Regression

By
Dr Ravi Prakash Verma
Professor
Department of CSAI
ABESIT

Regression

- Regression analysis is a common tool used in business, finance and other fields to study variable dependency.
- It can helps a professional in these areas understand the relationship between key variables.
- Learning about regression and its various methods can help you gain the analytic skills necessary to succeed in a data-driven position.
- Regression is a **statistical and machine learning technique** used to model and analyze the relationship between **independent variables (inputs/features)** and **dependent variables (outputs/targets)**.
- The primary goal is to **predict continuous values** based on input data.

Regression Types

- **1 Simple regression**
  - Simple regression methods help you estimate the relationship between a dependent variable and one independent variable.
  - For example, you might use simple regression to compare the connection between umbrella sales (dependent variable) and the amount of rain a meteorologist forecasts (independent variable).
  - Other examples of this variable relationship might be how much money someone earns based on their level of education or how high lumber prices get during labor shortages.
  - One independent variable predicts one dependent variable.
  - Example: Predicting a person’s weight based on height.

Regression Types

- **2 Multiple regression**
  - Multiple regression analysis methods help you determine the relationship between a dependent variable and more than one independent variable.
  - Adding more independent variables makes for a more complex regression analysis study, though it often generates more specific and realistic results.
  - For example, you might evaluate if more umbrellas sell when the meteorologist forecasts rainy weather specifically during spring or comparatively across all seasons.
  - Or you might review salary earnings for education, experience and proximity to a metropolitan area.

  - Multiple independent variables predict a dependent variable.
  - Example: Predicting house prices based on area, number of bedrooms, and location.

Regression

- **3 Linear regression**
  - Linear regression analysis is a simple regression type that requires you to create a hypothetical line that best connects all data points.
  - You determine the best fit line with linear regression and establish a predictor error between the predicted value based on the line and what's actually observed.
  - The disadvantage of linear regression is the potential for outliers in the data so it's frequently used for small data pools of information or predictions.
  - This is because some data points may not fit neatly into the regression line.

Regression

- **4. Multiple linear regression**
  - Similar to linear regression, multiple linear regression shows the direct or linear correlation between variables. The difference is that it involves more than one dependent variable.
  - Even though multiple linear regression may involve more dependent variables, it's also best used for smaller batches of data to prevent accuracy issues with outliers.
- **5. Logistic regression**
  - Logistic regression helps measure the relationship between dependent and independent variables, though it doesn't correlate between independent variables.
  - You often have a large data set when using logistic regression, and the dependent variable is usually discrete, meaning that you can count all values within a finite amount of time.
  - With logic regression, the target variable typically only has two values, and a sigmoid curve shows the correlation.

Regression

- **6. Ridge regression**
- **7. Lasso regression**
- **8. Polynomial regression**
- **9. Bayesian linear regression**
- **10. Jackknife regression**
- **11. Elastic net regression**
- **12. Ecological regression**
- **13. Stepwise regression**
- **14. Decision Tree and Random Forest Regression**
  - Uses tree-based models for non-linear relationships.
  - Example: Predicting housing prices using decision rules.
- **15. Support Vector Regression (SVR)**
  - Uses support vector machines for regression tasks.
  - Example: Predicting stock market trends.

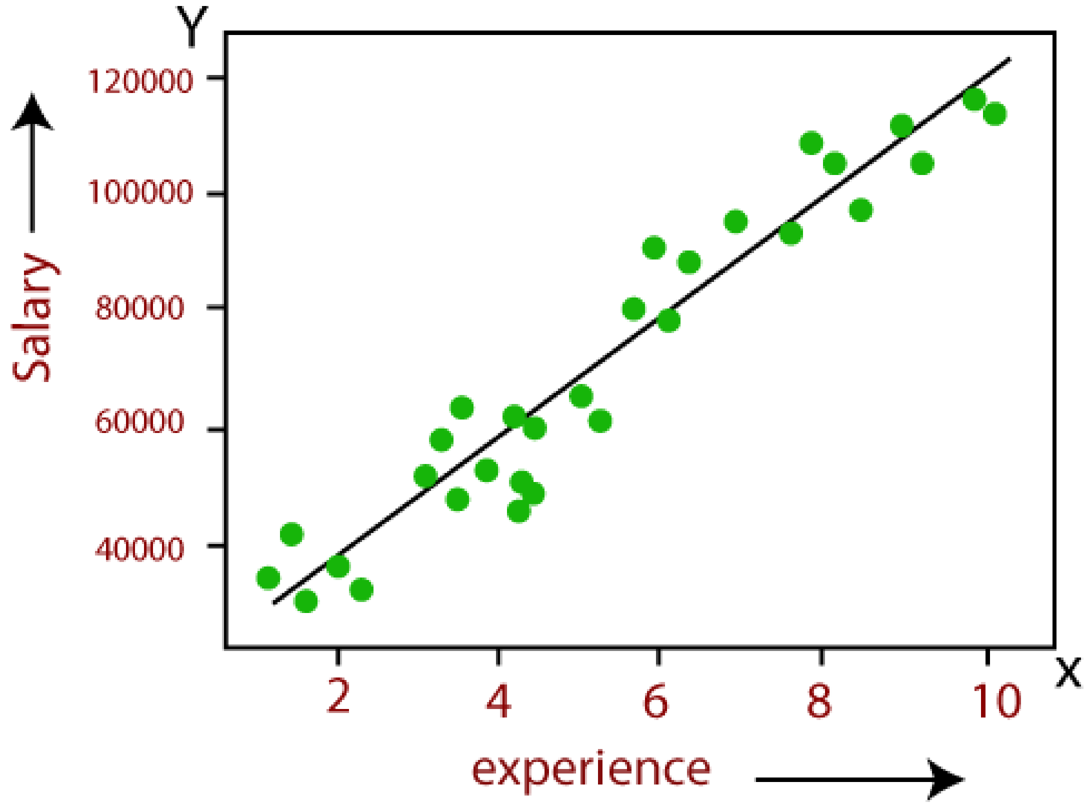
Linear Regression

- Linear Regression is a **supervised learning algorithm** used for **predicting continuous values**.
- It models the relationship between an **independent variable (input)** and a **dependent variable (output)** by fitting a straight line to the data.

Linear Regression

- **Equation of a Linear Model**
  - The equation of **Simple Linear Regression** (one independent variable) is:
    - y=mx+c
  - where:
    - y = **Predicted output (dependent variable)**
    - x = **Input feature (independent variable)**
    - m = **Slope of the line (coefficient)**
    - c = **Intercept (constant term)**

Linear Regression



Multiple Linear Regression

- For **Multiple Linear Regression** (more than one independent variable), the equation becomes:
  - $y=b_0+b_1x_1+b_2x_2+...+b_nx_n$
  - where:
    - y = **Predicted value (independent variables)**
    - b_0 = **Intercept**
    - $b_1,b_2,...,b_n$ = **Coefficients for each feature**
    - $x_1,x_2,...,x_n$ = **Input features (independent variables)**

Linear Regression

- The goal of Linear Regression is to find the **best-fitting line** that minimizes the error between the predicted values and actual values.
- This is done using **Ordinary Least Squares (OLS)**, which minimizes the **Mean Squared Error (MSE):**
  - $MSE=(1/n)\sum(y_{actual}-y_{predicted})^2$
- where:
  - y_{actual} = True value from dataset
  - $y_{predicted}$ = Predicted value using the regression equation
  - n = Number of data points

Linear Regression

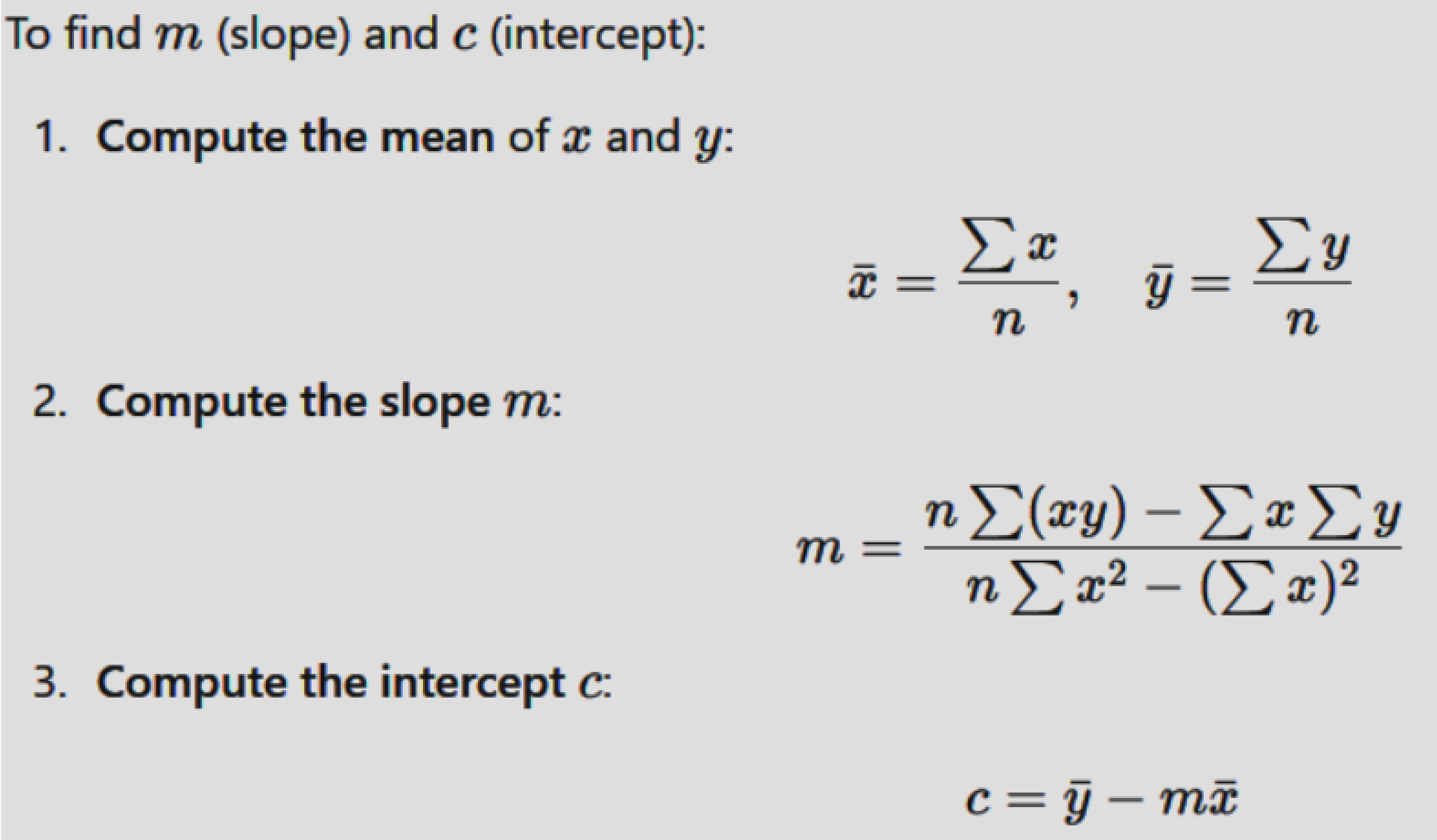
- **Steps to Perform Linear Regression**
  - **Step 1: Collect Data**
    - Gather historical data with known inputs (x) and outputs (y).
  - **Step 2: Plot the Data**
    - Visualize the data to check if a **linear trend** exists.
  - **Step 3: Compute the Regression Line**
    - Use the **Least Squares Method** to find the best-fit line by determining mmm (slope) and c (intercept).
  - **Step 4: Make Predictions**
    - Use the equation y=mx+c to predict new values.
  - **Step 5: Evaluate the Model**
    - Calculate metrics like:
      - **Mean Squared Error (MSE)**
      - **R-squared (R^2) Score** (measures how well the model fits the data)

Linear Regression

- **Step-by-Step Calculation**
  - **Step 1: Define the Relationship**
    - Identify the dependent variable y and independent variable(s) x.
  - **Step 2: Gather Data**
    - Collect historical data of x and y.
  - **Step 3: Choose a Regression Model**
    - If the data follows a **straight-line trend**, use **Linear Regression**.
    - If the data is **curved**, use **Polynomial Regression**.
    - If predicting **Yes/No outcomes**, use **Logistic Regression**.
  - **Step 4: Compute Regression Coefficients**
    - For **Simple Linear Regression**, the formula for the best-fit line is:
    - $y=mx+c$

Linear Regression

- **Step-by-Step Calculation**



To find m (slope) and c (intercept):

