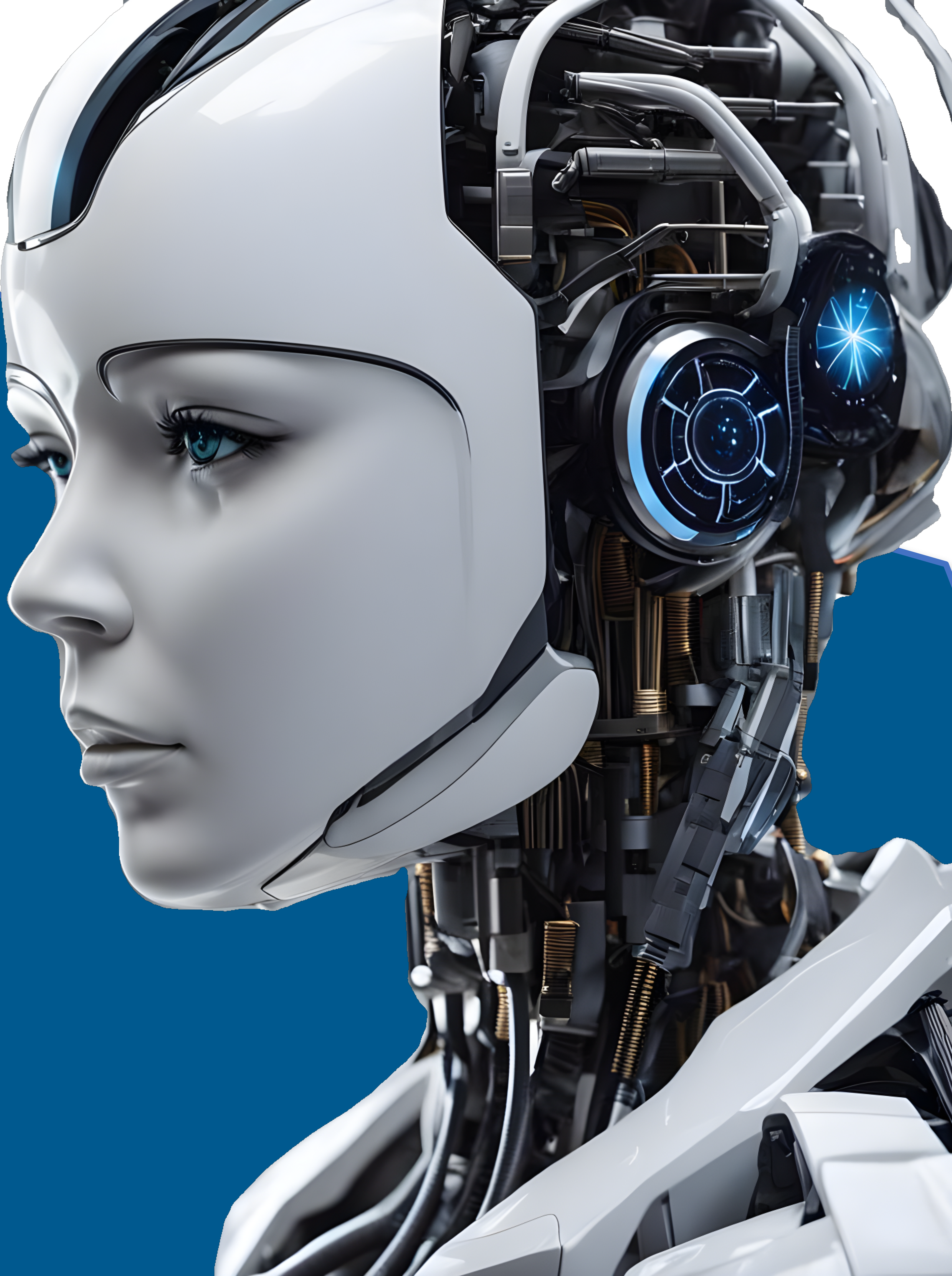
# Introduction to Machine Learning

## Regression 02

Lecture 8

Spring 2024

Hemin Ibrahim, PhD
hemin.ibrahim@tiu.edu.iq



# Outline

- Estimating the Slop Coefficient
- Evaluate metrics for regression
  - MAE
  - MSE
  - $R^2$
- Polynomial Regression

