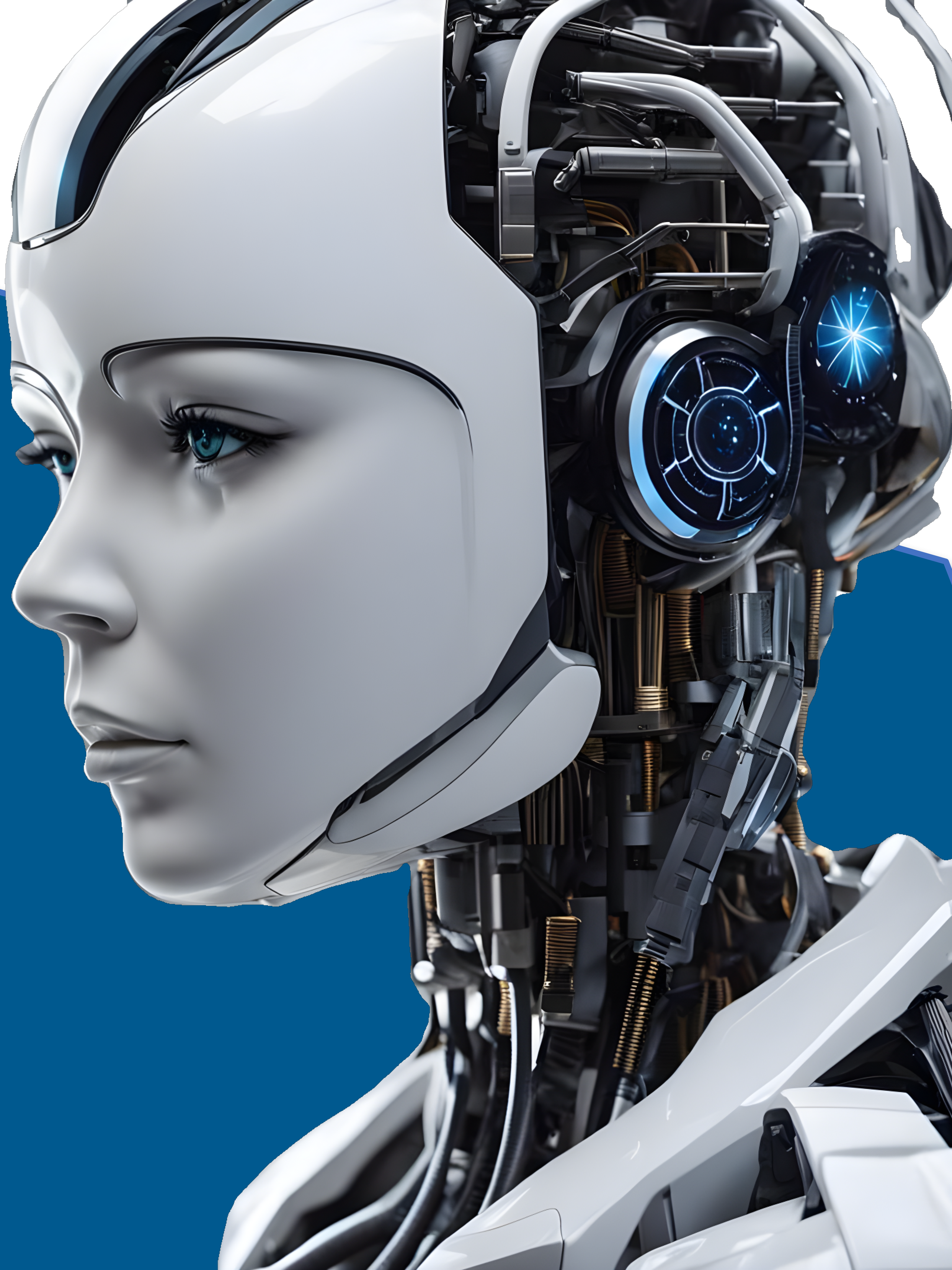
Tishk International University
IT Department
Course Code: IT-344/A

Introduction to Machine Learning

Regression

Lecture 7



Spring 2024

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

Outline

- Regression
- Applications of Regression
- Case study
- Linear Regression
- Linear function
- Linear Regression Equation

Objectives

- Understand the concept of regression analysis, a statistical method used to study the relationship between a dependent variable and one or more independent variables, with a focus on prediction and modeling.
- Explore various real-world scenarios and industries where regression analysis is applied.
- Define and understand a linear function in the context of mathematics and statistics.
- Derive and analyze the linear regression equation, which models the relationship between the dependent variable and one or more independent variables.

Regression

- It is a type of supervised learning task in machine learning and statistics where the goal is to develop a predictive model that maps an input variable or set of input variables to a continuous output value.
- Unlike classification, where the output is categorical, regression is concerned with predicting numerical or continuous outcomes.
- Find a relationship between independent variables (features) and a dependent variable (target).
  - **Example**: Predicting a house's sale price (target) based on its size, location, and number of bedrooms (features).

Regression (Cont.)

- Accurate prediction for new, unseen data.
  - **Example**: Training a model on historical housing data and using it to estimate the price of a newly listed home.
- Various techniques can be used to develop regression models.
  - **Examples**: Linear regression, decision trees, random forests, neural networks, etc.

Regression vs Classification

Example

- A company may track how much they spend on various digital marketing channels (social media ads) and revenue numbers from those efforts.