1. **Compute the mean of x and y:**

$$\bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}$$

2. **Compute the slope m:**

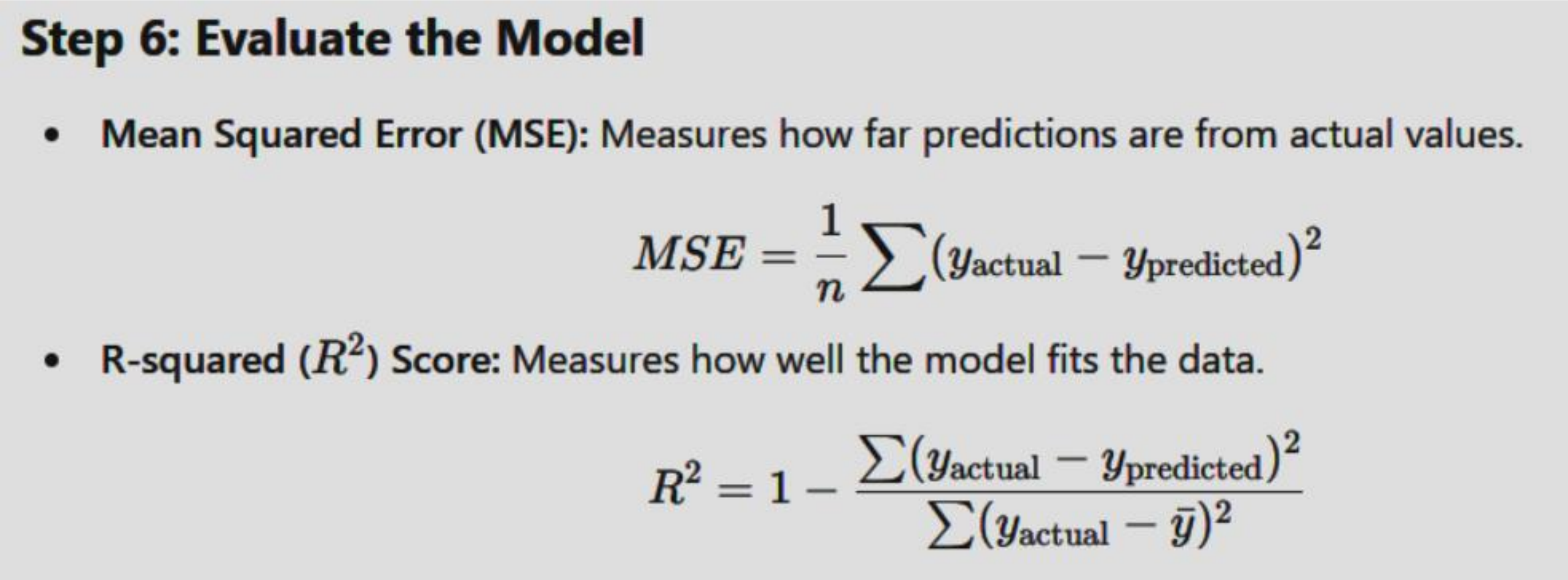
$$m = \frac{n\sum(xy) - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

3. **Compute the intercept c:**

$$c = \bar{y} - m\bar{x}$$

Linear Regression

- **Step-by-Step Calculation**
  - **Step 5: Make Predictions**
    - Once the equation is determined, predict values using:
    - y=mx+c



Step 6: Evaluate the Model

- **Mean Squared Error (MSE):** Measures how far predictions are from actual values.

$$MSE = \frac{1}{n}\sum(y_{\text{actual}} - y_{\text{predicted}})^2$$

- **R-squared (R^2) Score:** Measures how well the model fits the data.

$$R^2 = 1 - \frac{\sum(y_{\text{actual}} - y_{\text{predicted}})^2}{\sum(y_{\text{actual}} - \bar{y})^2}$$

Linear Regression

- Example
- Given Data:

X (Hours Studied)	Y (Exam Score)
1	50
2	55
3	60
4	65
5	70

Linear Regression

- **Step 1: Compute the Slope (m)**
- $m=(N\sum(xy)-\sum x\sum y)/\ (N\sum x^2-(\sum x)^2)$
- Using calculations:
- $m=5(1*50+2*55+3*60+4*65+5*70)-(1+2+3+4+5)(50+55+60+65+70)/\ 5(1^2+2^2+3^2+4^2+5^2)-(1+2+3+4+5)^2$
- $m=5$

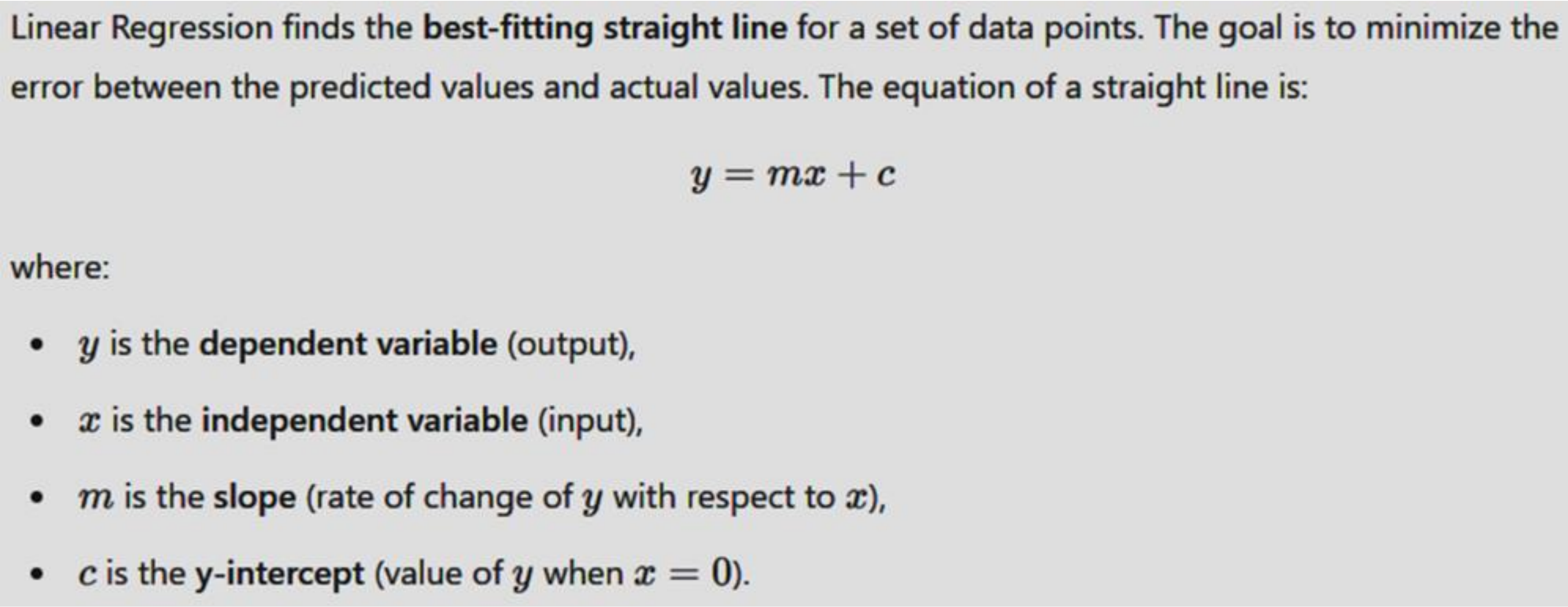
Linear Regression

- **Step 2: Compute the Intercept (c)**
  - C =(∑y−m∑x)/N
  - C = ((50+55+60+65+70)−5(1+2+3+4+5))/5
  - C = 45
- **Step 3: Regression Equation**
  - y=5x+45
- **Step 4: Make a Prediction**
- If a student studies **6 hours**, the predicted exam score:
  - y=5(6)+45=75

Linear Regression

- **Advantages of Linear Regression**
  - ✅ **Simple and easy to interpret**
  - ✅ **Computationally efficient**
  - ✅ **Works well for small to medium-sized datasets**
- **Limitations of Linear Regression**
  - ❌ **Assumes a linear relationship** (won't work well for non-linear patterns)
    ❌ **Sensitive to outliers**
    ❌ **Assumes no multicollinearity** in multiple regression

Derivation of Linear Regression



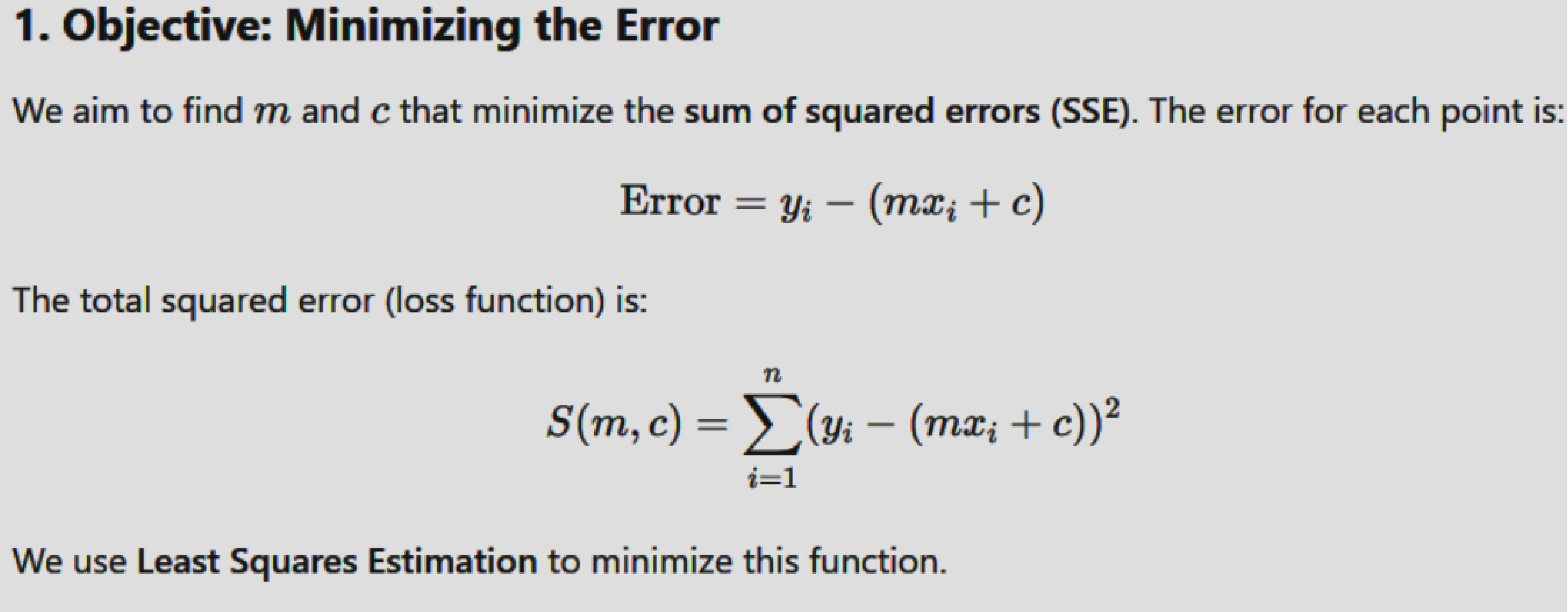
Linear Regression finds the **best-fitting straight line** for a set of data points. The goal is to minimize the error between the predicted values and actual values. The equation of a straight line is:

$$y = mx + c$$

where:

- y is the **dependent variable** (output),
- x is the **independent variable** (input),
- m is the **slope** (rate of change of y with respect to x),
- c is the **y-intercept** (value of y when $x = 0$).

Derivation of Linear Regression



1. Objective: Minimizing the Error

We aim to find m and c that minimize the **sum of squared errors (SSE)**. The error for each point is:

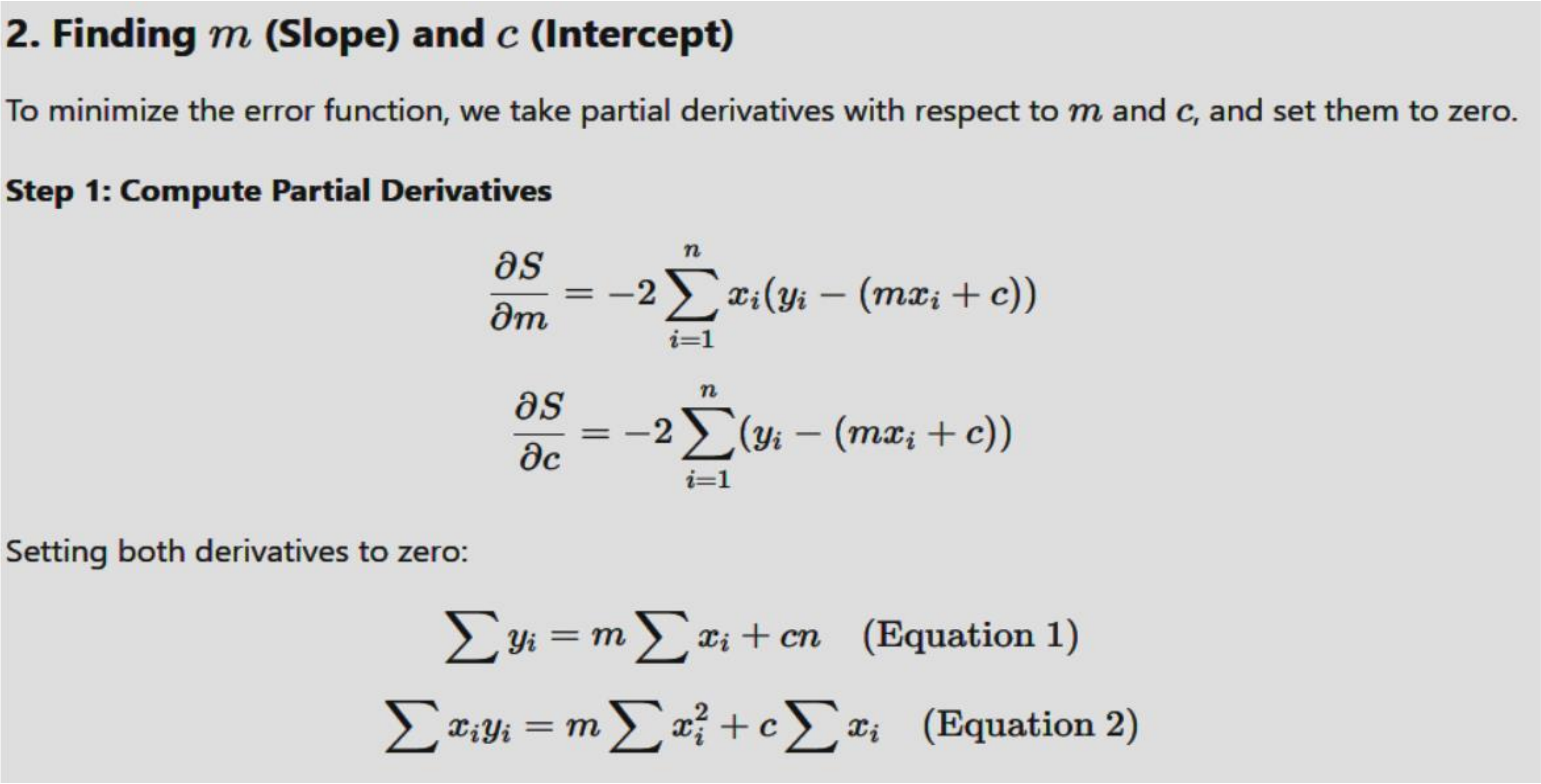
$$\text{Error} = y_i - (mx_i + c)$$

The total squared error (loss function) is:

$$S(m, c) = \sum_{i=1}^{n} (y_i - (mx_i + c))^2$$

We use **Least Squares Estimation** to minimize this function.