# Estimating the Slope Coefficient ($\beta_1$)

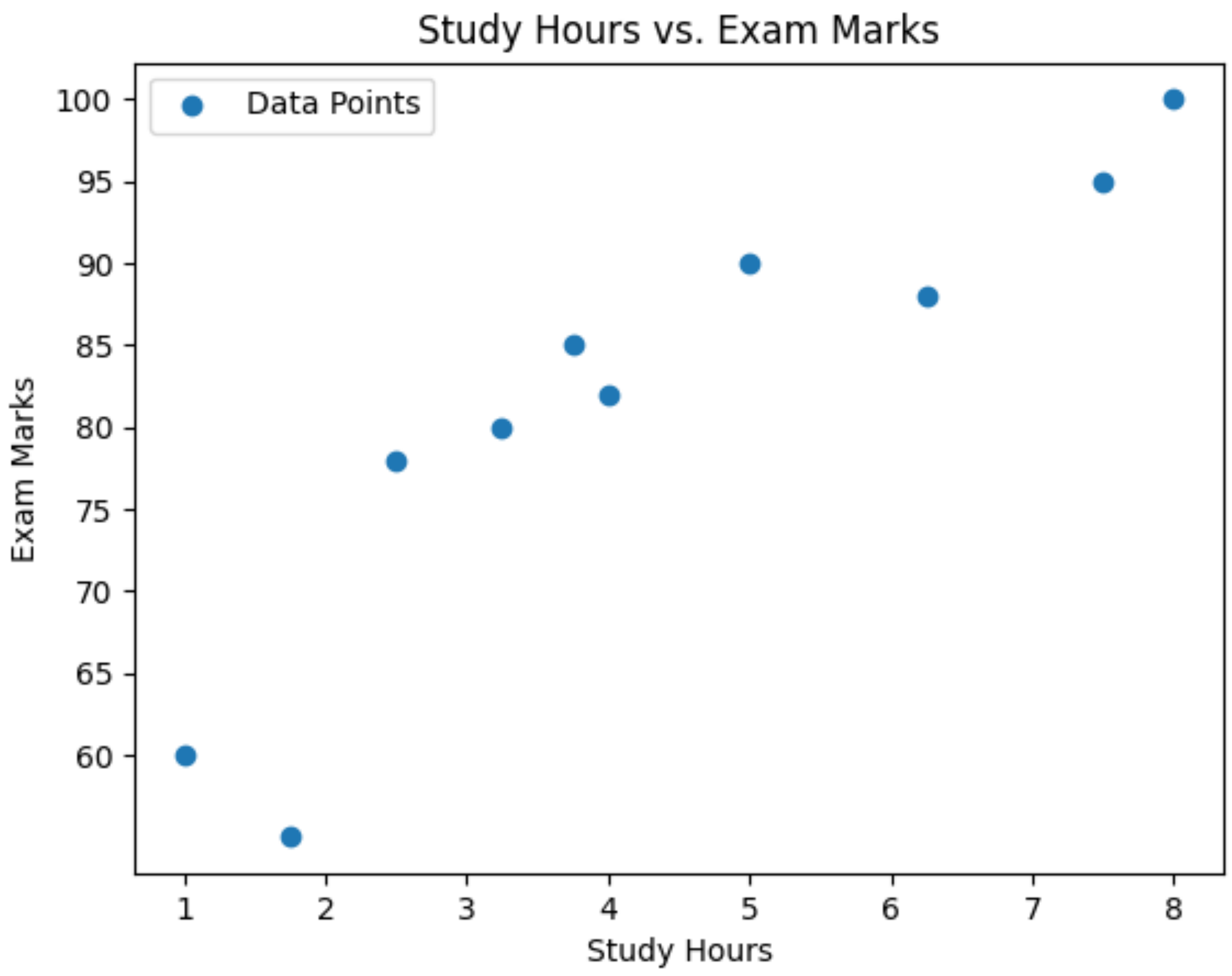
- The equation for estimating the slope coefficient $\beta_1$ is

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

- The most common method to estimate the slope coefficient ($\beta_1$) is the least squares method.
- This method minimizes the sum of the squared differences (errors) between the observed values and the values predicted by the linear regression model.

# Example

| Study Hours (x) | Exam Marks (y) |
|---|---|
| 1.00 | 60 |
| 2.50 | 78 |
| 3.75 | 85 |
| 5.00 | 90 |
| 6.25 | 88 |
| 7.50 | 95 |
| 8.00 | 100 |
| 4.00 | 82 |
| 3.25 | 80 |
| 1.75 | 55 |



# Example

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

For the Mean of $x$ =>

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1 + 2.5 + 3.75 + 5 + 6.25 + 7.5 + 8 + 4 + 3.25 + 1.75}{10} = \frac{42}{10} = 4.2$$

For the Mean of y

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{y} = \frac{60 + 78 + 85 + 90 + 88 + 95 + 100 + 82 + 80 + 55}{10} = \frac{813}{10} = 81.3$$