$f(socialads)$

$f(1000) = 18000$

$f(1800) = 32500$

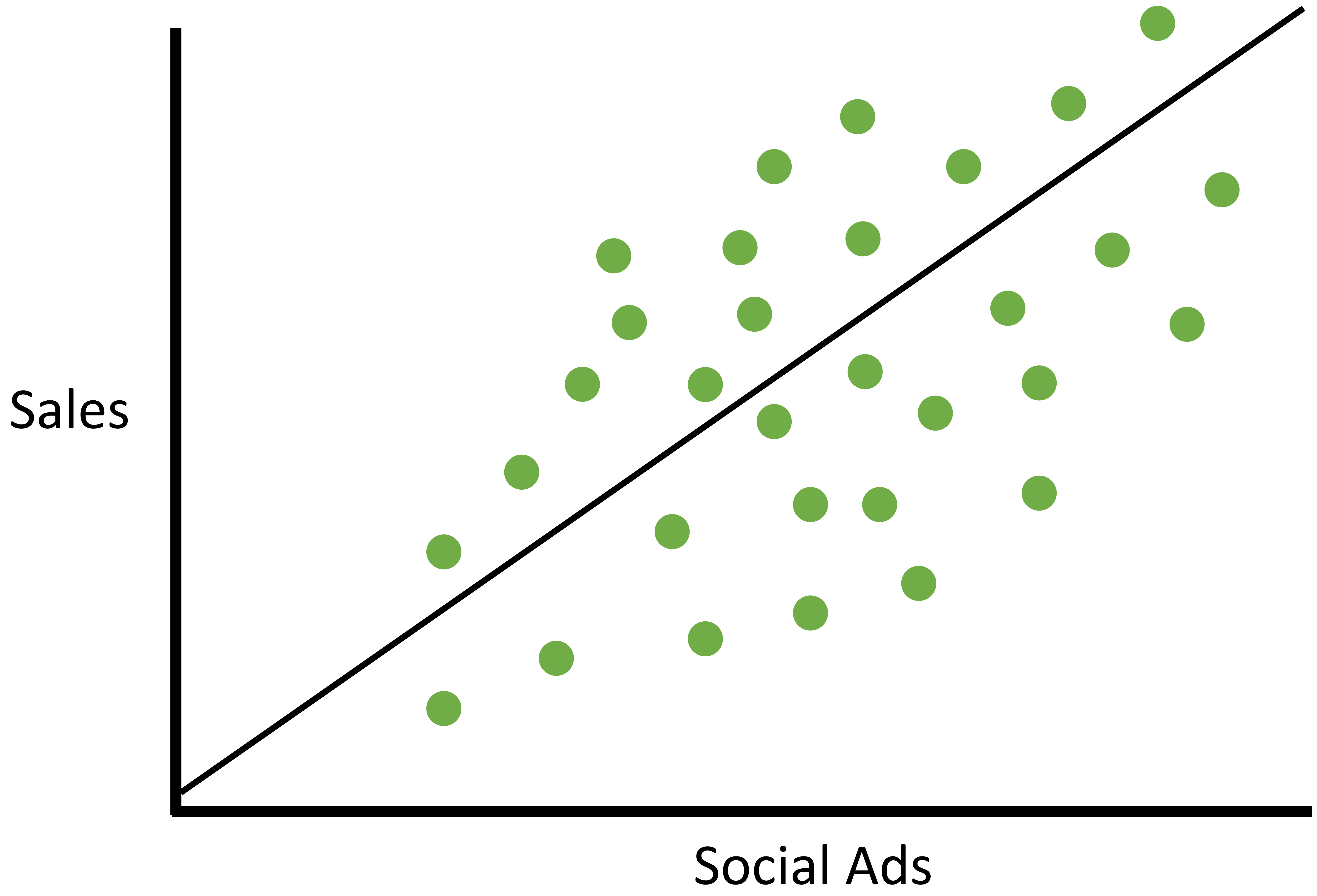
$f(2900) = 44000$

$f(x)$ represents a function mapping social media advertising spend x to revenue returns

$h(1400) = ?$

This data could then be used to build regression models to understand the relationship between marketing spend and business metrics

Example



Applications of Regression

House Prices Prediction

- Regression analysis in real estate is used to determine the value of a property based on factors such as <u>size</u>, <u>location</u>, and <u>features</u>.
- Linear regression is often employed to <u>estimate a property's price based on historical sales data</u> and <u>property characteristics</u>.

- Square_Meter: [100, 150, 200, 250]
- Number_of_Bedrooms: [2, 3, 4, 5]
- Age_of_the_house: [2,10,7,9]
- Crime_rates: [33,10,15,4]
- Price: [60000, 100000, 150000, 200000]

Applications of Regression

Health Risk Assessment

- In healthcare, regression models are used to predict health outcomes and assess risk factors. This information is critical for preventive care and designing treatment plans.
- A healthcare provider might use regression to analyze the relationship between patient characteristics (age, weight, blood pressure, cholesterol levels, etc.) and health outcomes like the risk of heart disease.