Derivation of Linear Regression



2. Finding m (Slope) and c (Intercept)

To minimize the error function, we take partial derivatives with respect to m and c, and set them to zero.

Step 1: Compute Partial Derivatives

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

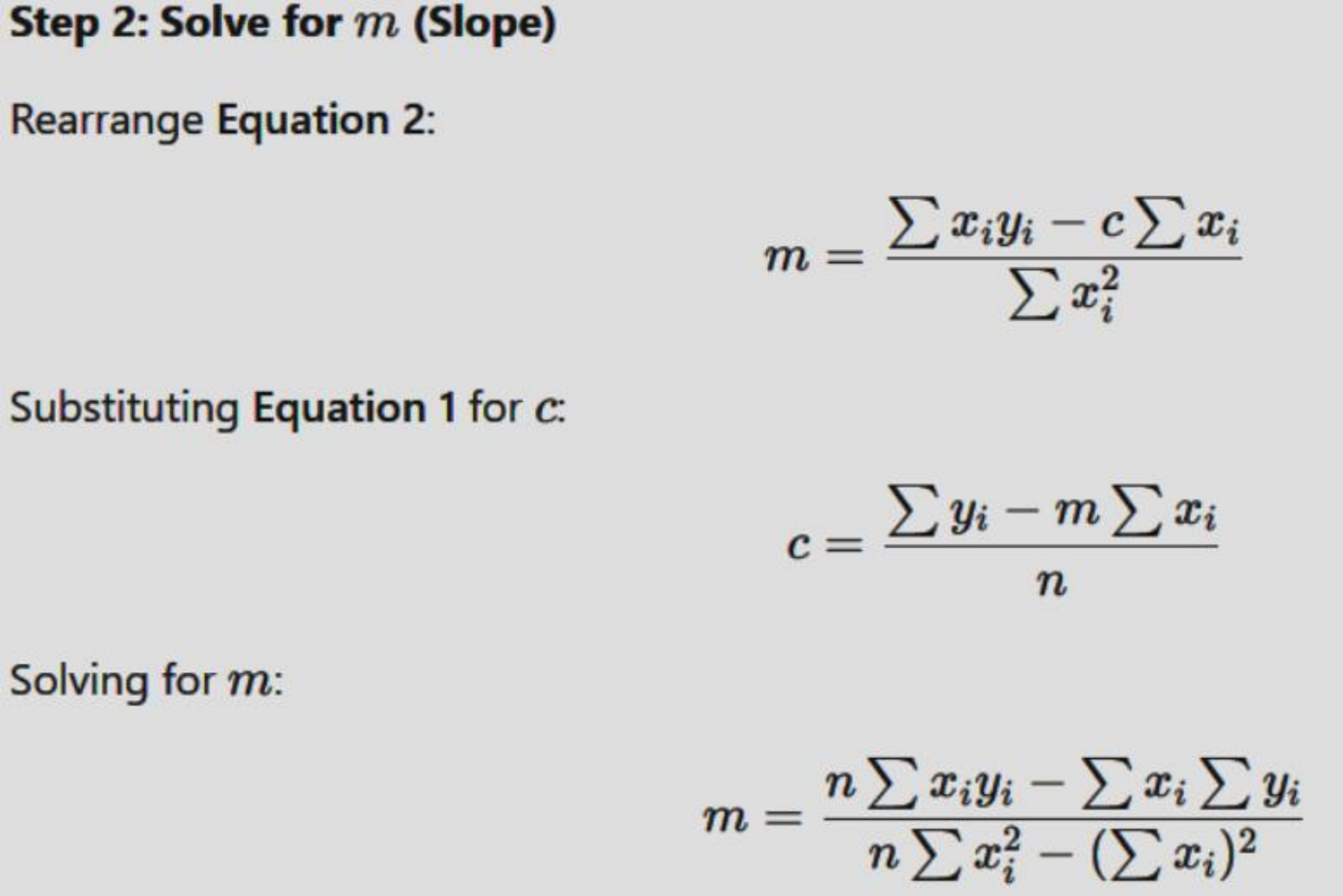
$$\frac{\partial S}{\partial c} = -2\sum_{i=1}^{n} (y_i - (mx_i + c))$$

Setting both derivatives to zero:

$$\sum y_i = m\sum x_i + cn \quad \text{(Equation 1)}$$

$$\sum x_i y_i = m\sum x_i^2 + c\sum x_i \quad \text{(Equation 2)}$$

Derivation of Linear Regression



Step 2: Solve for m (Slope)

Rearrange **Equation 2**:

$$m = \frac{\sum x_i y_i - c \sum x_i}{\sum x_i^2}$$

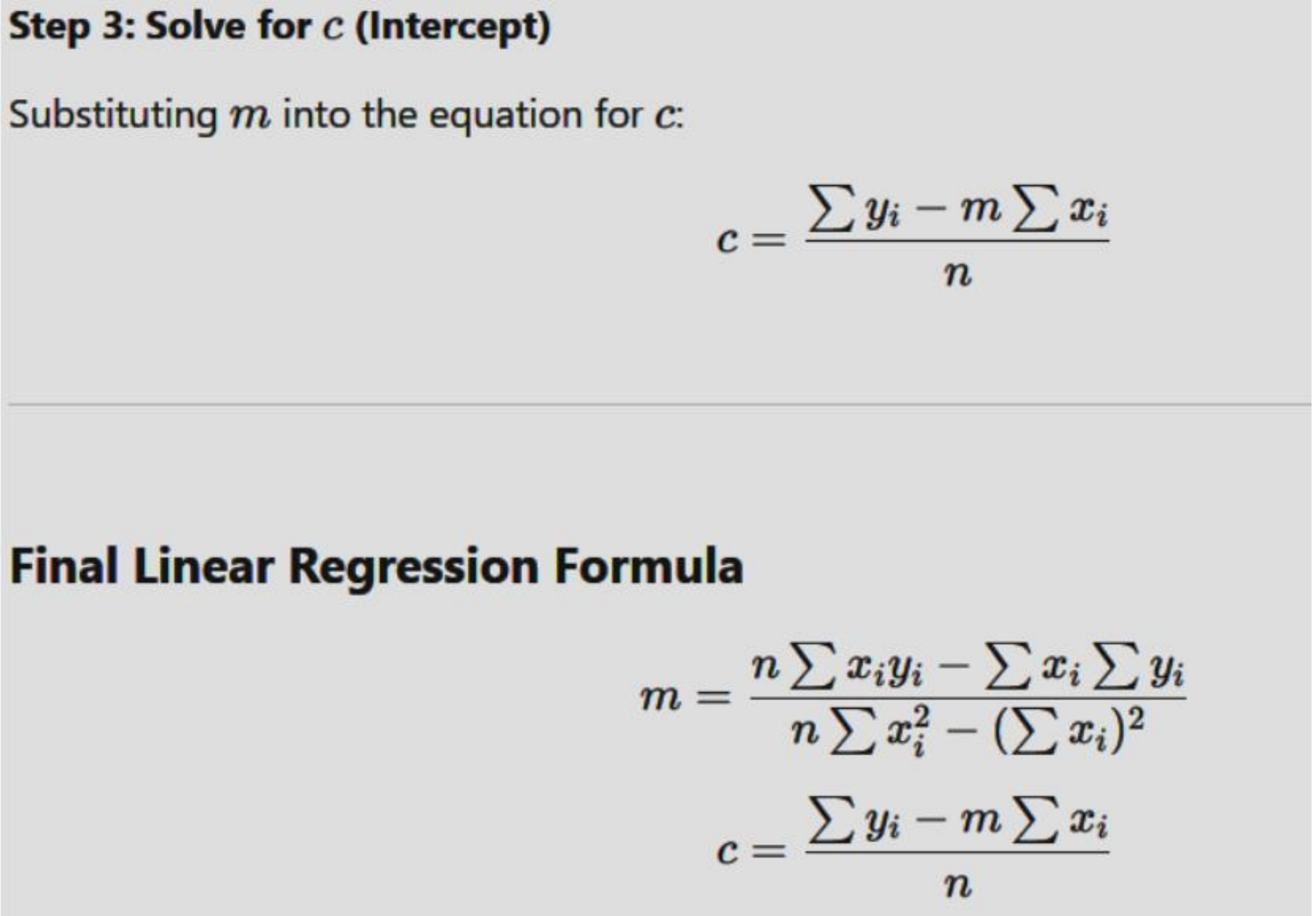
Substituting **Equation 1** for c:

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

Solving for m:

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

Derivation of Linear Regression



Step 3: Solve for c (Intercept)

Substituting m into the equation for c:

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

Final Linear Regression Formula

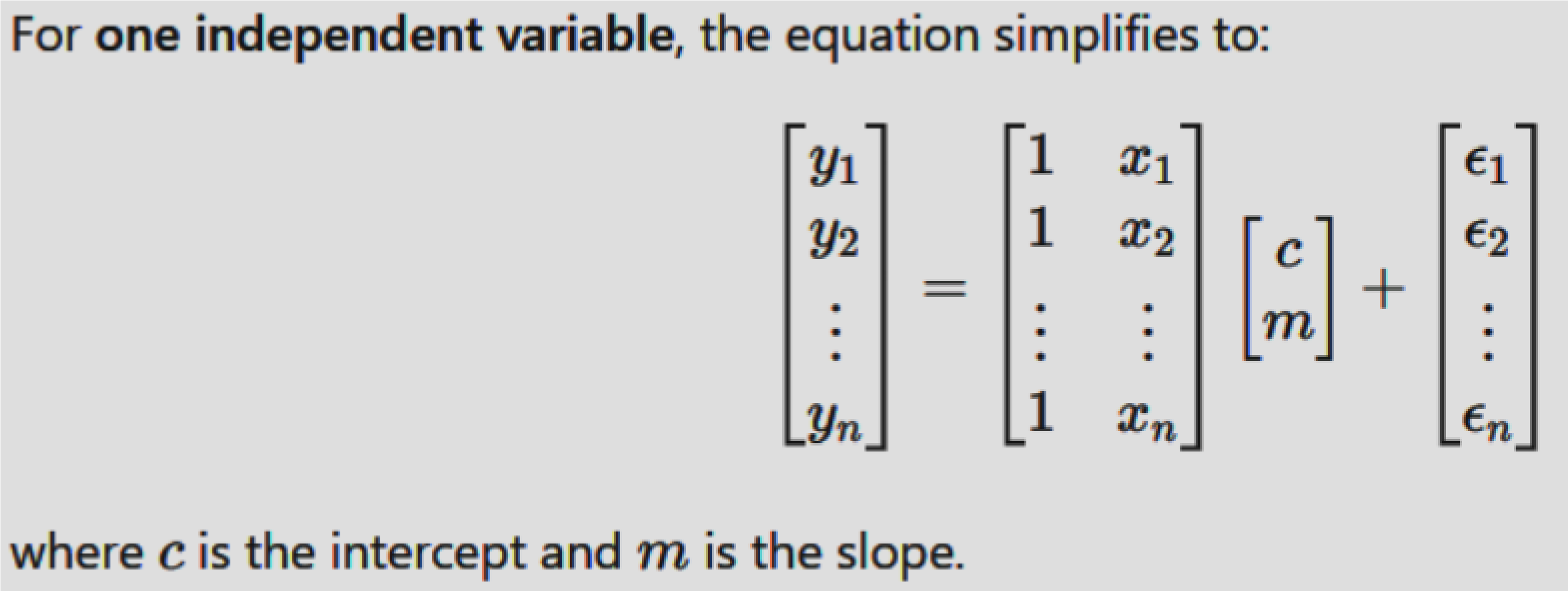
$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

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

Linear Regression in Matrix Form

- **Linear Regression in Matrix Form**
  - Linear regression can also be derived and solved using **matrix notation** for multiple input variables (**Multivariate Linear Regression**).
  - **1. Linear Regression Equation in Matrix Form**
  - For multiple input variables, the linear regression model is:
  - $Y=X\beta+\epsilon$
- where:
  - Y is the **n×1** vector of output values (dependent variable),
  - X is the **n×(p+1)** matrix of input features (independent variables), including a column of ones for the intercept,
  - β is the **(p+1)×1** vector of coefficients (parameters: slope and intercept),
  - ϵ is the **n×1** vector of errors (differences between predictions and actual values).