# Example

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

$\bar{x} = 4.2$

$\bar{y} = 81.3$

| $x$ | $y$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 1.00 | 60 | -3.20 | -21.30 | 68.16 | 10.24 |
| 2.50 | 78 | -1.70 | -3.30 | 5.61 | 2.89 |
| 3.75 | 85 | -0.45 | 3.70 | -1.665 | 0.2025 |
| 5.00 | 90 | 0.80 | 8.70 | 6.96 | 0.64 |
| 6.25 | 88 | 2.05 | 6.70 | 13.735 | 4.2025 |
| 7.50 | 95 | 3.30 | 13.70 | 45.21 | 10.89 |
| 8.00 | 100 | 3.80 | 18.70 | 71.06 | 14.44 |
| 4.00 | 82 | -0.20 | 0.70 | -0.14 | 0.04 |
| 3.25 | 80 | -0.95 | -1.30 | 1.235 | 0.9025 |
| 1.75 | 55 | -2.45 | -26.30 | 64.435 | 6.0025 |

# Example

The total sum for the numerator in the slope calculation is:

$$\sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y}) = 274.645$$

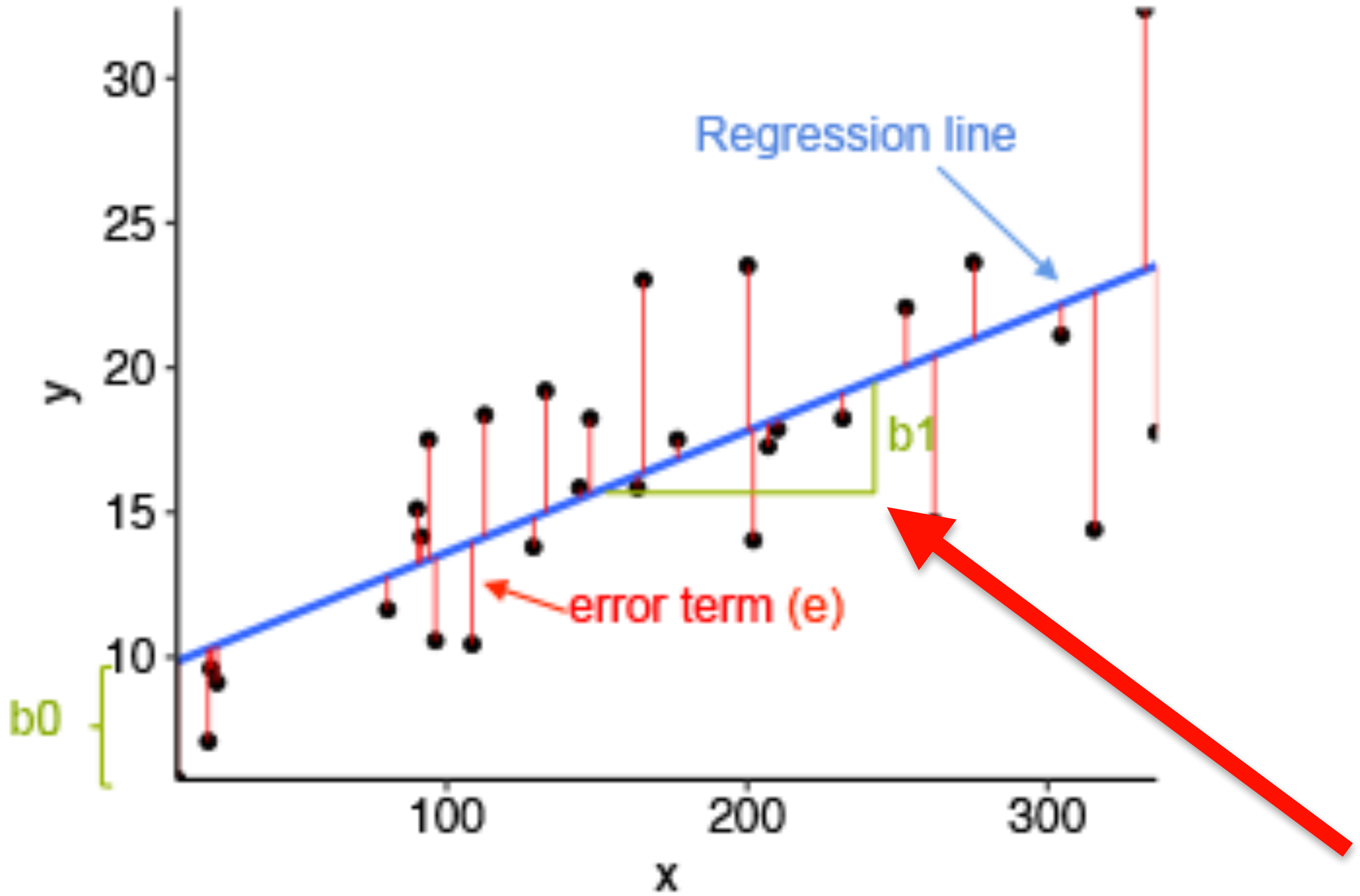
The total sum for the denominator in the slope calculation is:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 60.4475$$

