Hello Everyone,

I hope you all are fine and doing well.

Linear Regression

線性迴歸

What is Linear Regression?

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (y) and one or more independent variables ($x_1, x_2,\ldots, x_p$). The primary goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple variables) that predicts the value of the dependent variable based on the independent variables.

線性迴歸是一種基本的統計和機器學習技術，用於對因變數 (y) 和一個或多個自變數 ($x_1, x_2,\ldots, x_p$) 之間的關係進行建模。線性迴歸的主要目標是找到基於自變數預測因變數值的最佳擬合線（或多變量情況下的超平面）

Linear regression assumes that there is a **linear** relationship between the dependent variable and the independent variables. This means that the change in the dependent variable is proportional to the change in the independent variables.

線性迴歸假設因變數和自變數之間存在線性關係。這意味著因變數的變化與自變數的變化成正比

Types of Linear Regression

1) Simple Linear Regression
2) Multiple Linear Regression

1) Simple Linear Regression

A. In simple linear regression, there is only **one independent variable.**

B. The relationship between the dependent variable y and the independent variable x is modeled by a straight line.

A. 在簡單線性迴歸中，只有**一個自變數**。
B. 因變數y和自變數x之間的關係以直線建模。

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

Where:

1) y is the dependent variable (what we are trying to predict),

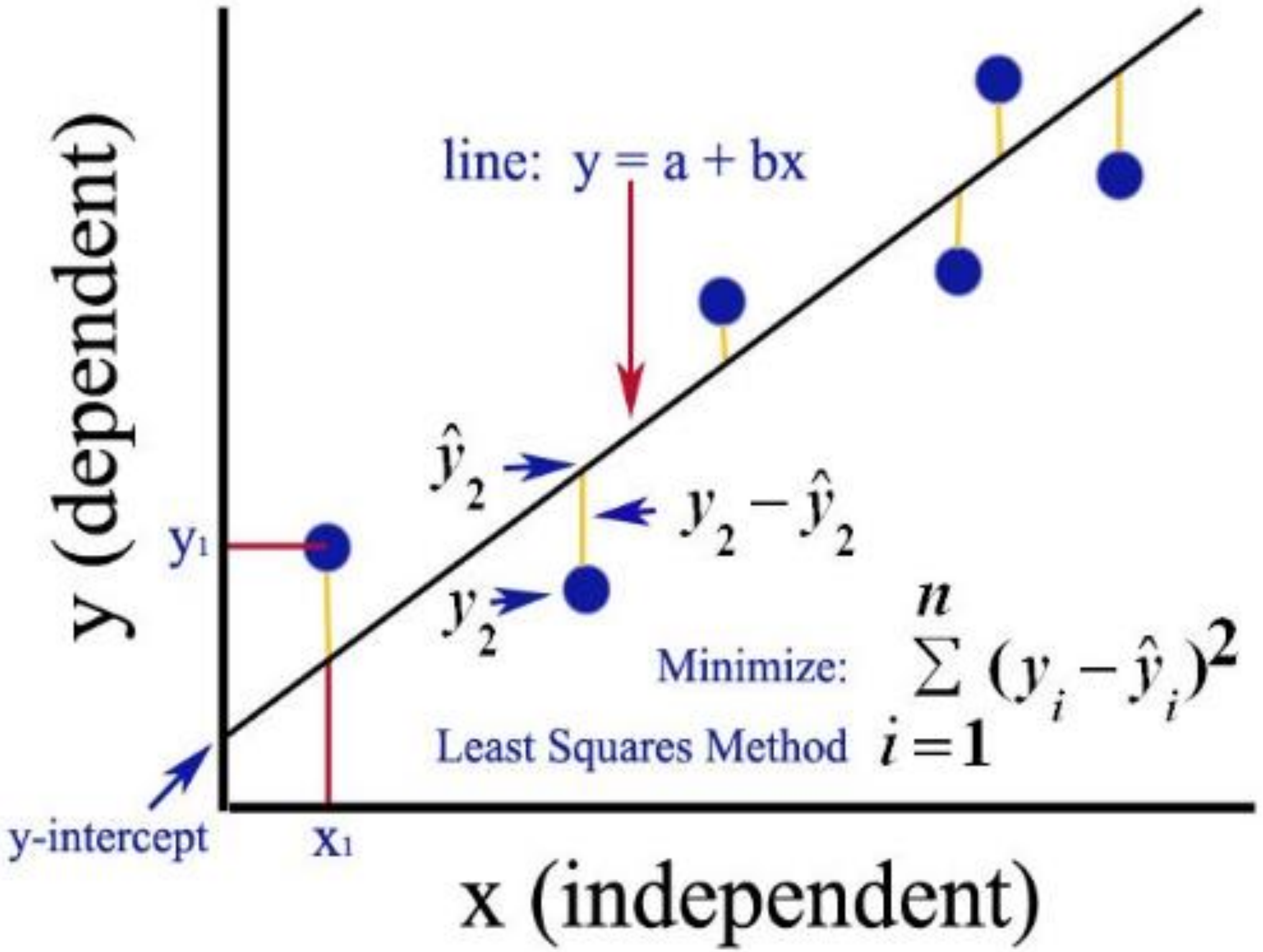
2) x is the independent variable (the predictor),