Linear Regression in Matrix Form

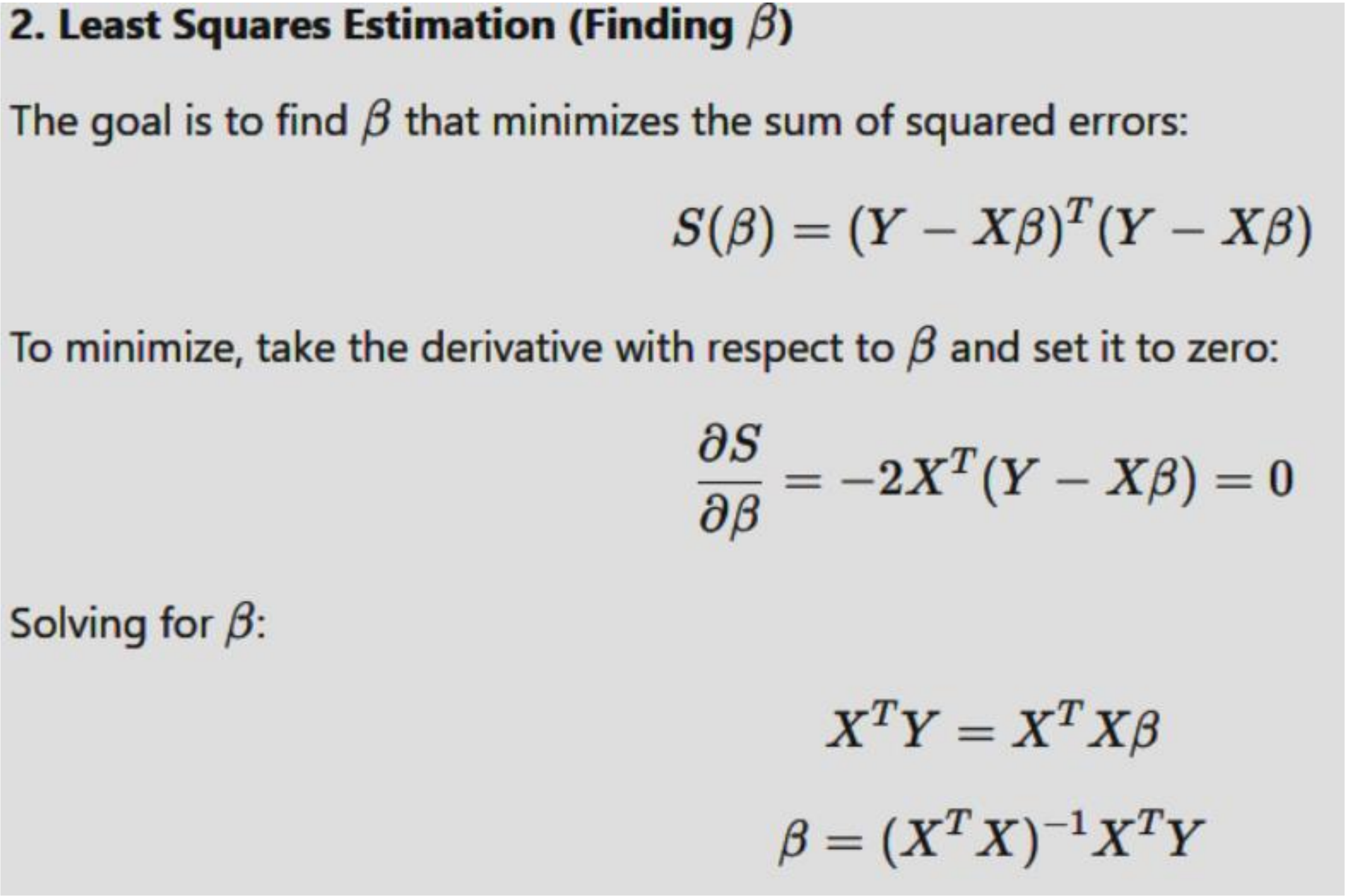


For **one independent variable**, the equation simplifies to:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} c \\ m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where c is the intercept and m is the slope.

Linear Regression in Matrix Form



2. Least Squares Estimation (Finding β)

The goal is to find β that minimizes the sum of squared errors:

$$S(\beta) = (Y - X\beta)^T(Y - X\beta)$$

To minimize, take the derivative with respect to β and set it to zero:

$$\frac{\partial S}{\partial \beta} = -2X^T(Y - X\beta) = 0$$

Solving for β:

$$X^TY = X^TX\beta$$

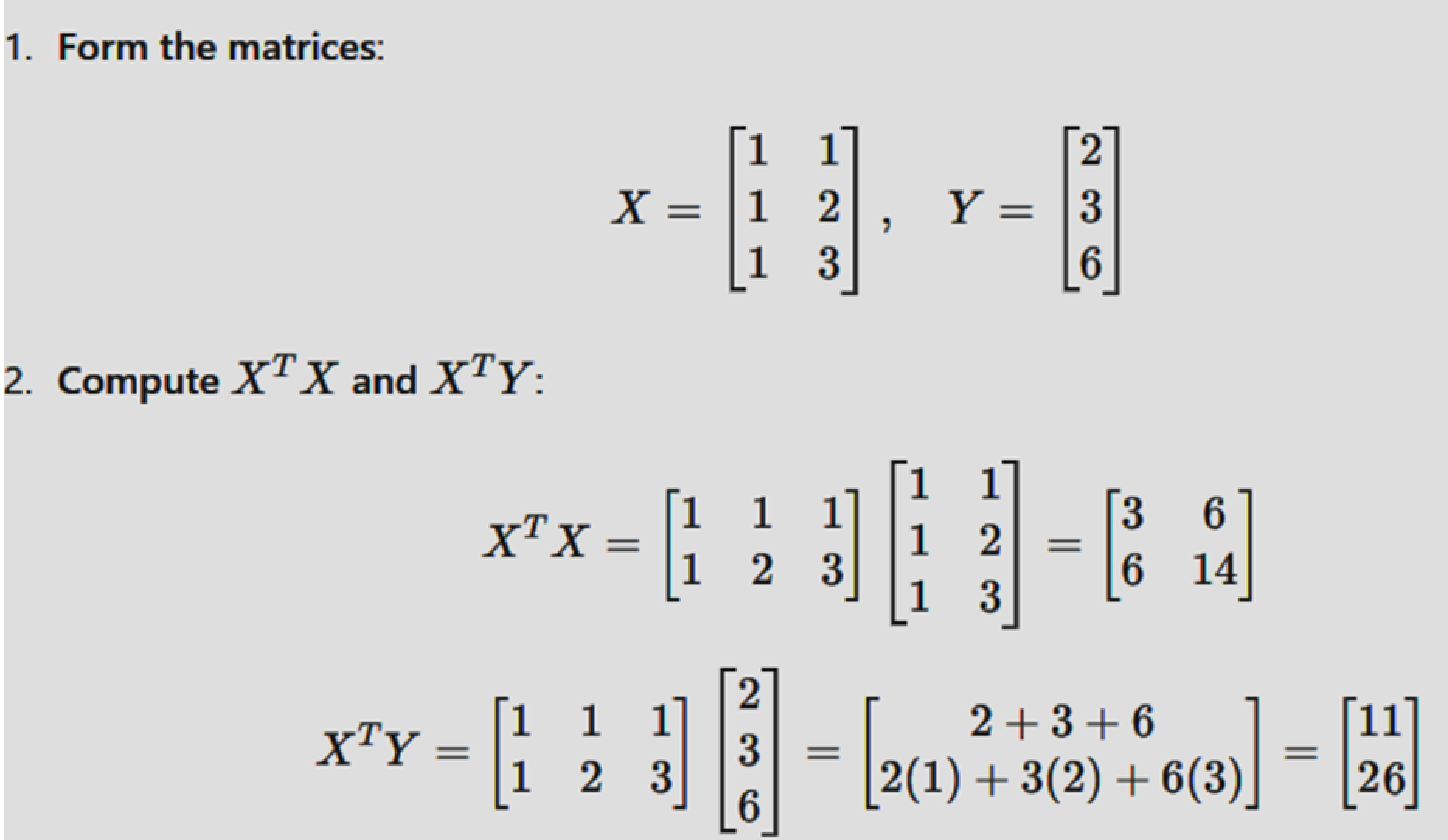
$$\beta = (X^TX)^{-1}X^TY$$

Linear Regression in Matrix Form

- Example
- Let’s take a simple dataset:

x	y
1	2
2	3
3	6

Linear Regression in Matrix Form



1. **Form the matrices:**

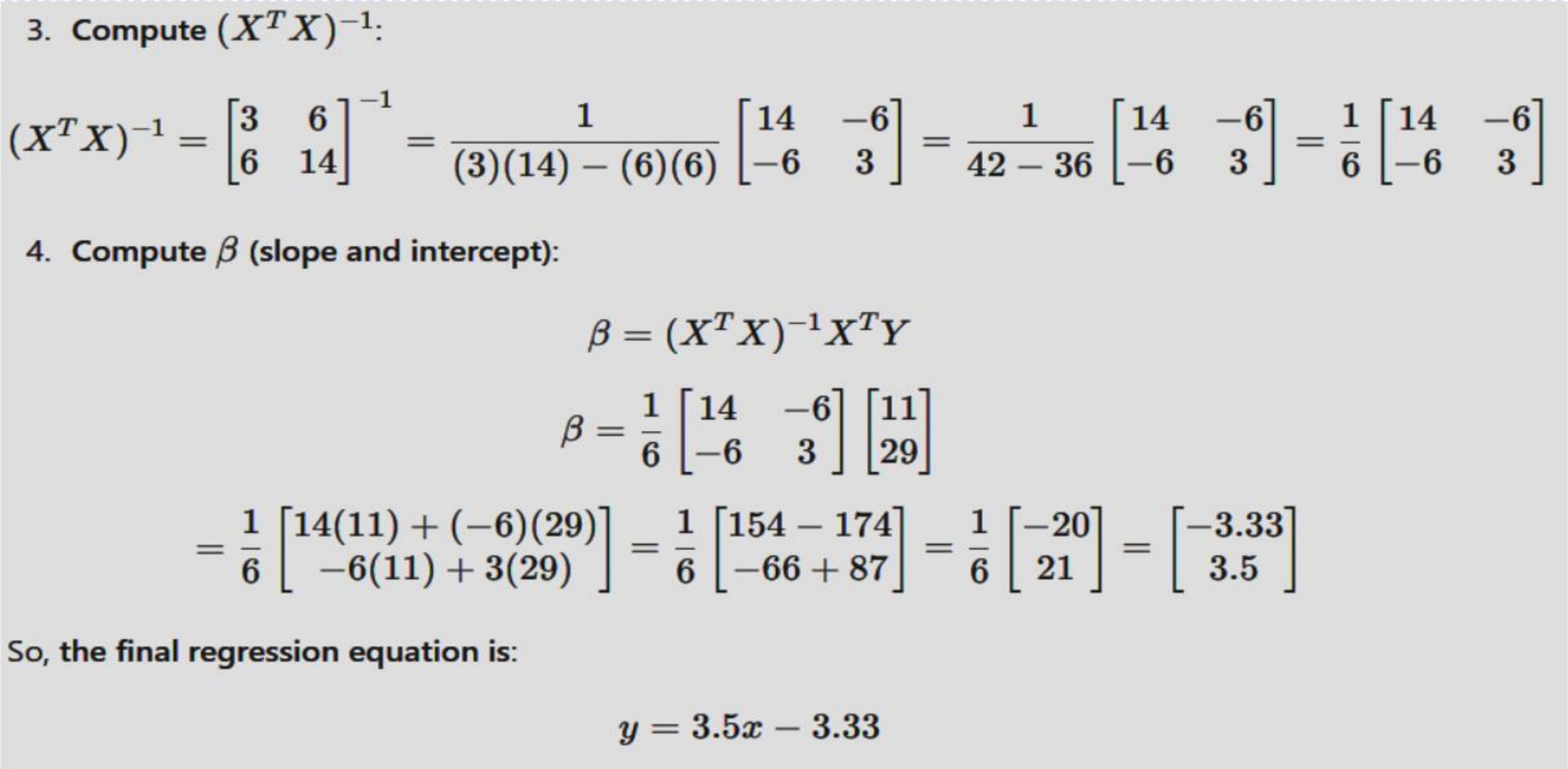
$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad Y = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix}$$

2. **Compute $X^T X$ and $X^T Y$:**

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 2+3+6 \\ 2(1)+3(2)+6(3) \end{bmatrix} = \begin{bmatrix} 11 \\ 26 \end{bmatrix}$$

Linear Regression in Matrix Form



3. **Compute $(X^TX)^{-1}$:**

$$(X^TX)^{-1} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}^{-1} = \frac{1}{(3)(14)-(6)(6)} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \frac{1}{42-36} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix}$$

4. **Compute β (slope and intercept):**

$$\beta = (X^TX)^{-1}X^TY$$

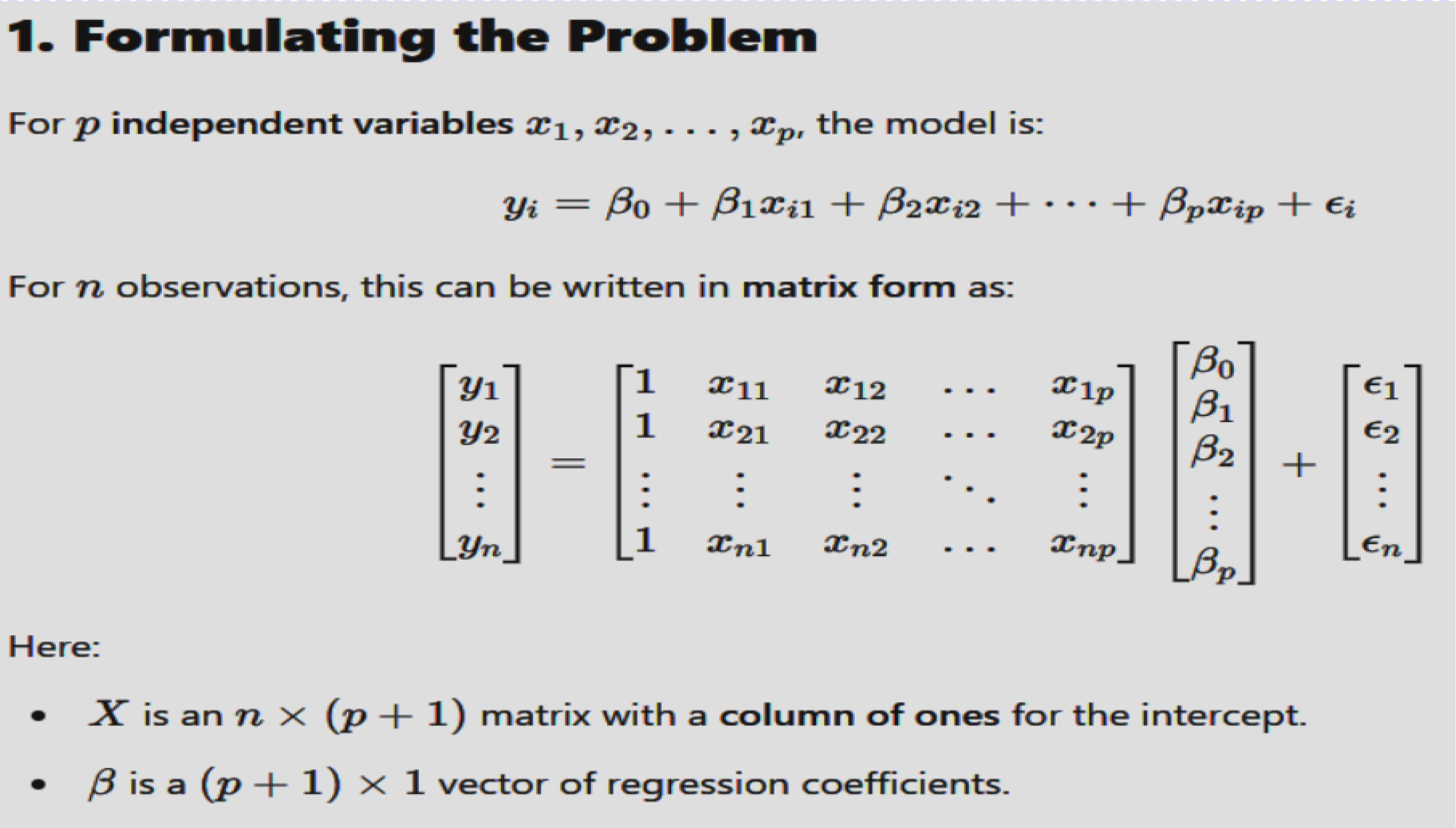
$$\beta = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} \begin{bmatrix} 11 \\ 29 \end{bmatrix}$$

$$= \frac{1}{6} \begin{bmatrix} 14(11) + (-6)(29) \\ -6(11) + 3(29) \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 154 - 174 \\ -66 + 87 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} -20 \\ 21 \end{bmatrix} = \begin{bmatrix} -3.33 \\ 3.5 \end{bmatrix}$$