The slope coefficient ( $\beta_1$ ) is calculated as:

$$\beta_1 = \frac{\sum (x_i - \bar{x}) \times (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{274.645}{60.4475} \approx 4.54$$

# Example



# Example

Calculate the Intercept $\beta_0$

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

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

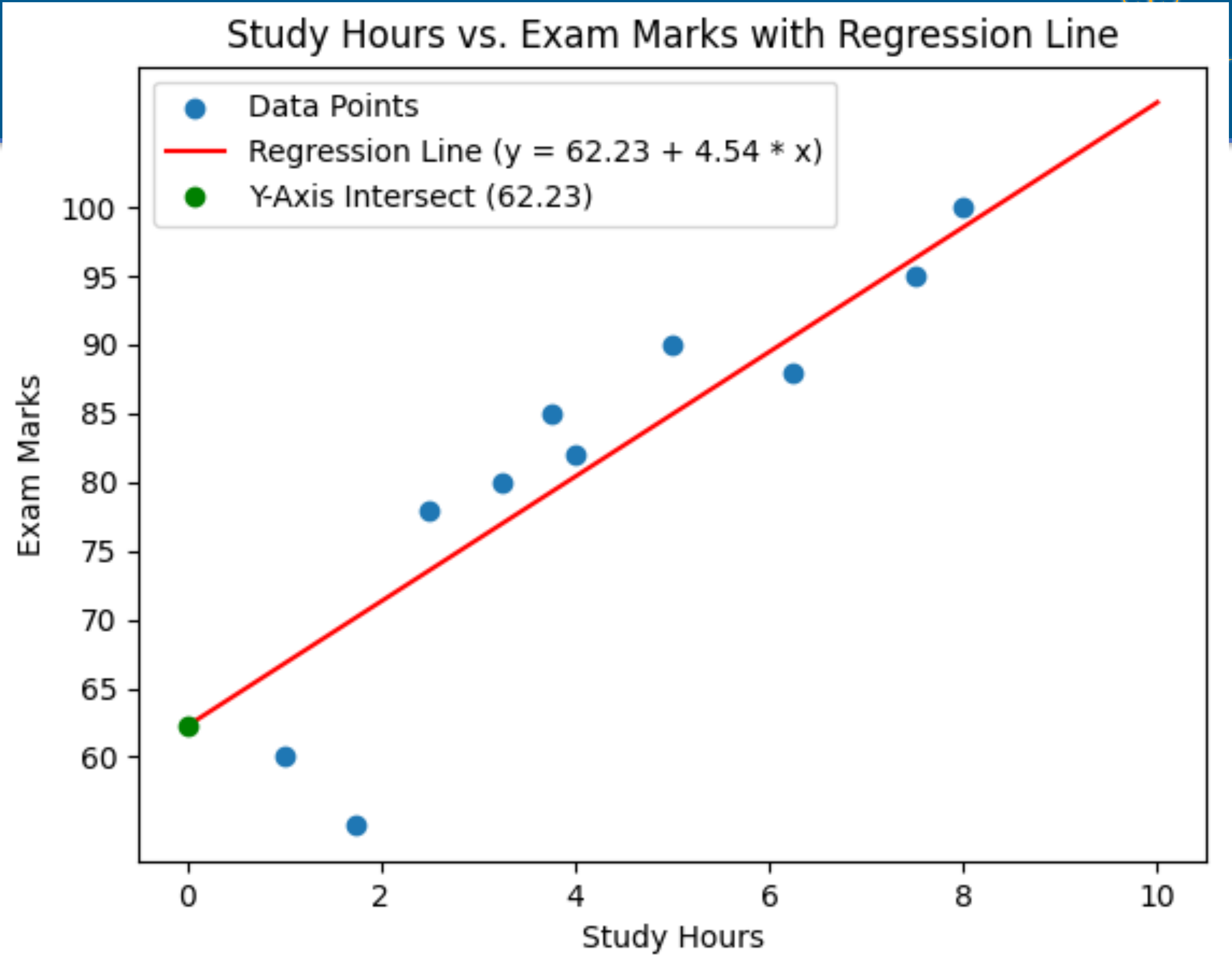
$Given\ \beta_1 = 4.54\ , and\ \bar{x} = 4.2,\ \bar{y} = 81.3\ the\ intercept\ (\beta_0)\ is\ calculated\ as:$

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

$$\beta_0 = 81.3 - 4.54 \times 4.2$$

Rounding to two decimal places, the intercept $\beta_0 = 62.23$

# Example



Study Hours vs. Exam Marks with Regression Line
Data Points
Regression Line (y = 62.23 + 4.54 * x)
Y-Axis Intersect (62.23)
Exam Marks
100
95
90
85
80
75
70
65
60
0
2
4
6
8
10
Study Hours

# Example

If a student studies for **6 hours**, what would be their expected exam mark, based on a simple linear regression model with an intercept of 62.23 and a slope of 4.54?