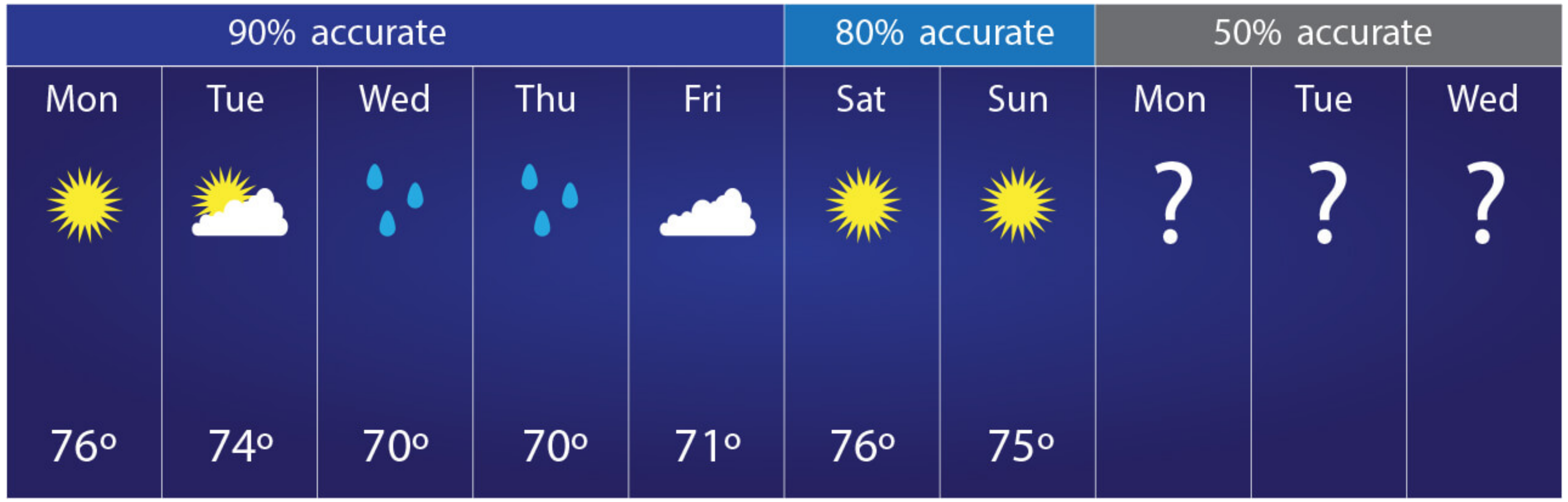
- Age: [25, 35, 45, 55]
- Cholesterol_Level: [180, 200, 220, 250]
- Risk_of_Heart_Disease: [0.1, 0.15, 0.25, 0.35]

Applications of Regression

Weather Prediction

- Weather forecasting relies on complex regression models to predict future weath conditions. These models analyze a variety of meteorological data points to mak accurate predictions.

- Temperature: [15, 20, 25, 30]
- Humidity: [50, 60, 70, 80]
- Chance_of_Rain: [0.1, 0.2, 0.3, 0.4]



Case study

A large technology company wants to make sure it's offering competitive salaries to its IT Specialists. To help with this, you, a data scientist, are tasked with creating a model that predicts IT Specialist salaries based on several factors.

- Which information (features) do you need?
- What features are most predictive of IT Specialist salaries?
- What types of bias might be present in this dataset, and how can you reduce them?

Case study

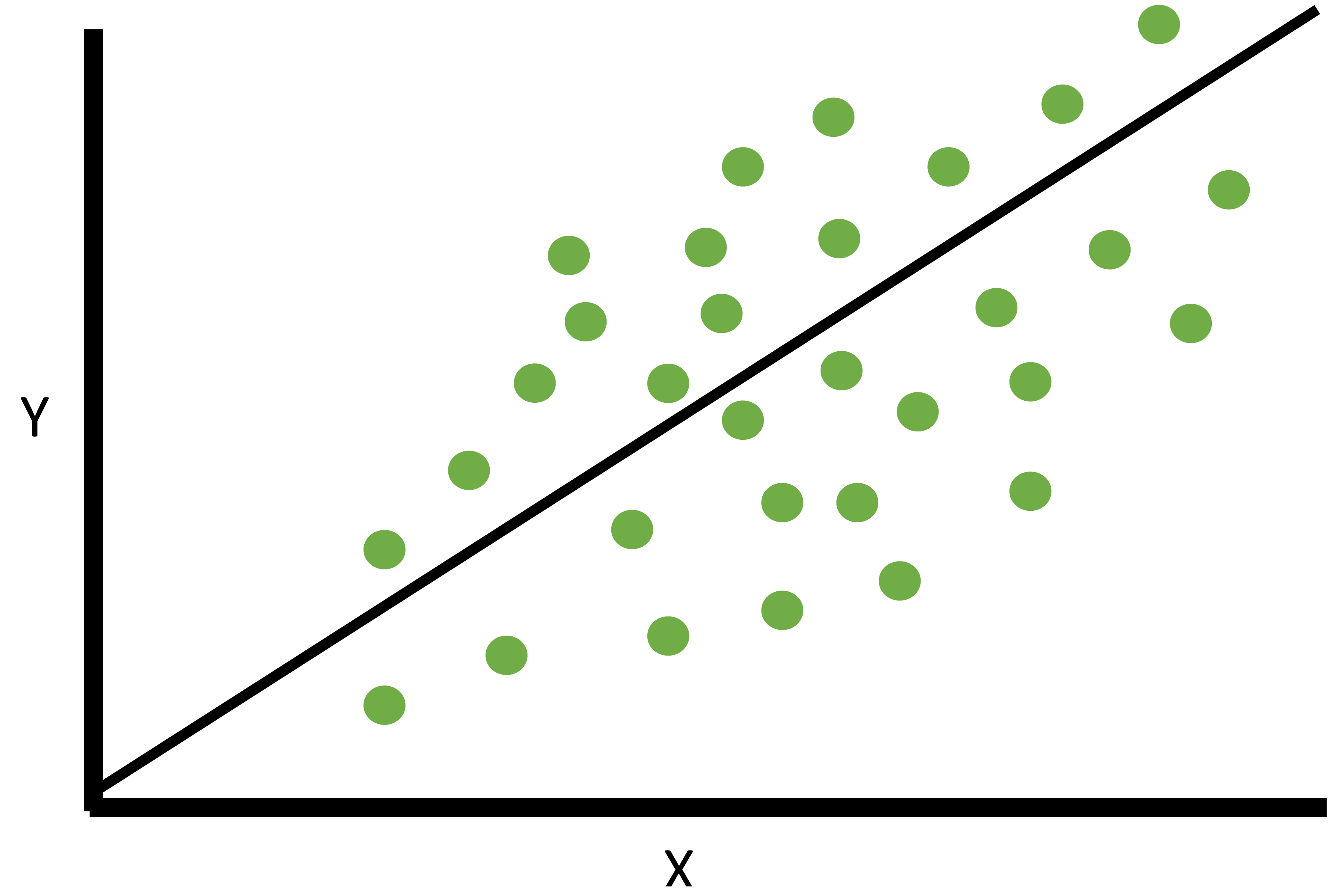
- **Features**:

Salaries, Years of experience, Education levels, Specializations, Locations, Company sizes, Certifications, Programming language skills.

- Explore the dataset and understand the various features and their potential impact on salaries.
- Preprocess the data, handle missing values, and encode categorical variables appropriately.

Linear Regression

- Linear regression is a technique to estimate an outcome (Y) based on one or more input factors (X). It's useful when you need to predict a continuous value.
  - **For example**: estimating someone's salary based on their years of experience.
- It models the relationship by fitting a linear equation to the observed data.
- The linear equation describes how the dependent variable changes as the independent variables change.



Linear Functions

A linear function describes a relationship between two variables in a straight line. It is typically written as

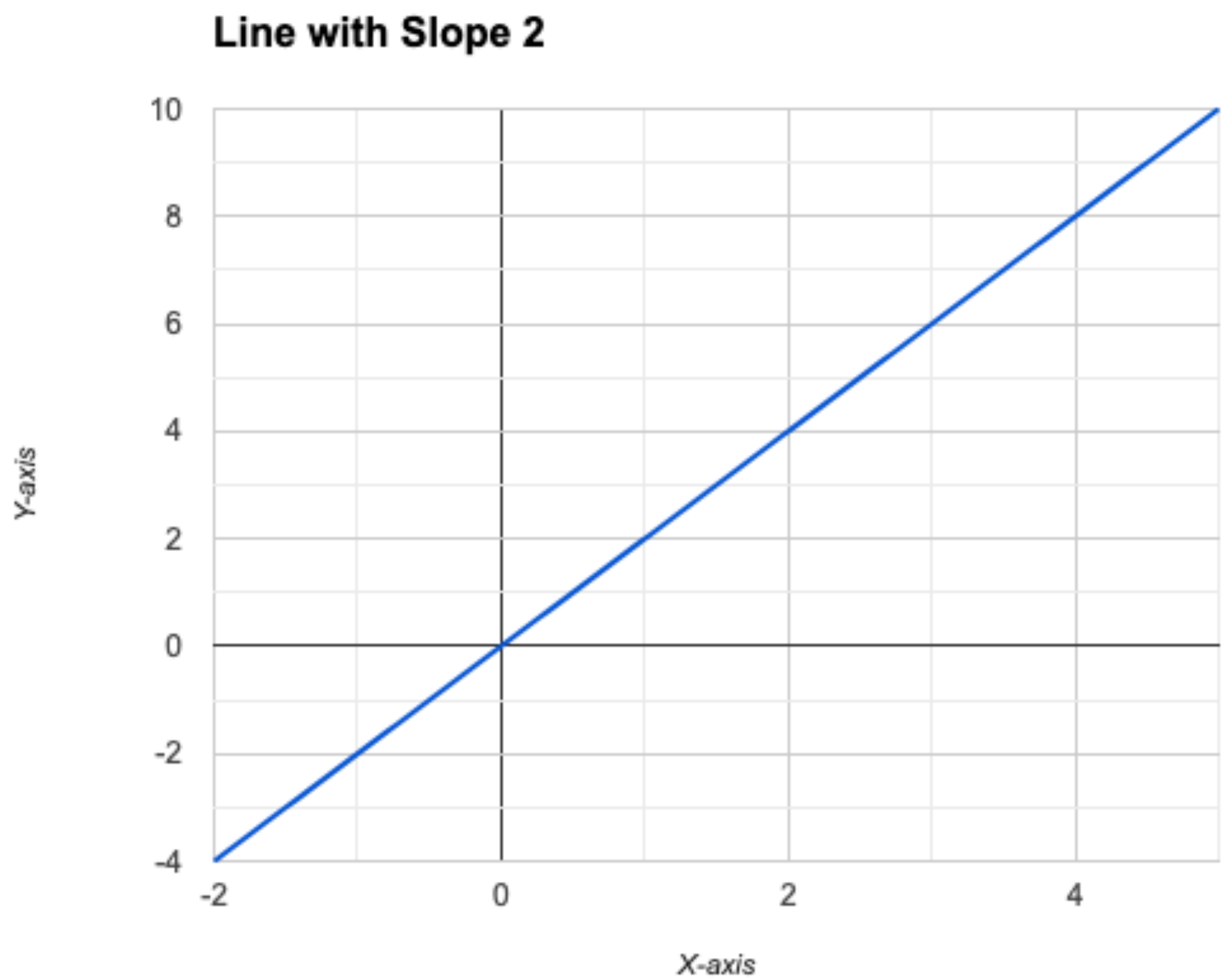
$$y = mx + b$$

where:

- y is the dependent variable.
- x is the independent variable.
- m is the slope of the line, indicating how much y changes for a one-unit change in x.
- b is the y-intercept, the value of y when x = 0.