So, the final regression equation is:

$$y = 3.5x - 3.33$$

Multiple Linear Regression



1. Formulating the Problem

For p **independent variables** $x_1, x_2, \ldots, x_p$, the model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

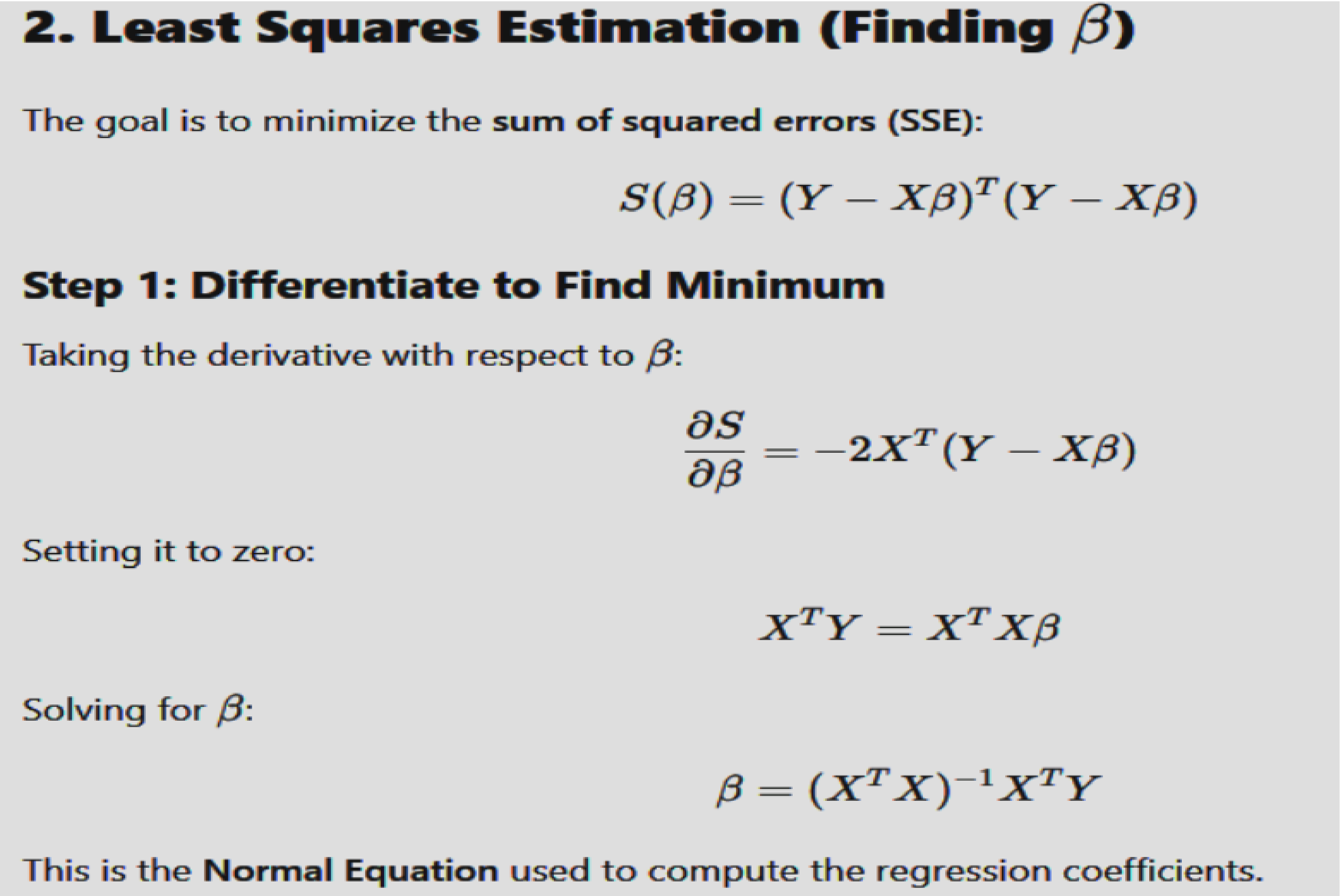
For n observations, this can be written in **matrix form** as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Here:

- X is an $n \times (p+1)$ matrix with a **column of ones** for the intercept.
- β is a $(p+1) \times 1$ vector of regression coefficients.

Multiple Linear Regression



2. Least Squares Estimation (Finding β)

The goal is to minimize the **sum of squared errors (SSE)**:

$$S(\beta) = (Y - X\beta)^T(Y - X\beta)$$

Step 1: Differentiate to Find Minimum

Taking the derivative with respect to β:

$$\frac{\partial S}{\partial \beta} = -2X^T(Y - X\beta)$$

Setting it to zero:

$$X^TY = X^TX\beta$$

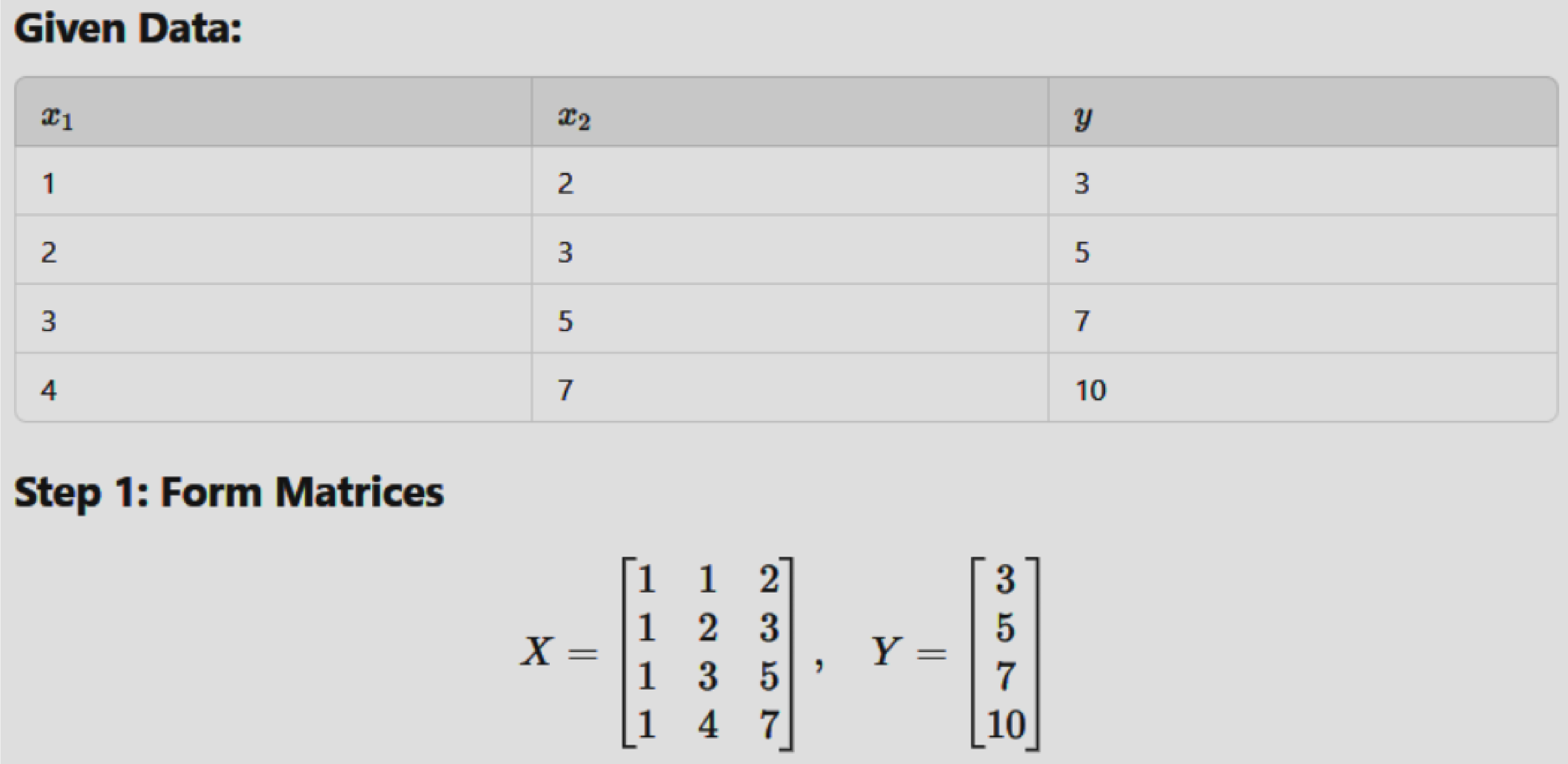
Solving for β:

$$\beta = (X^TX)^{-1}X^TY$$

This is the **Normal Equation** used to compute the regression coefficients.

Multiple Linear Regression

- Example



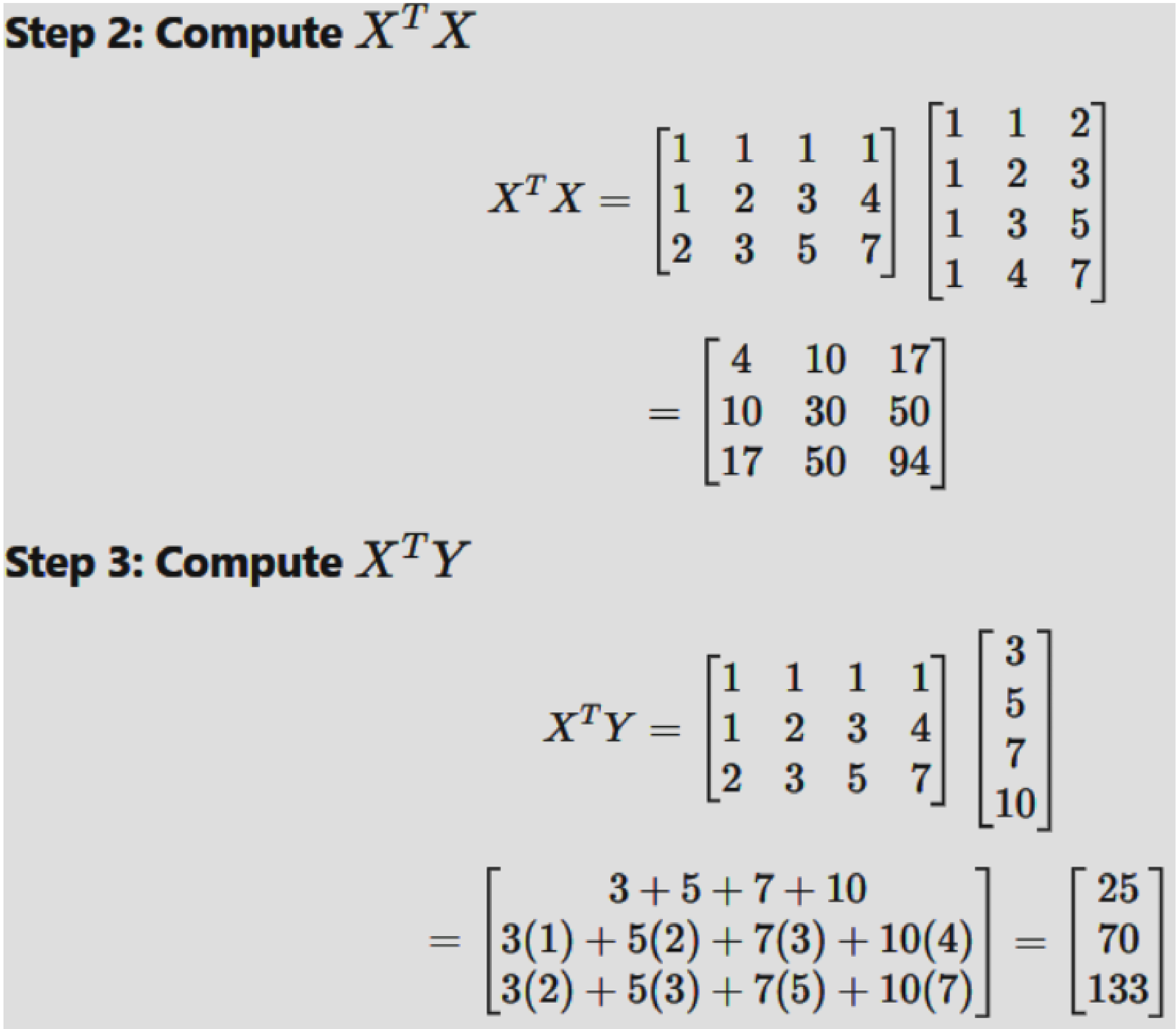
Given Data:

x_1	x_2	y
1	2	3
2	3	5
3	5	7
4	7	10

Step 1: Form Matrices

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 5 \\ 1 & 4 & 7 \end{bmatrix}, \quad Y = \begin{bmatrix} 3 \\ 5 \\ 7 \\ 10 \end{bmatrix}$$