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

$$\beta_0 = 62.23 \ (the \ intercept)$$

$$\beta_1 = 4.54 \ (the \ slope)$$

$$x = 6 \ (hours \ studied)$$

$$y = 62.23 + (4.54 \times 6)$$

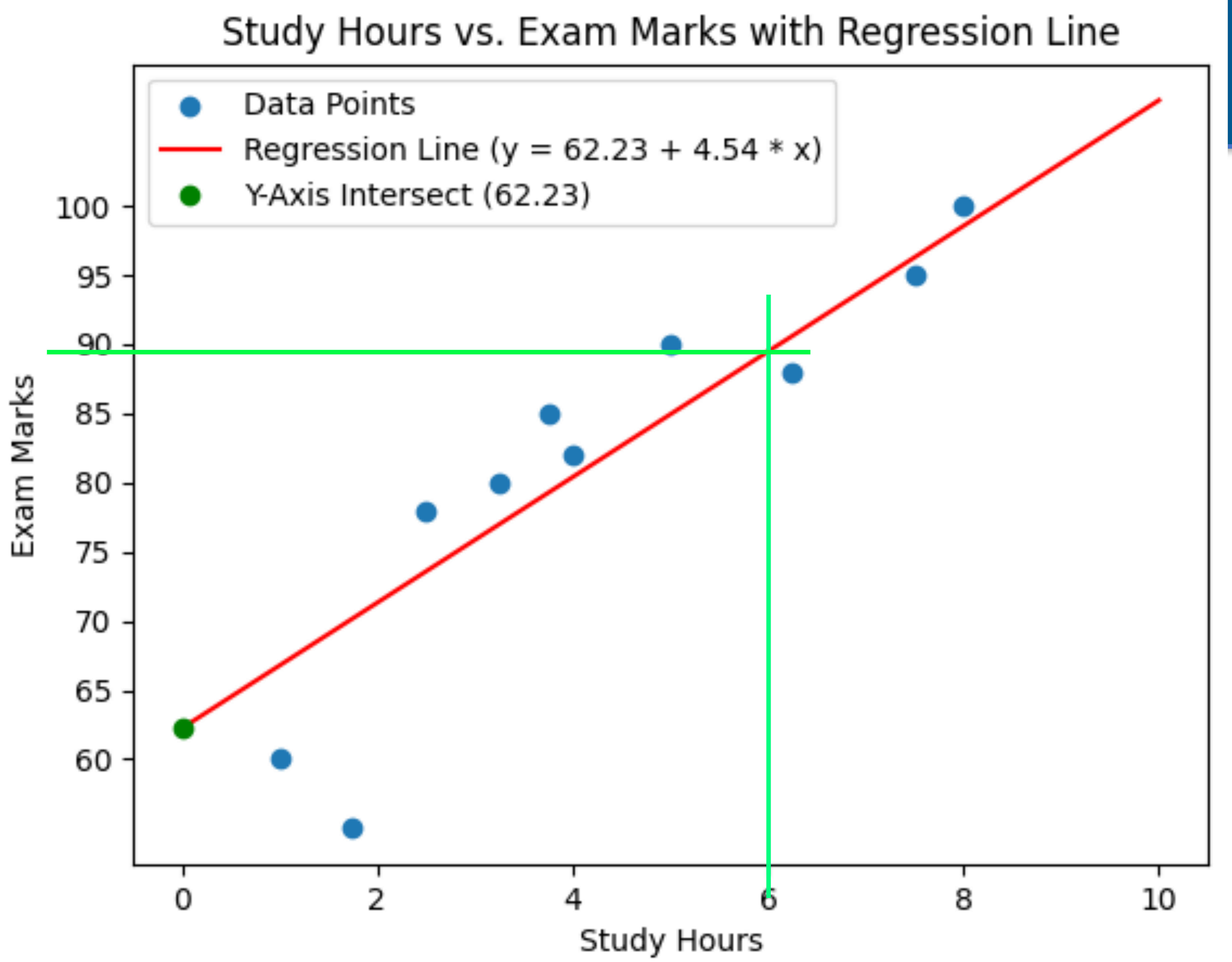
$$y = 89.47$$

Thus, the expected exam mark for a student who studies for 6 hours is approximately 89.47

# Example

$x = 6\ (hours\ studied)$

$y = 89.47$



# Example

If a student studies for **1 hour**, what would be their expected exam mark, based on a simple linear regression model with an intercept of 62.23 and a slope of 4.54?