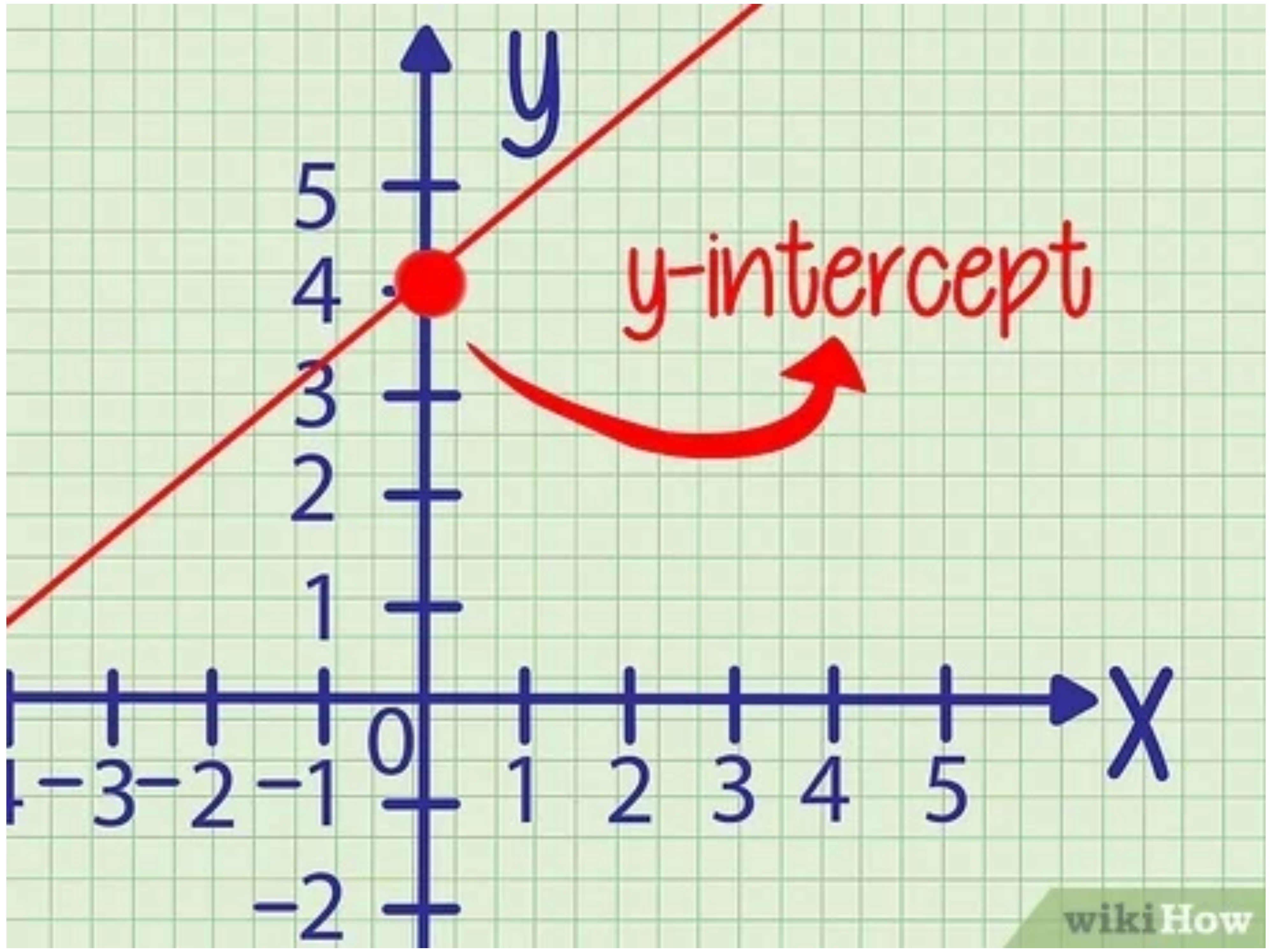
Understanding the Slope

- The slope (m) determines the angle and direction of the line.
- A positive slope means the line goes upward as x increases.
- A negative slope means the line goes downward as x increases.
- An example with a slope of 2: For every unit increase in x, y increases by 2.
- A horizontal line has a slope of 0.



The Y-Intercept

- The y-intercept (b) is the point where the line crosses the y-axis.
- This value represents the starting point of the linear function when x = 0.
- Example: If b = 4, the line starts at y = 4 when x = 0.