Multiple Linear Regression



Step 2: Compute $X^T X$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 5 & 7 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 5 \\ 1 & 4 & 7 \end{bmatrix}$$

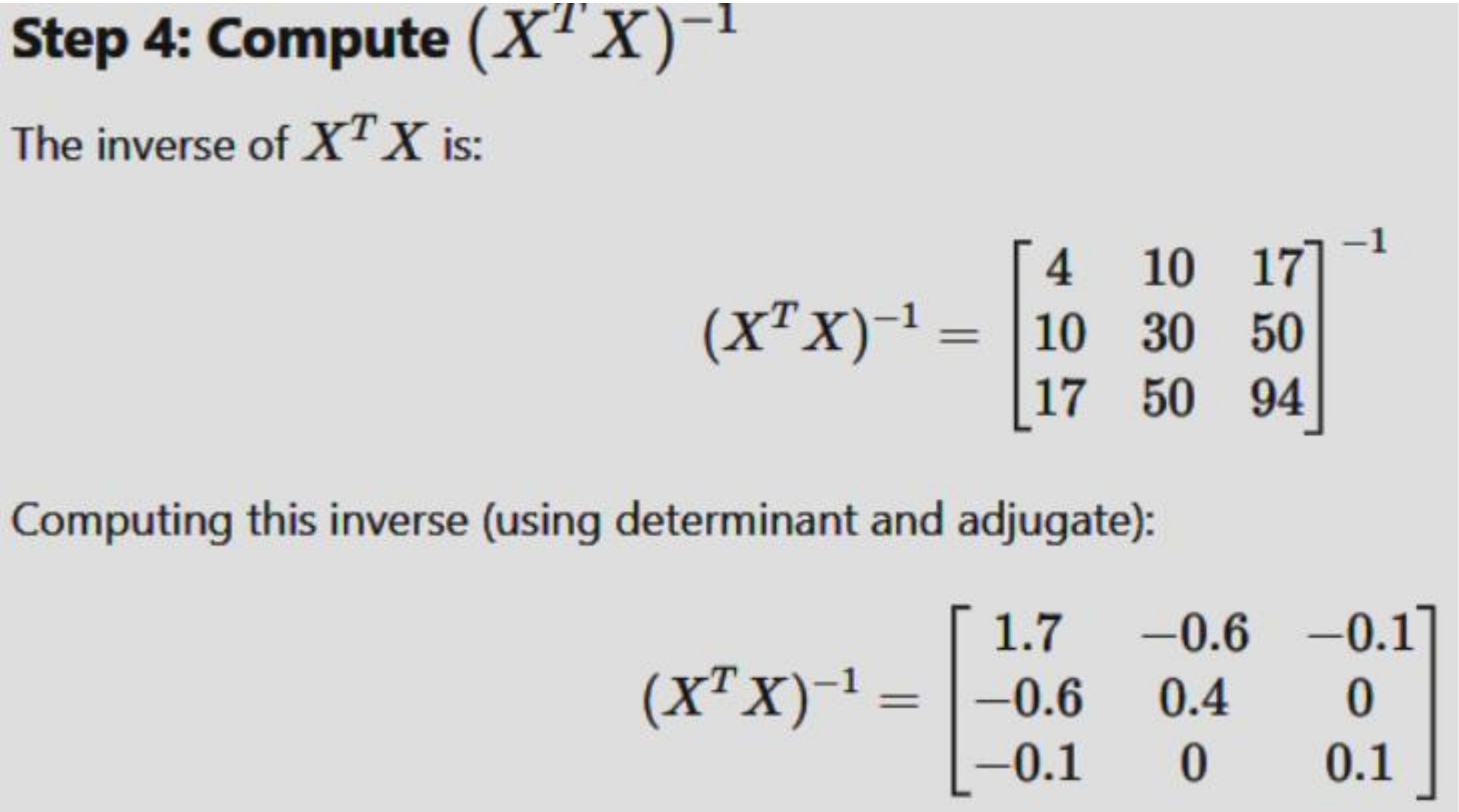
$$= \begin{bmatrix} 4 & 10 & 17 \\ 10 & 30 & 50 \\ 17 & 50 & 94 \end{bmatrix}$$

Step 3: Compute $X^T Y$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 5 & 7 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 7 \\ 10 \end{bmatrix}$$

$$= \begin{bmatrix} 3+5+7+10 \\ 3(1)+5(2)+7(3)+10(4) \\ 3(2)+5(3)+7(5)+10(7) \end{bmatrix} = \begin{bmatrix} 25 \\ 70 \\ 133 \end{bmatrix}$$

Multiple Linear Regression



Step 4: Compute $(X^TX)^{-1}$

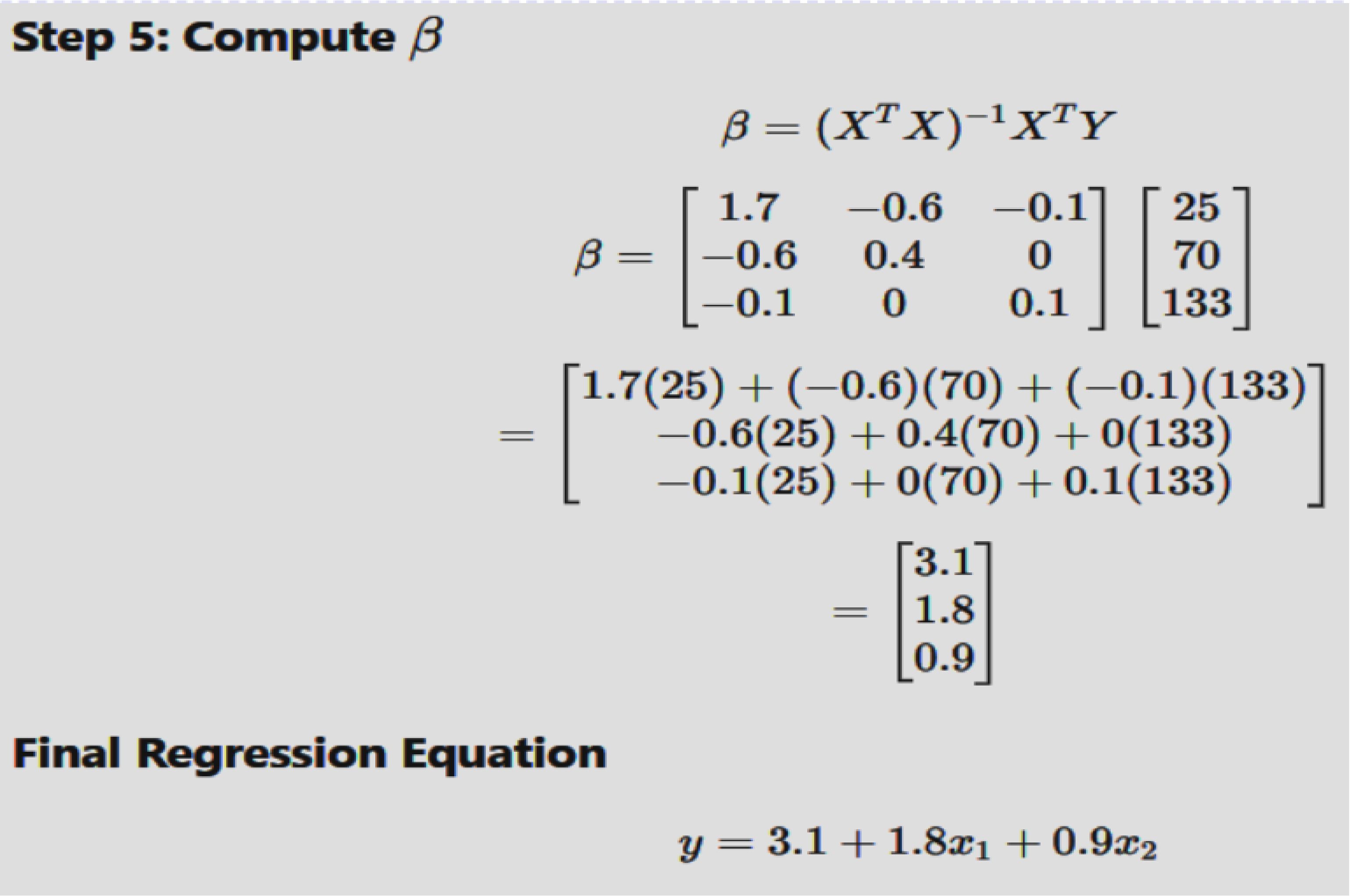
The inverse of X^TX is:

$$(X^TX)^{-1} = \begin{bmatrix} 4 & 10 & 17 \\ 10 & 30 & 50 \\ 17 & 50 & 94 \end{bmatrix}^{-1}$$

Computing this inverse (using determinant and adjugate):

$$(X^TX)^{-1} = \begin{bmatrix} 1.7 & -0.6 & -0.1 \\ -0.6 & 0.4 & 0 \\ -0.1 & 0 & 0.1 \end{bmatrix}$$

Multiple Linear Regression



Step 5: Compute β

$$\beta = (X^T X)^{-1} X^T Y$$

$$\beta = \begin{bmatrix} 1.7 & -0.6 & -0.1 \\ -0.6 & 0.4 & 0 \\ -0.1 & 0 & 0.1 \end{bmatrix} \begin{bmatrix} 25 \\ 70 \\ 133 \end{bmatrix}$$

$$= \begin{bmatrix} 1.7(25) + (-0.6)(70) + (-0.1)(133) \\ -0.6(25) + 0.4(70) + 0(133) \\ -0.1(25) + 0(70) + 0.1(133) \end{bmatrix}$$

$$= \begin{bmatrix} 3.1 \\ 1.8 \\ 0.9 \end{bmatrix}$$

Final Regression Equation

$$y = 3.1 + 1.8x_1 + 0.9x_2$$

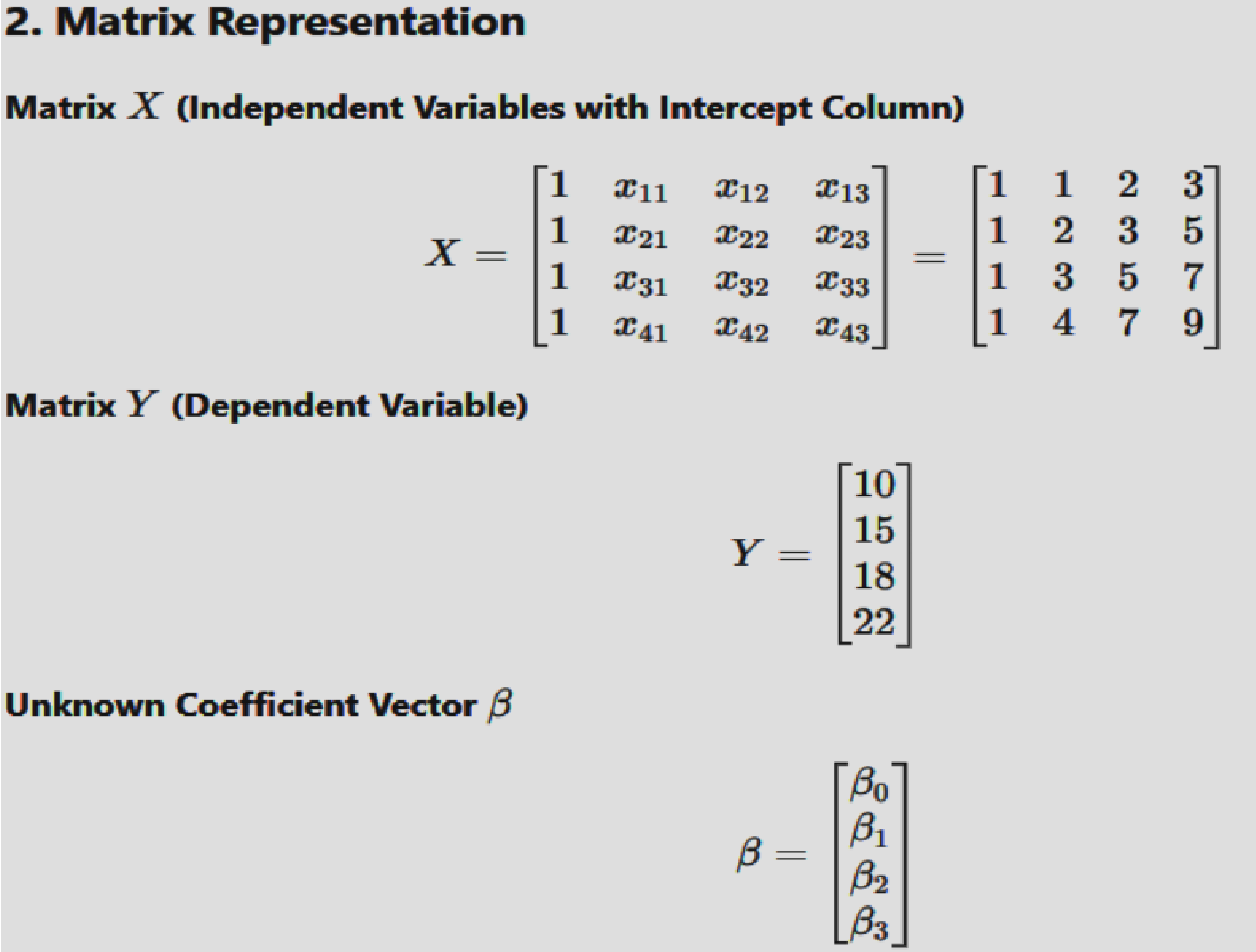
Multiple Linear Regression

- **Conclusion**
  - **Matrix form** allows for efficient computation of regression coefficients, especially for **multiple features**.
  - The key formula is:
  - $\beta = (X^T X)^{-1} X^T Y$
  - This method generalizes well for **any number of independent variables**.