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

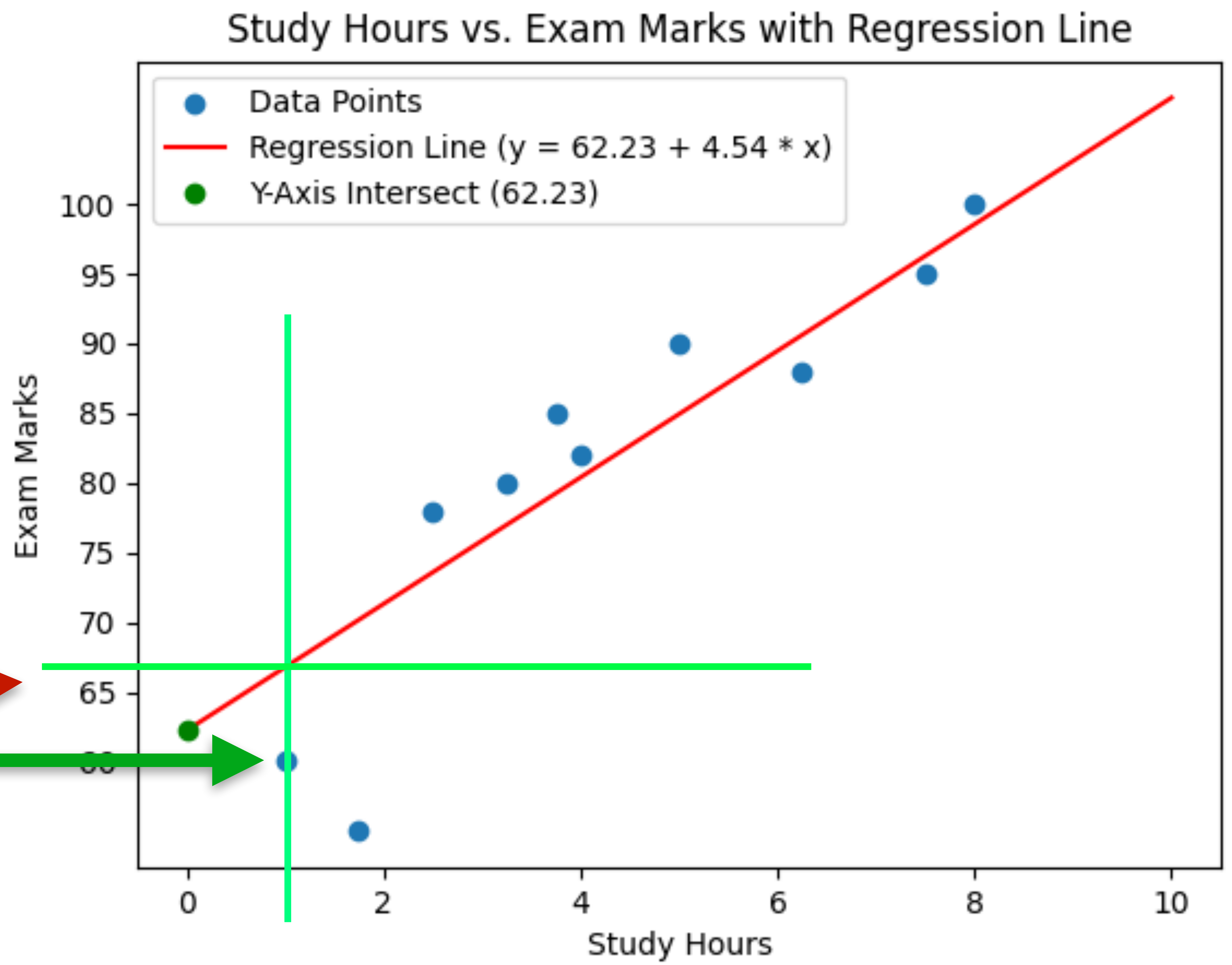
$$\beta_0 = 62.23 \ (\textit{the intercept})$$

$$\beta_1 = 4.54 \ (\textit{the slope})$$

$$x = 1 \ (\textit{hour studied})$$

$$y = 62.23 + (4.54 \times 1)$$

$$y = 66.77$$



# Evaluation Metrics for Regression

- Evaluating the performance of regression models is crucial for assessing their accuracy and effectiveness.
- Several evaluation metrics are commonly used to quantify the differences between predicted and actual values.
- Here are some key evaluation metrics for regression:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Coefficient of Determination ($R^2$)