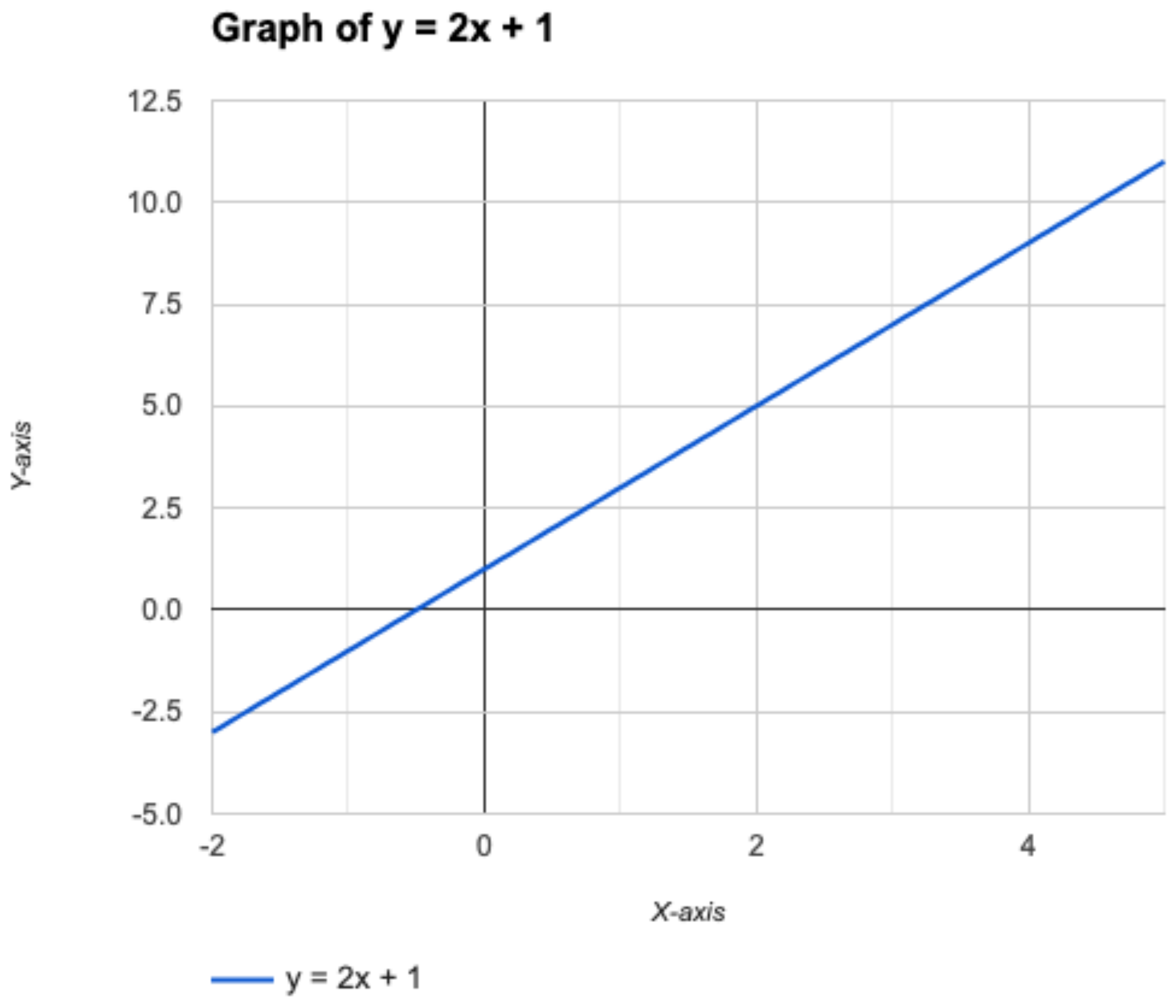


How to Graph a Linear Function?

- Identify the y-intercept (b).
- From the y-intercept, use the slope (m) to determine other points on the line.
- Draw a straight line through the points.

- **Example**: Graph the line $y = 2x + 1$:
  - Start at y = 1 (y-intercept).
  - Use the slope to find the next point: Up 2 units and right 1 unit.
  - Draw a line through these points.