Multiple Linear Regression

x_1	x_2	x_3	y
1	2	3	10
2	3	5	15
3	5	7	18
4	7	9	22

Multiple Linear Regression



2. Matrix Representation

Matrix X (Independent Variables with Intercept Column)

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 3 & 5 \\ 1 & 3 & 5 & 7 \\ 1 & 4 & 7 & 9 \end{bmatrix}$$

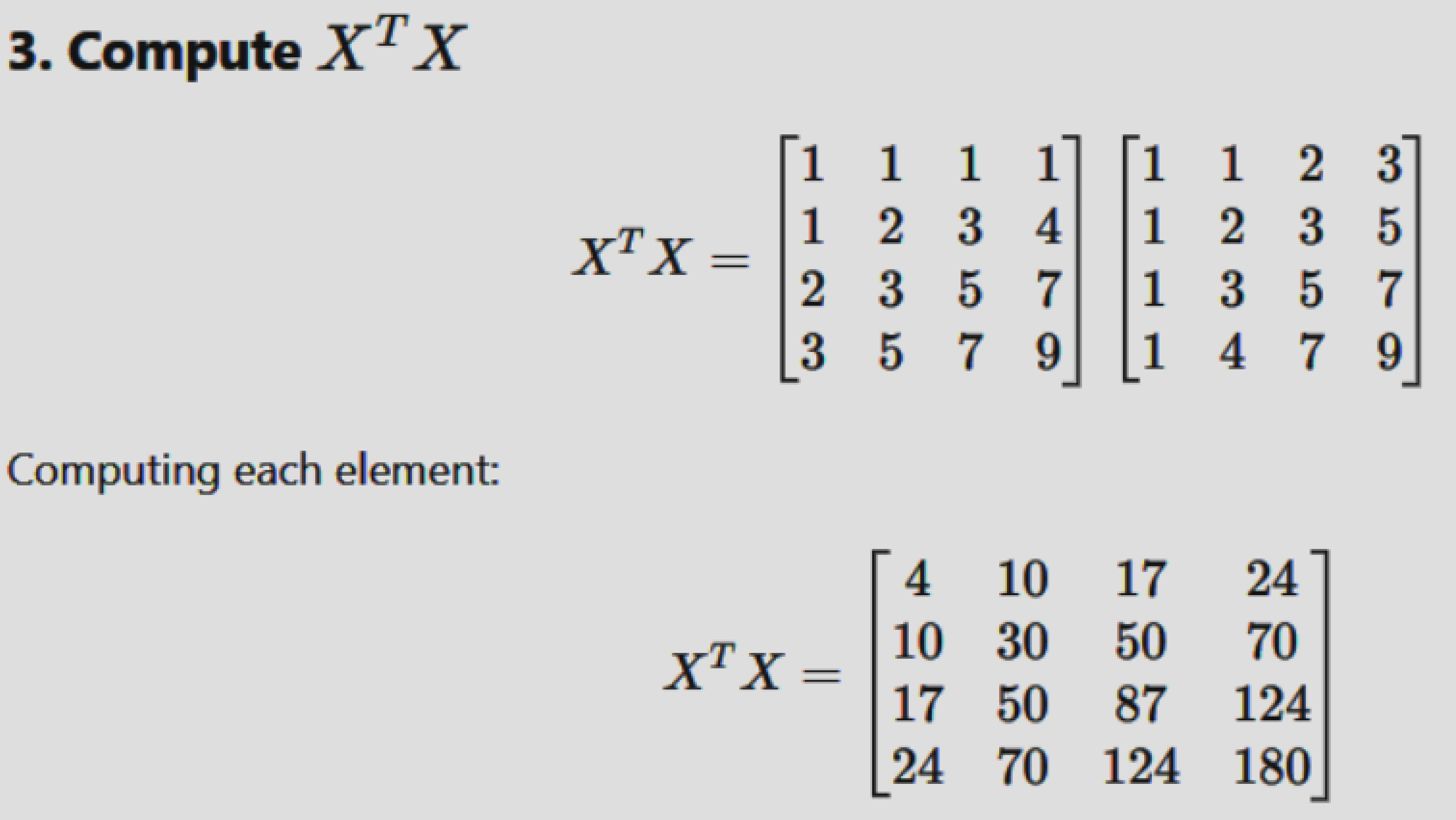
Matrix Y (Dependent Variable)

$$Y = \begin{bmatrix} 10 \\ 15 \\ 18 \\ 22 \end{bmatrix}$$

Unknown Coefficient Vector β

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Multiple Linear Regression



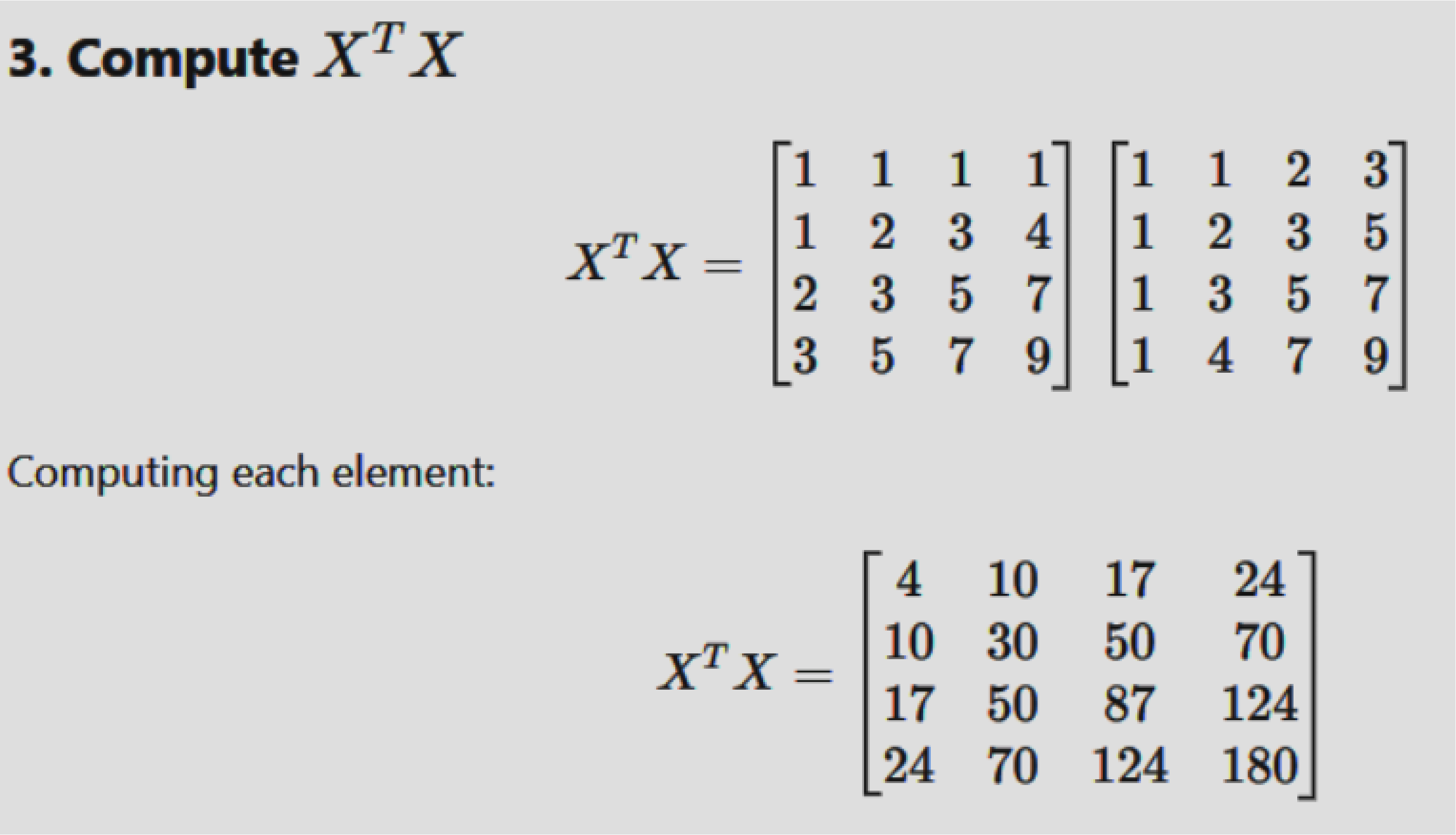
3. Compute $X^T X$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 5 & 7 \\ 3 & 5 & 7 & 9 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 3 & 5 \\ 1 & 3 & 5 & 7 \\ 1 & 4 & 7 & 9 \end{bmatrix}$$

Computing each element:

$$X^T X = \begin{bmatrix} 4 & 10 & 17 & 24 \\ 10 & 30 & 50 & 70 \\ 17 & 50 & 87 & 124 \\ 24 & 70 & 124 & 180 \end{bmatrix}$$

Multiple Linear Regression



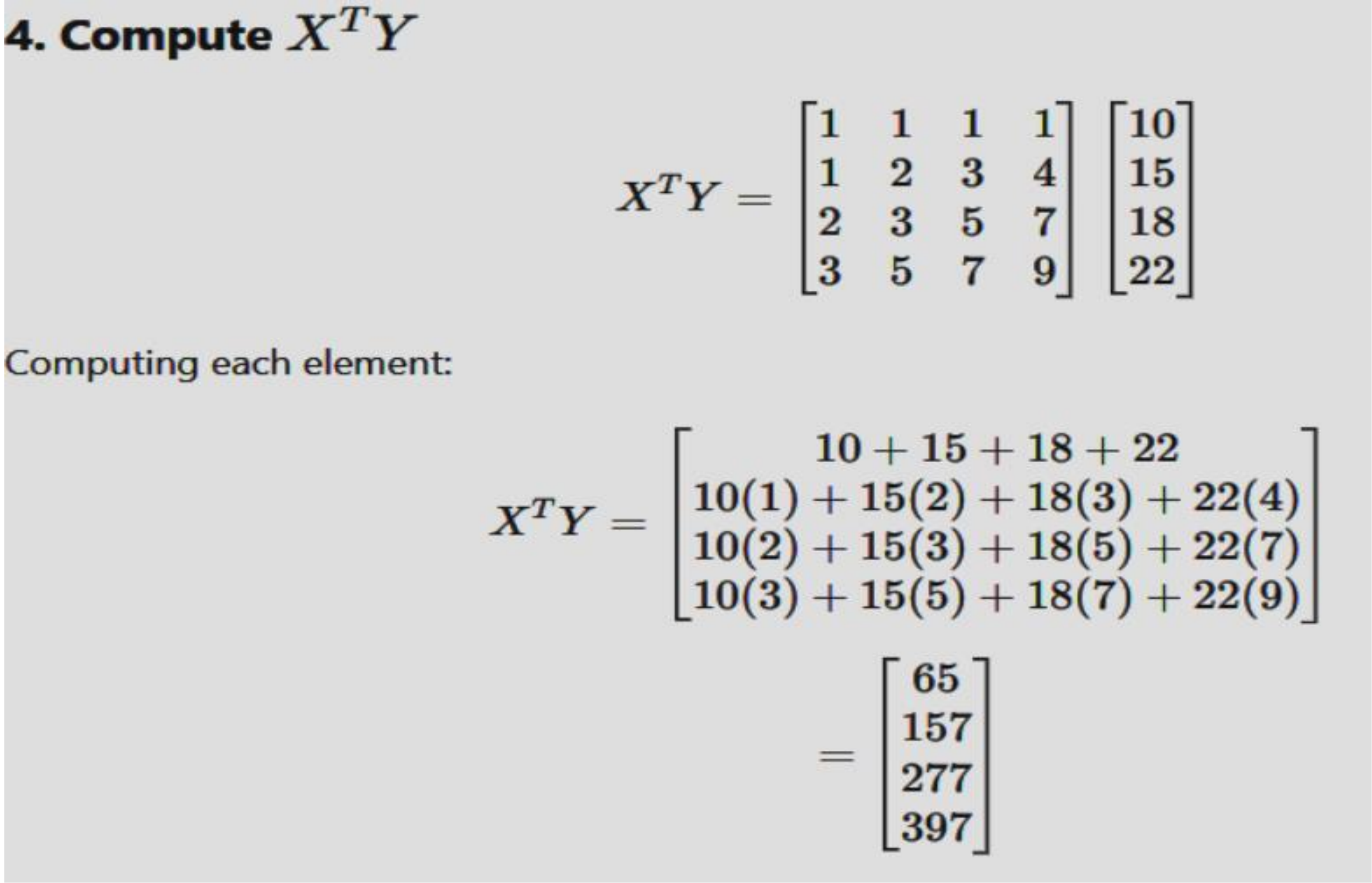
3. Compute $X^T X$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 5 & 7 \\ 3 & 5 & 7 & 9 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 3 & 5 \\ 1 & 3 & 5 & 7 \\ 1 & 4 & 7 & 9 \end{bmatrix}$$

Computing each element:

$$X^T X = \begin{bmatrix} 4 & 10 & 17 & 24 \\ 10 & 30 & 50 & 70 \\ 17 & 50 & 87 & 124 \\ 24 & 70 & 124 & 180 \end{bmatrix}$$