# Mean Absolute Error (MAE)

- MAE measures the average absolute difference between predicted and actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\text{MAE} = \frac{1}{10}\sum_{i=1}^{10} \left|y_i - \hat{y}_i\right|$$

$$\text{MAE} = 4.854$$

| $x$ | $y$ | $\hat{y}$ | $\lvert y - \hat{y}\rvert$ |
|---|---|---|---|
| 1.00 | 60 | 66.77 | 6.77 |
| 2.50 | 78 | 73.08 | 4.92 |
| 3.75 | 85 | 78.97 | 6.03 |
| 5.00 | 90 | 84.77 | 5.23 |
| 6.25 | 88 | 90.57 | 2.57 |
| 7.50 | 95 | 96.37 | 1.37 |
| 8.00 | 100 | 98.89 | 1.11 |
| 4.00 | 82 | 80.39 | 1.61 |
| 3.25 | 80 | 76.61 | 3.39 |
| 1.75 | 55 | 70.54 | 15.54 |

# Mean Squared Error (MSE)

- To compute MSE, square the differences between actual and predicted values

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{MSE} = \frac{1}{10}\sum_{i=1}^{10}(y_i - \hat{y}_i)^2$$

$$\text{MSE} = 3.998$$

| $x$ | $y$ | $\hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|
| 1.00 | 60 | 66.77 | 45.92 |
| 2.50 | 78 | 73.08 | 24.18 |
| 3.75 | 85 | 78.97 | 36.36 |
| 5.00 | 90 | 84.77 | 27.34 |
| 6.25 | 88 | 90.57 | 6.60 |
| 7.50 | 95 | 96.37 | 1.88 |
| 8.00 | 100 | 98.89 | 1.23 |
| 4.00 | 82 | 80.39 | 2.60 |
| 3.25 | 80 | 76.61 | 11.46 |
| 1.75 | 55 | 70.54 | 242.26 |

# Coefficient of Determination ( $R^2$)

- To compute $R^2$, find the explained variance (squared differences between predicted and actual values) and the total variance (squared differences between actual values and the mean), then compute the ratio:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

# Coefficient of Determination ( $R^2$)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{405.47}{1826.11} \approx 0.778$$

| $x$ | $y$ | $\hat{y}$ | $(y-\hat{y})^2$ | $(y-\bar{y})^2$ |
|---|---|---|---|---|
| 1.00 | 60 | 66.77 | 45.92 | 454.89 |
| 2.50 | 78 | 73.08 | 24.18 | 10.89 |
| 3.75 | 85 | 78.97 | 36.36 | 13.69 |
| 5.00 | 90 | 84.77 | 27.34 | 75.69 |
| 6.25 | 88 | 90.57 | 6.60 | 44.89 |
| 7.50 | 95 | 96.37 | 1.88 | 187.69 |
| 8.00 | 100 | 98.89 | 1.23 | 348.49 |
| 4.00 | 82 | 80.39 | 2.60 | 0.49 |
| 3.25 | 80 | 76.61 | 11.46 | 1.69 |
| 1.75 | 55 | 70.54 | 242.26 | 687.69 |

# What does it mean?

$$MAE = 4.85$$

$$MSE = 3.998$$

$$R^2 = 0.778$$

# What does it mean?

$$MAE = 4.85 \quad MSE = 3.998 \quad R^2 = 0.778$$

Mean Absolute Error (MAE) = 4.85

- The MAE value of 4.85 indicates that, on average, the predicted values deviate from the actual values by approximately 4.85 units (in the same scale as the target variable).
- A lower MAE value is desirable, as it means the predictions are closer to the actual values.

# What does it mean?

$$MAE = 4.85 \quad MSE = 3.998 \quad R^2 = 0.778$$

Mean Squared Error (MSE) = 3.998

- The MSE value of 3.998 represents the average squared difference between the predicted and actual values.
- <u>A lower MSE value is preferred</u>, as it indicates smaller differences between predictions and actual values.

# What does it mean?

$$MAE = 4.85 \quad MSE = 3.998 \quad R^2 = 0.778$$

Coefficient of Determination ($R^2$) = 0.778

- The $R^2$ value of 0.778 means that approximately 77.8% of the variation in the target variable is explained by the linear regression model.
- $R^2$ ranges from 0 to 1, with higher values indicating a better fit of the model to the data.
- An $R^2$ value of 0.778 suggests a reasonably good fit, but there is still room for improvement in the model's predictive ability.