3) β_0 is the intercept (the value of y when $x = 0$),

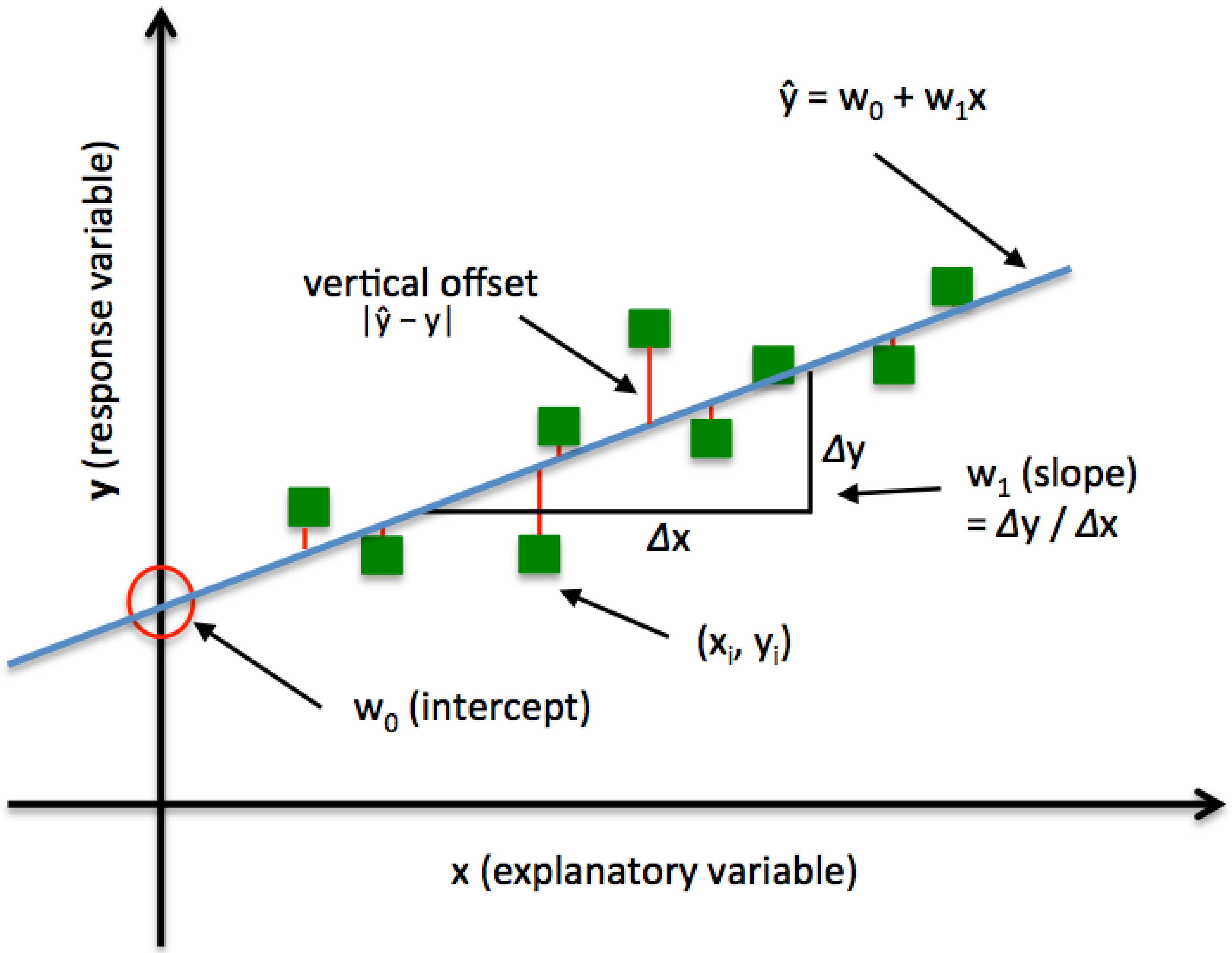
4) β_1 is the slope or coefficient, which shows how much y changes with a one-unit change in x,