Multiple Linear Regression



4. Compute X^TY

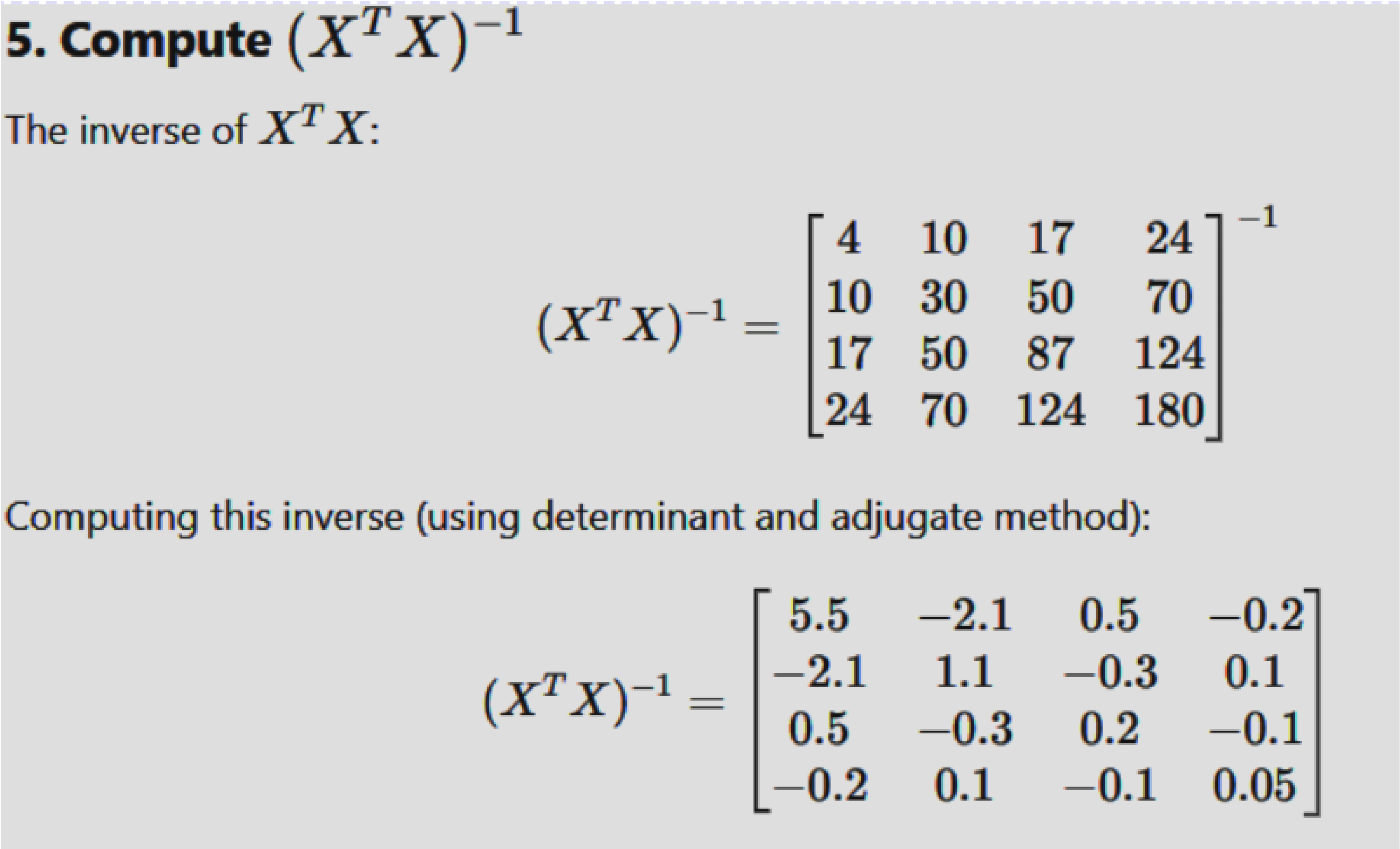
$$X^TY = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 2 & 3 & 5 & 7 \\ 3 & 5 & 7 & 9 \end{bmatrix} \begin{bmatrix} 10 \\ 15 \\ 18 \\ 22 \end{bmatrix}$$

Computing each element:

$$X^TY = \begin{bmatrix} 10 + 15 + 18 + 22 \\ 10(1) + 15(2) + 18(3) + 22(4) \\ 10(2) + 15(3) + 18(5) + 22(7) \\ 10(3) + 15(5) + 18(7) + 22(9) \end{bmatrix}$$

$$= \begin{bmatrix} 65 \\ 157 \\ 277 \\ 397 \end{bmatrix}$$

Multiple Linear Regression



5. Compute $(X^TX)^{-1}$

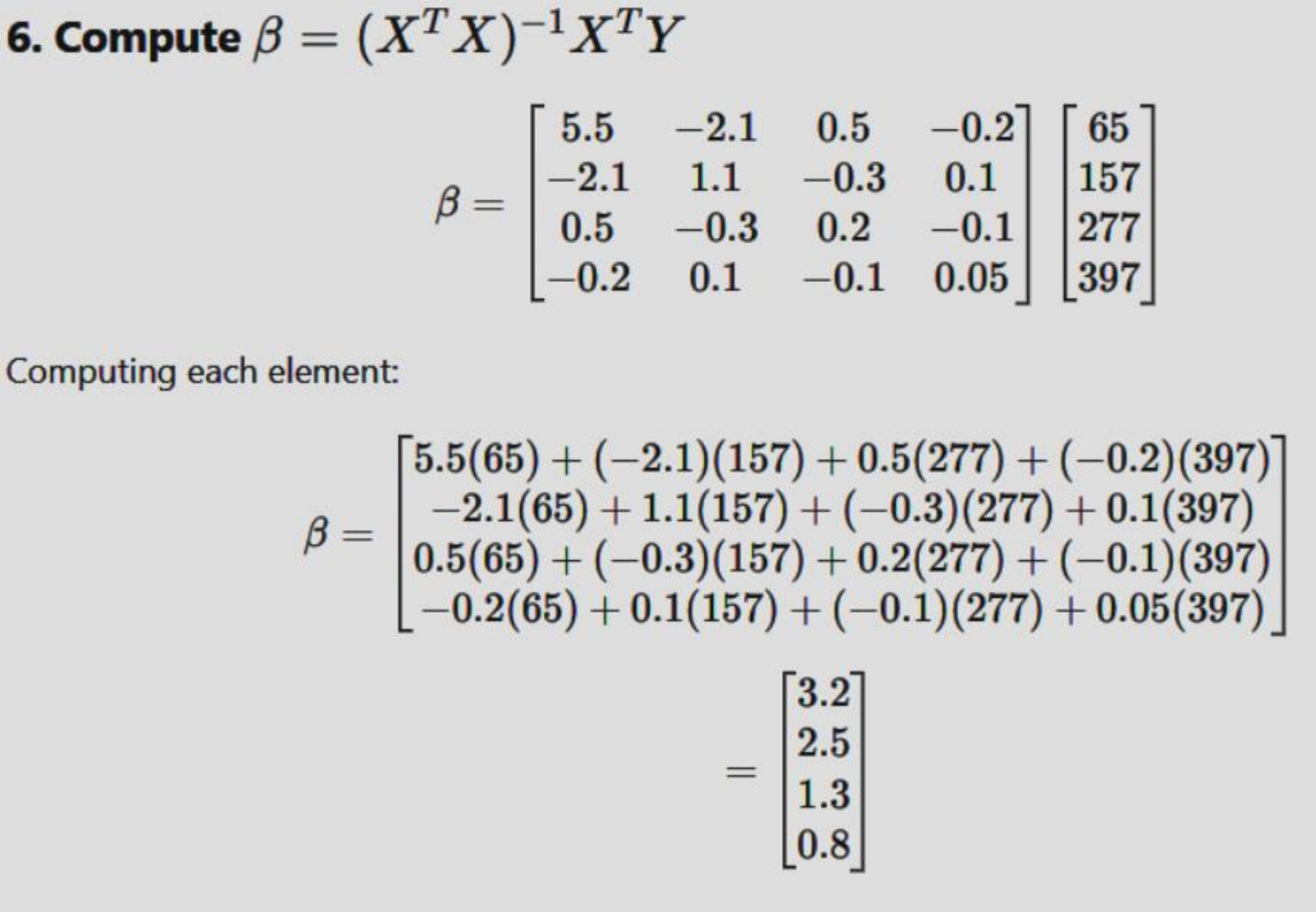
The inverse of X^TX:

$$(X^TX)^{-1} = \begin{bmatrix} 4 & 10 & 17 & 24 \\ 10 & 30 & 50 & 70 \\ 17 & 50 & 87 & 124 \\ 24 & 70 & 124 & 180 \end{bmatrix}^{-1}$$

Computing this inverse (using determinant and adjugate method):

$$(X^TX)^{-1} = \begin{bmatrix} 5.5 & -2.1 & 0.5 & -0.2 \\ -2.1 & 1.1 & -0.3 & 0.1 \\ 0.5 & -0.3 & 0.2 & -0.1 \\ -0.2 & 0.1 & -0.1 & 0.05 \end{bmatrix}$$

Multiple Linear Regression



6. Compute $\beta = (X^T X)^{-1} X^T Y$

$$\beta = \begin{bmatrix} 5.5 & -2.1 & 0.5 & -0.2 \\ -2.1 & 1.1 & -0.3 & 0.1 \\ 0.5 & -0.3 & 0.2 & -0.1 \\ -0.2 & 0.1 & -0.1 & 0.05 \end{bmatrix} \begin{bmatrix} 65 \\ 157 \\ 277 \\ 397 \end{bmatrix}$$

Computing each element:

$$\beta = \begin{bmatrix} 5.5(65) + (-2.1)(157) + 0.5(277) + (-0.2)(397) \\ -2.1(65) + 1.1(157) + (-0.3)(277) + 0.1(397) \\ 0.5(65) + (-0.3)(157) + 0.2(277) + (-0.1)(397) \\ -0.2(65) + 0.1(157) + (-0.1)(277) + 0.05(397) \end{bmatrix}$$

$$= \begin{bmatrix} 3.2 \\ 2.5 \\ 1.3 \\ 0.8 \end{bmatrix}$$

7. Final Regression Equation

$y=3.2+2.5x_1+1.3x_2+0.8x_3$

Linear Regression -Assumptions

- In linear regression, several **assumptions** need to hold true for the model to produce reliable and valid results.
- If these assumptions are violated, it can affect the accuracy of predictions and lead to biased or incorrect conclusions.
- Below are the key assumptions of **linear regression**:
- **1. Linearity**
  - **Assumption**: The relationship between the independent variables (x) and the dependent variable (y) should be linear. This means that the effect of each independent variable on the dependent variable is constant and additive.
  - **Why Important**: If the relationship is non-linear, linear regression will not accurately model the data, leading to misleading results.
  - **How to Check**: You can plot the residuals (errors) against the fitted values (predictions) to check for non-linearity. If the residuals show a pattern, the assumption may be violated.

Linear Regression -Assumptions

- **2. Independence of Errors**
- **Assumption**: The residuals (errors) should be independent of each other. This means that the error for one observation should not be correlated with the error for another observation.
- **Why Important**: If the errors are correlated (e.g., in time series data), it suggests that there is some unaccounted information in the model, leading to biased estimates of the coefficients.
- **How to Check**: You can check for autocorrelation (serial correlation) of residuals using the **Durbin-Watson test** or by plotting the residuals in time order.

Linear Regression -Assumptions

- **3. Homoscedasticity (Constant Variance of Errors)**
- **Assumption**: The variance of the residuals should be constant across all levels of the independent variables. This means that the spread of the residuals (errors) should not increase or decrease with the fitted values.
- **Why Important**: If the variance of the residuals is not constant (heteroscedasticity), it can lead to inefficient estimates and affect hypothesis testing (e.g., confidence intervals, p-values).
- **How to Check**: Plot the residuals against the fitted values. If the plot shows a pattern (like a funnel shape), heteroscedasticity might be present. You can also perform the **Breusch-Pagan test** or **White test**.

Linear Regression -Assumptions

- **4. Normality of Errors**
- **Assumption**: The residuals (errors) should be approximately normally distributed, especially for the purpose of conducting statistical inference (e.g., hypothesis testing, confidence intervals).
- **Why Important**: If the errors are not normally distributed, it can affect the validity of statistical tests (like t-tests for coefficients).
- **How to Check**: You can use **Q-Q plots** (quantile-quantile plots) to visually check normality, or perform tests like the **Shapiro-Wilk test** or **Anderson-Darling test**.

Linear Regression -Assumptions

- **5. No Multicollinearity**
- **Assumption**: The independent variables should not be highly correlated with each other. Multicollinearity occurs when one independent variable can be linearly predicted from others with a high degree of accuracy.
- **Why Important**: High multicollinearity makes it difficult to assess the individual effect of each independent variable on the dependent variable. It can also inflate the standard errors of the coefficients, making them unreliable.
- **How to Check**: You can calculate the **Variance Inflation Factor (VIF)** for each independent variable. If the VIF is greater than 10, it suggests high multicollinearity.

Linear Regression -Assumptions

- **6. No Influential Outliers**
- **Assumption**: There should be no influential outliers that disproportionately affect the model. Outliers can have a large influence on the regression coefficients, leading to biased or unreliable estimates.
- **Why Important**: Outliers can distort the regression line and affect predictions.
- **How to Check**: You can use **Cook's distance** or **Leverage statistics** to identify influential data points. Points with high leverage or large Cook's distance might be influencing the regression results.

Linear Regression -Assumptions

1. **Linearity**: The relationship between independent and dependent variables is linear.
2. **Independence of Errors**: Errors are independent.
3. **Homoscedasticity**: Errors have constant variance.
4. **Normality of Errors**: Errors are normally distributed.
5. **No Multicollinearity**: Independent variables are not highly correlated.
6. **No Influential Outliers**: No points disproportionately affecting the model.

Linear Regression -Assumptions

- **What Happens if Assumptions are Violated?**
- **Linearity**: If violated, linear regression will give biased predictions.
- **Independence of Errors**: If violated, it could lead to incorrect conclusions about the significance of coefficients.
- **Homoscedasticity**: If violated, the standard errors of the coefficients may be biased, leading to incorrect significance tests.
- **Normality of Errors**: If violated, the p-values and confidence intervals may not be reliable, especially for small datasets.
- **Multicollinearity**: If violated, it can make the coefficients unstable and hard to interpret.
- **Outliers**: If violated, it can distort the regression coefficients and predictions.

Linear Regression