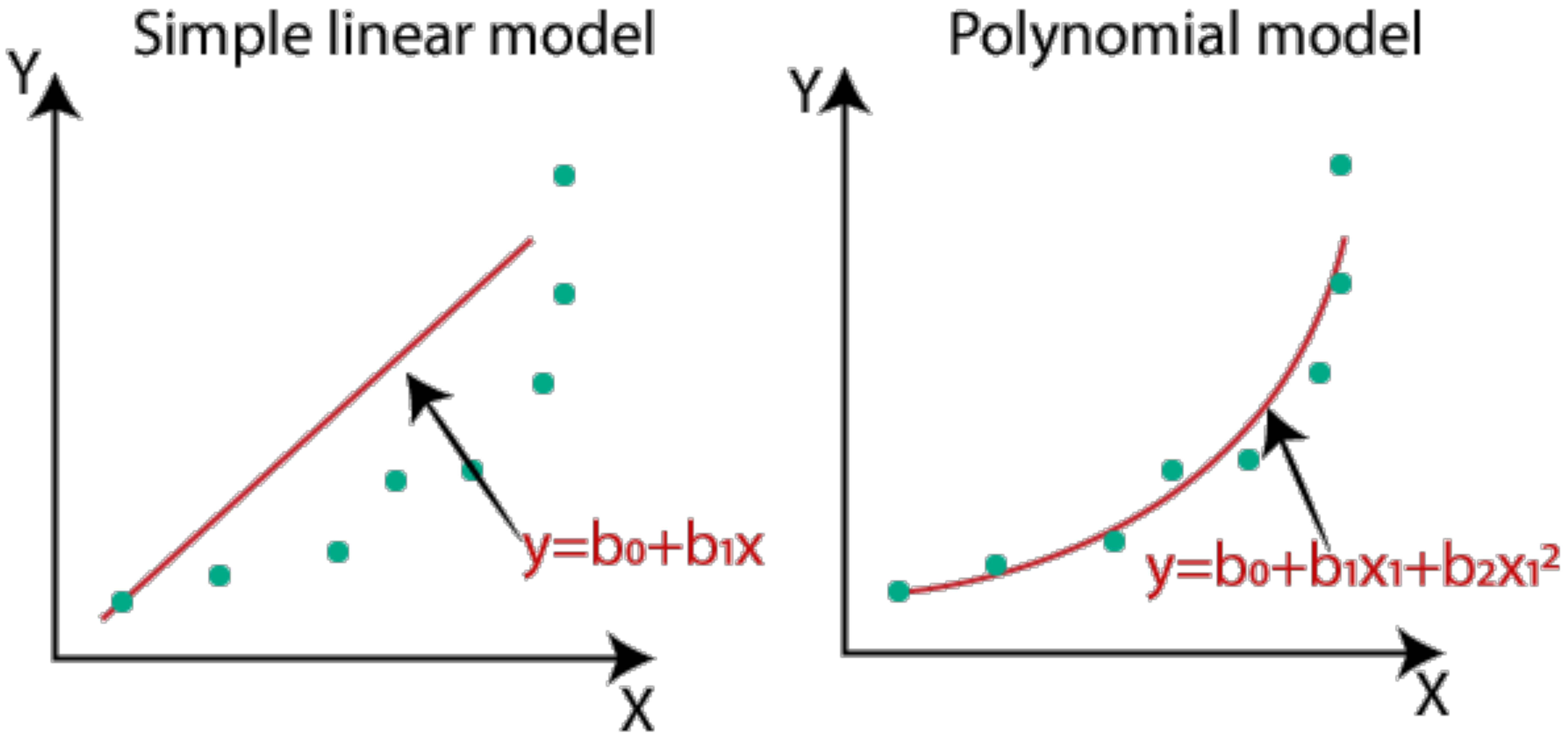
# Polynomial Regression

Polynomial Regression is a type of regression analysis where the relationship between the independent (x) variable(s) and the dependent variable (y) is modeled as an nth-degree polynomial.

It is an extension of linear regression, allowing for more complex relationships and providing a flexible way to fit nonlinear data.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_n X^n + \epsilon$$

# Polynomial Regression



# Polynomial Regression

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_n X^n + \epsilon$$

Where

- y: Dependent variable (what we're trying to predict)
- x: Independent variable (what we're basing our prediction on)
- $\beta_0$ (y-intercept): This is the constant term that indicates the value of $y$ y when x=0. It represents the starting point of the curve on the y-axis.
- $\beta_1$ to $\beta_n$: Coefficients to be estimated from the data
- n: Degree of the polynomial (how many times x is multiplied by itself)

# Polynomial Regression

**Advantages**:

**Flexibility**: Polynomial regression can model a wide range of non-linear relationships.

**Interpretability**: While more complex than linear regression, however, instead of fitting a straight line to the data, polynomial regression fits a curved line.

**Disadvantages:**

**Overfitting**: Higher-degree polynomials can fit the training data too closely, leading to poor generalization to new data.

**Numerical Instability**: Using high-degree polynomials can make things unstable, causing the curve to wobble a lot or create very large values.

Thank You

TISHK INTERNATIONAL UNIVERSITY
TIU
ERBIL 2008