5) ϵ is the error term (residual), which accounts for the variability in y that can't be explained by x.

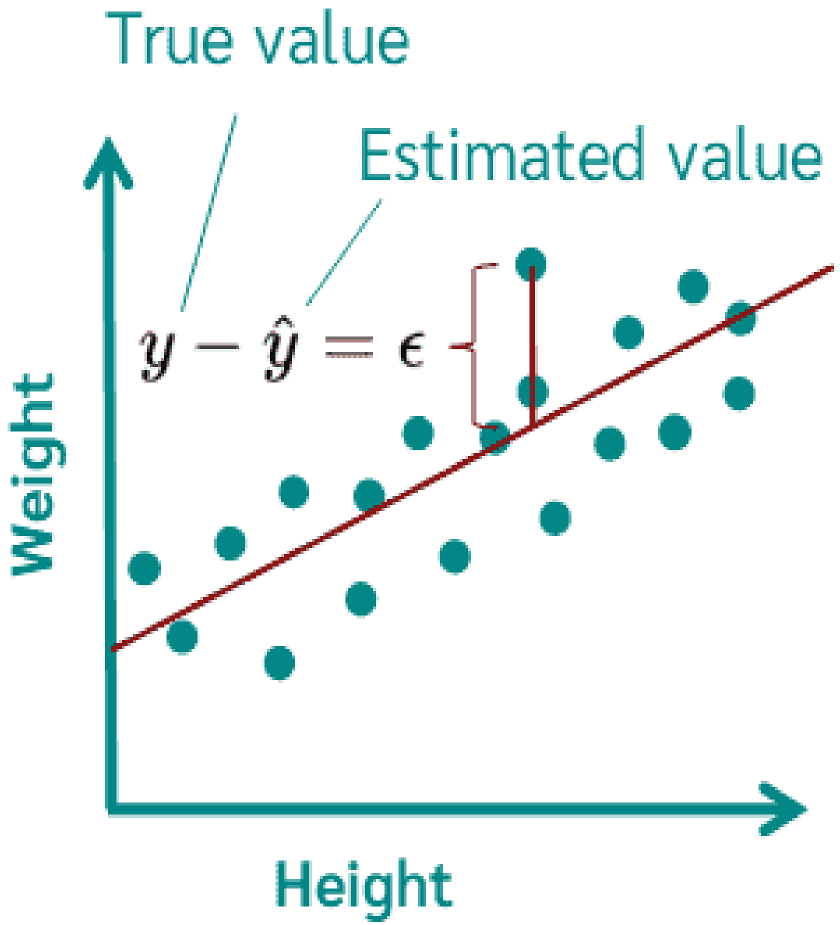
在哪裡：
1) y 是因變數（我們試圖預測的變數），
2) x 是自變數（預測變數），
3) β_0 為截距（當$x = 0$ 時 y 的值），
4) β_1是斜率或係數，表示x每變化一個單位，y會發生多少變化，
5) ϵ 是誤差項（殘差），它解釋了 y 中無法用 x 解釋的變異性



line: y = a + bx
$\hat{y}_2$
$y_2 - \hat{y}_2$
y₁
y_2
Minimize: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Least Squares Method
y-intercept
x₁
y (dependent)
x (independent)



$$y = \beta_0 = \mathrm{W}_0 + \beta_1 = \mathrm{W}_1 x + \epsilon = \mathbf{0}$$



Error epsilon

$$y = b \cdot x + a + \boxed{\epsilon}$$

2) Multiple Linear Regression

A. In multiple linear regression, there are **two or more independent variables.**

B. The relationship between the dependent variable and the independent variables is modeled as a linear equation in multiple dimensions.