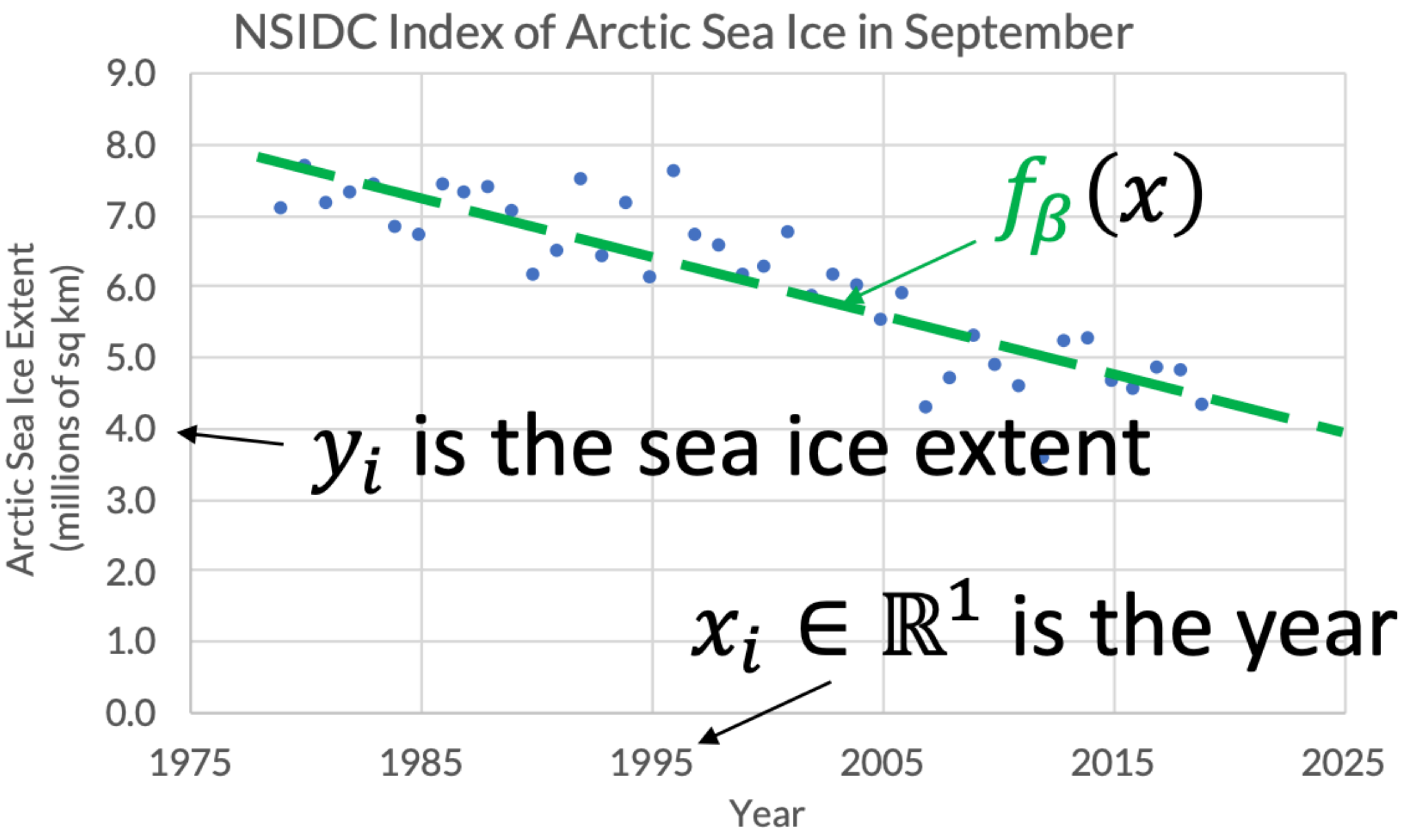
How to Graph a Linear Function?



Linear Regression Equation

- The goal is to determine the slope (m) and y-intercept (b) that minimize the error between the line and the data points.
- We don't know the slope and y-intercept beforehand, and we need to estimate them from the data.
- The general formula for a simple linear regression is $y = \beta_0 + \beta_1 x$, where:
  - y is the predicted outcome.
  - x is the independent variable or predictor.
  - β 1 is the slope, indicating how much y changes with a unit increase in x.
  - β 0 is the y-intercept, indicating the starting value of y when x = 0.