A. 在多元線性迴歸中，有**兩個或多個自變數**。
B. 因變數和自變數之間的關係被建模為多維線性方程式。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Where:

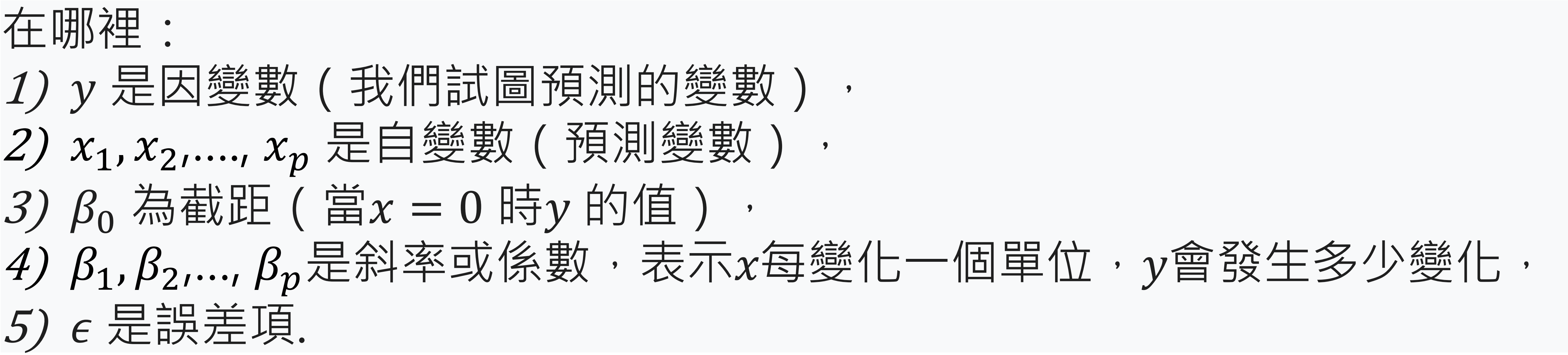
1) y is the dependent variable (what we are trying to predict),

2) $x_1, x_2, \ldots,$ x_p is the independent variables (the predictors),

3) β_0 is the intercept (the value of y when $x = 0$),

4) $\beta_1, \beta_2, \ldots,$ β_p is the slope or coefficient, which shows how much y changes with a one-unit change in x,

5) ϵ is the error term



在哪裡：
1) y 是因變數（我們試圖預測的變數），
2) $x_1, x_2, \ldots,$ x_p 是自變數（預測變數），
3) β_0 為截距（當$x = 0$ 時y 的值），
4) $\beta_1, \beta_2, \ldots,$ β_p是斜率或係數，表示x每變化一個單位，y會發生多少變化，
5) ϵ 是誤差項.

Simple Linear Regression

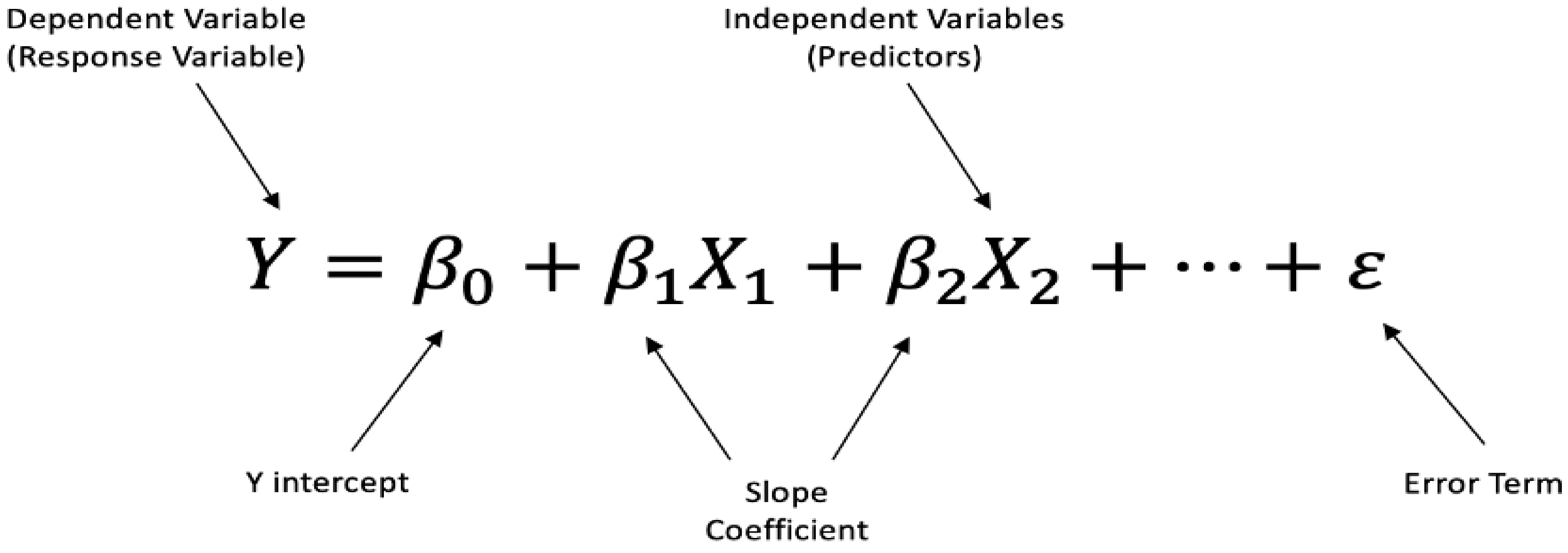
$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$