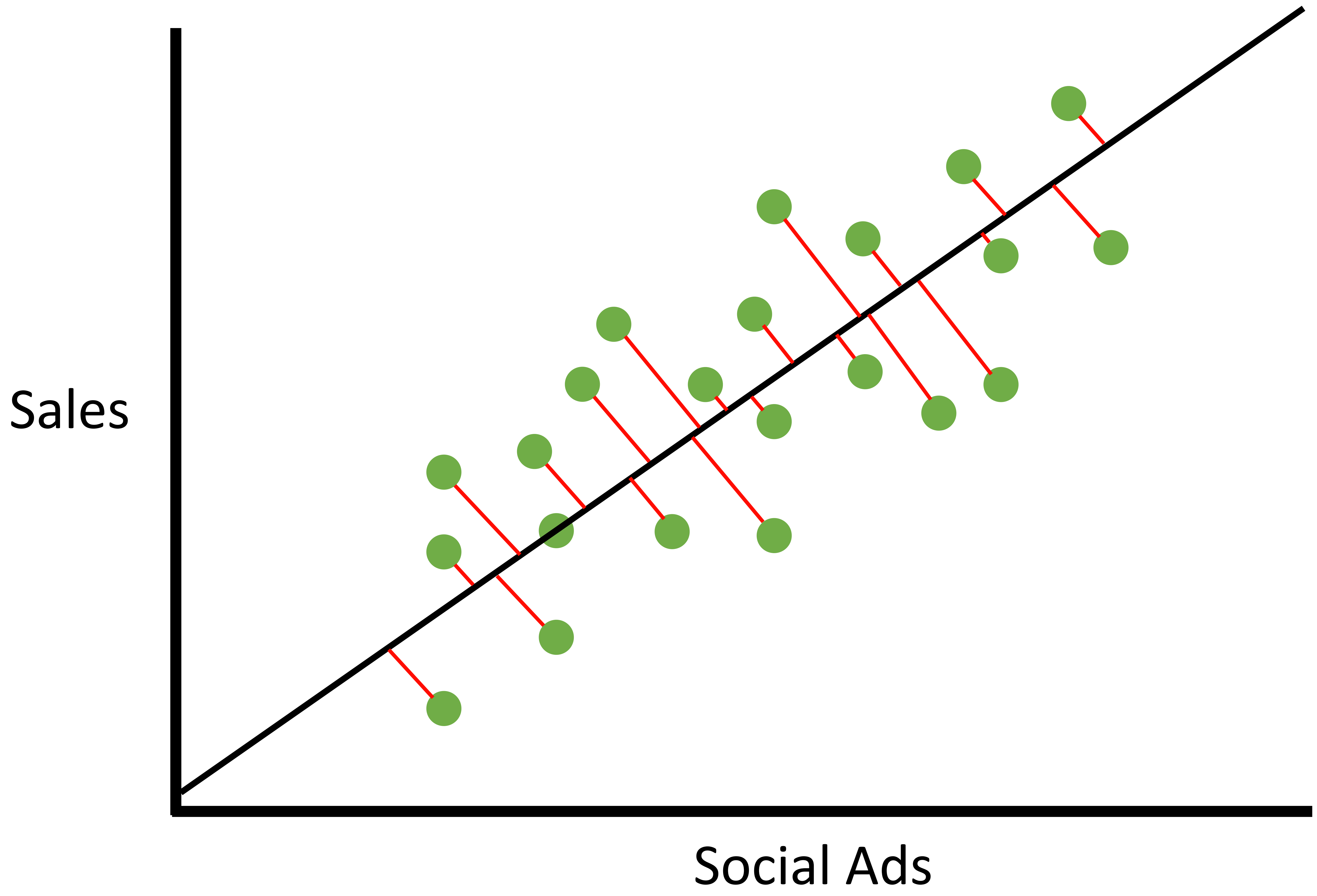
Example



A linear function $f_\beta(x) = \beta^\top x$ such that $y_i \approx \beta^\top x_i$

https://nsidc.org/arcticseaicenews/sea-ice-tools/

Error in Linear Regression

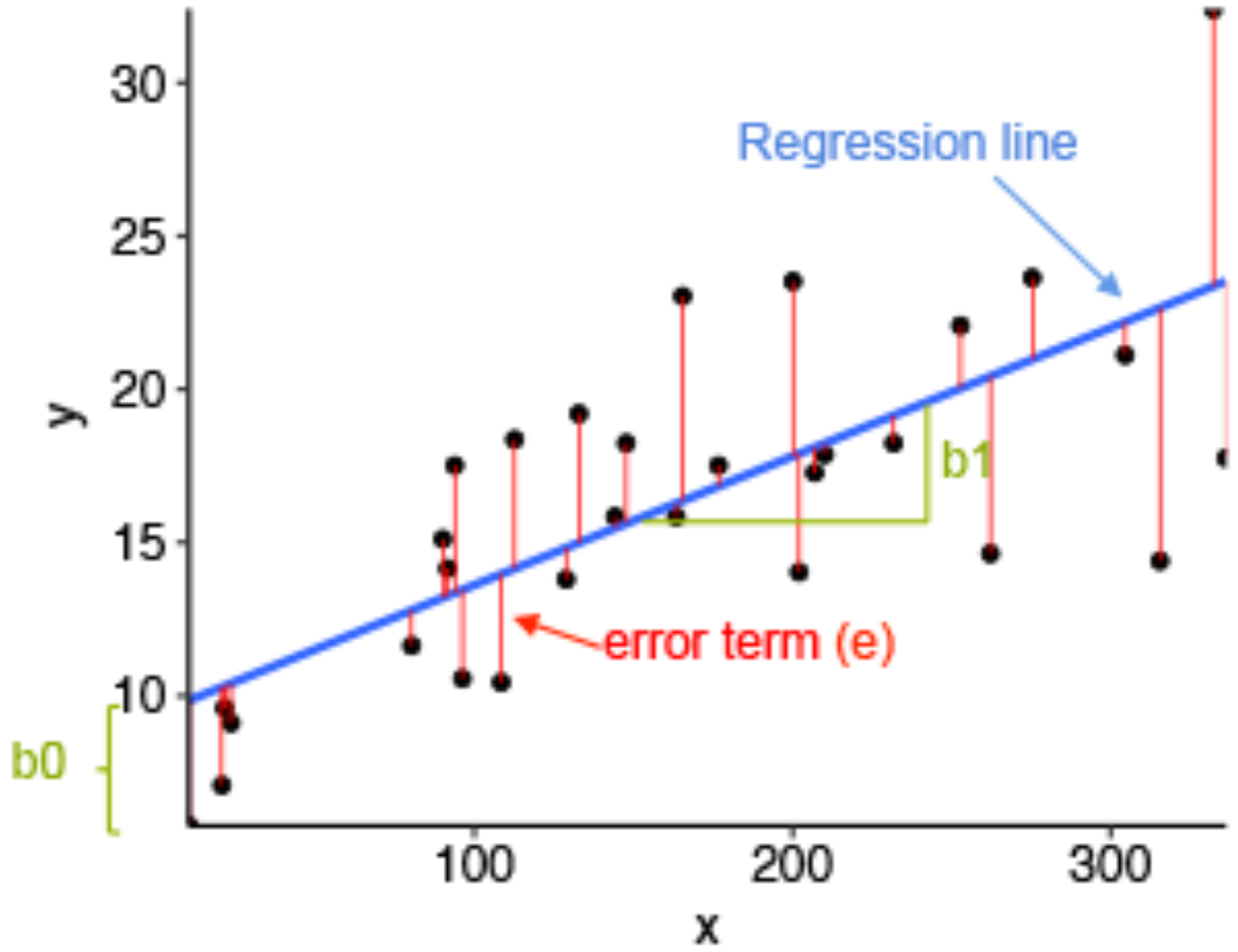


Error in Linear Regression

- The error (also known as residual) is the difference between the actual data point and the predicted value from the linear function.
- The objective in linear regression is to minimize the sum of squared errors
- Given a dataset with n pairs of observed values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the goal is to find the best values for β_0 and β_1 that minimize the sum of squared residuals (SSR), which is calculated as:

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

Where $\hat{y}_i = \beta_0 + \beta_1 x_i$ is the predicted value of y given x_i



Estimating the Slope Coefficient (β_1)

- The equation for estimating the slope coefficient β_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 (β_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.

Thank You