Dependent Variable (Response Variable)

Independent Variables (Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Y intercept

Slope Coefficient

Error Term

Mathematical Form of Linear Regression

Linear regression is a statistical method for modeling the relationship between a dependent variable y and one or more independent variables $x_1, x_2, \dots, x_p$. The goal is to find the best-fitting line (or hyperplane) that minimizes the difference between the predicted and actual values of the dependent variable.

線性迴歸是一種統計方法，用於對因變數 y 和一個或多個自變數 $x_1, x_2, \dots, x_p$ 之間的關係進行建模。目標是找到最小化因變數的預測值和實際值之間的差異的最佳擬合線 (或超平面)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Where:

1) y is the dependent variable (what we are trying to predict),

2) x_1, x_2,…., x_p is the independent variables (the predictors),

3) β_0 is the intercept (the value of y when $x = 0$),

4) β_1, β_2,…, β_p is the slope or coefficient, which shows how much y changes with a one-unit change in x,

5) ϵ is the error term

在哪裡：
1) y 是因變數（我們試圖預測的變數），
2) x_1, x_2,…., x_p 是自變數（預測變數），
3) β_0 為截距（當$x = 0$ 時y 的值），
4) β_1, β_2,…, β_p是斜率或係數，表示x每變化一個單位，y會發生多少變化，
5) ϵ 是誤差項.

In **simple linear regression**, where there is only one independent variable, the equation simplifies to:

在只有一個自變數的**簡單線性迴歸**中，方程式簡化為：

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

Where:

- β_0 is the intercept,
- β_1 is the slope (coefficient),
- x is the independent variable.

在哪裡：

- β_0 是截距，
- β_1 是斜率（係數），
- x 是自變數

Objective

The goal of linear regression is to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ such that the model minimizes the **sum of squared errors (SSE)**, also known as **least squares error**:

線性迴歸的目標是估計係數 $\beta_0, \beta_1, \ldots, \beta_p$，使模型最小化**誤差平方和 (SSE)**，也稱為**最小平方法誤差**：

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

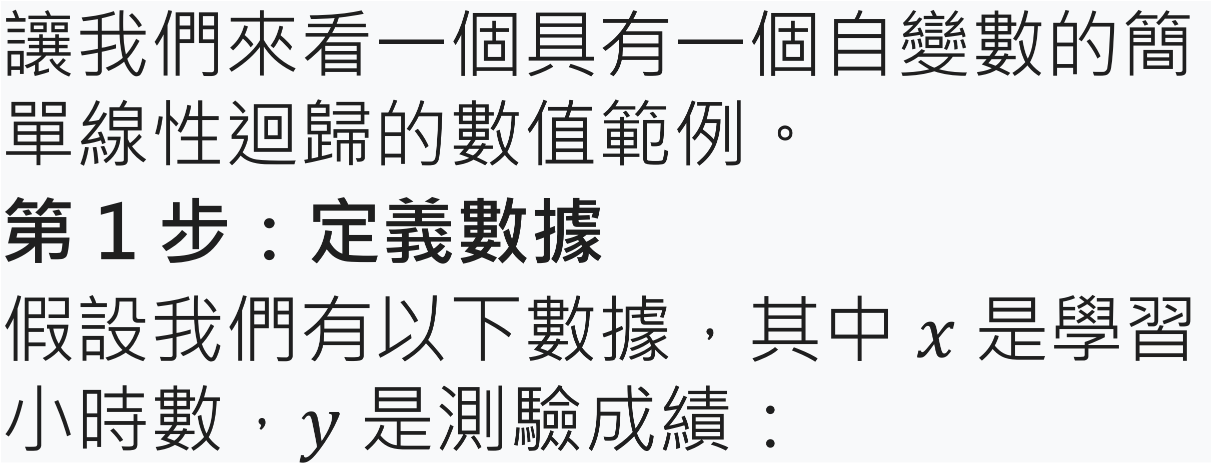
- Where:
- y_i is the actual value for the i-th data point,
- $\hat{y}_i$ is the predicted value for the i-th data point from the linear model.

Numerical Example of Simple Linear Regression

Let's work through a numerical example of **simple linear regression** with one independent variable.

Step 1: Define the Data

Suppose we have the following data, where x is the number of hours studied, and y is the test score:



讓我們來看一個具有一個自變數的簡單線性迴歸的數值範例。

第 1 步：定義數據

假設我們有以下數據，其中 x 是學習小時數，y 是測驗成績：

Hours Studied x	Test Score y
1	2
2	4
3	5
4	4
5	5

We want to find the line of best fit (i.e., the regression line) to predict y (test score) based on x (hours studied).

Step 2: Linear Regression Equation

The general equation for simple linear regression is:

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

Where:

- β_0 is the intercept,
- β_1 is the slope (the change in y for a unit change in x).

步驟2：線性迴歸方程

簡單線性迴歸的一般方程式為：

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

在哪裡：

- β_0 是截距，
- β_1 是斜率（ y 的變化對應於 x 的單位變化）

Step 3: Calculate the Slope ($\boldsymbol{\beta_1}$) and Intercept ($\boldsymbol{\beta_0}$)

The formulas for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ are:

$$\boldsymbol{\beta_1} = \frac{\boldsymbol{n \sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n}(x_i) \sum_{i=1}^{n}(y_i)}}{\boldsymbol{n \sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n}(x_i)\right)^2}}$$

$$\boldsymbol{\beta_0} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i}) - \boldsymbol{\beta_1} \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i})$$

Where n is the number of data points.

Calculate the necessary sums:

Sum of x values:

$$\sum_{i=1}^{5} x_i = 1 + 2 + 3 + 4 + 5 = 15$$

Sum of y values:

$$\sum_{i=1}^{5} y_i = 2 + 4 + 5 + 4 + 5 = 20$$

Sum of $\boldsymbol{x.y}$ values:

$$\sum_{i=1}^{5} x_i y_i = (1.2) + (2.4) + (3.5) + (4.4) + (5.5)$$

$$= 2 + 8 + 15 + 16 + 25 = 66$$

Sum of $\boldsymbol{x^2}$ values:

$$\sum_{i=1}^{5} x_i^2 = (1^2) + (2^2) + (3^2) + (4^2) + (5^2)$$

$$= 1 + 4 + 9 + 16 + 25 = 55$$

Now, we can calculate β_1 and β_0.

Step 4: Calculate the slope $\boldsymbol{\beta_1}$

$$\beta_1 = \frac{5(66) - 15(20)}{5(55) - (15)^2} = \frac{330 - 300}{275 - 225} = \frac{30}{50} = 0.6$$

Step 5: Calculate the slope $\boldsymbol{\beta_0}$

$$\beta_0 = \frac{1}{5}(20) - 0.6\left(\frac{1}{5}\right)(15) = 4 - (0.6)(3) = 4 - 1.8 = 2.2$$

Thus, the linear regression equation is:

$$y = 2.2 + 0.6x$$

Step 6: Make Predictions

Now, we can use this equation to make predictions. For example, if a student studies for 6 hours, the predicted test score would be:

步驟6：做出預測

現在，我們可以使用這個方程式來進行預測。例如，如果學生學習6小時，則預測的考試成績為：

$$y = 2.2 + 0.6(6) = 2.2 + 3.6 = 5.8$$

Conclusion

The **simple linear regression** formula is derived by estimating the best-fit line through the data points, using methods like least squares estimation. The process can be extended to **multiple linear regression** when there are multiple independent variables, and the goal is to predict a continuous outcome based on these features.