Simple Linear Regression

Dr. Mrityunjoy Barman, **mrityunjoybarman@soa.ac.in**
December 19, 2025

Introduction to Simple Linear Regression i

- Simple linear regression is used to estimate the relationship between a **predictor variable** (x) and a **response variable** (y).
- It provides a linear approximation of how y changes as x changes.
- **Example:** Estimating the nutritional rating of cereals based on their sugar content.

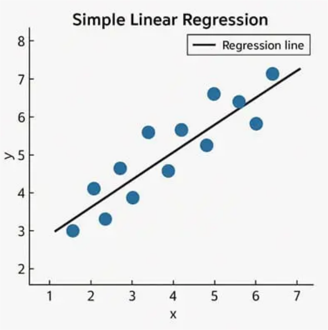

Figure 1: Data fitting with a straight line.

The Regression Equation

Let us consider the given data as:

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Table 1: The given data is given as $(x_i,\ y_i)$.

The estimated regression line is defined by

$$\hat{y} = b_0 + b_1 x \tag{1}$$

- $\hat{y}$: The estimated value of the response variable.
- b_0: The y-intercept (estimated value of y when $x = 0$).
- b_1: The slope (estimated change in y per unit increase in x).
- b_0 and b_1 are called the **regression coefficients**.

Method of Normal Equations

If the data points $(x_i,\ y_i)$ were lying on the regression line (1), then we will have

$$\hat{y}_i = b_0 + b_1 x_i$$

for all $i = 1, 2, ..., n$.

This can be viewed as

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}, \quad \text{which gives } Pq = Q. \tag{2}$$

Multiplying by P^T bothsides we get, $P^T Pq = P^T b$. This equations are known as the normal equations.

Now we can solve for the unknown vector $q = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$.

The goal is to minimize the **Sum of Squared Errors (SSE)**:

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

By differentiating with respect to b_0 and b_0 and setting to zero, we derive:

$$\frac{\partial}{\partial b_0}(SSE) = 0, \quad \frac{\partial}{\partial b_1}(SSE) = 0$$
$$\implies \sum y_i = nb_0 + b_1 \sum x_i, \quad \text{and} \sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2.$$

Slope Estimate (b_1)

$$b_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{cov(X, Y)}{var(X)}$$

.

Intercept Estimate (b_0)

$$b_0 = \bar{y} - b_1 \bar{x}$$

.

Example 1

Example
Consider the following data:

X	-1	1	2
Y	1	1	3

Table 2: The given data is given as $(x_i,\ y_i)$.

Here, we see $E(X) = 2/3,\ E(X^2) = 2,\ E(Y) = 5/3,\ E(XY) = 2$, and hence

$$cov(X, Y) = E(XY) - E(X)E(Y) = 2 - 10/9 = 8/9,$$
$$\text{and } var(X) = E(X^2) - E(X)^2 = 2 - 4/9 = 14/9,$$
$$\implies b_1 = \frac{cov(X, Y)}{var(X)} = \frac{8/9}{14/9} = 4/7.$$

Similary, $b_0 = \bar{y} - b_1\bar{x} = 5/3 - 4/7 \times 2/3 = 9/7$.

Thus, the regression line is given by $\hat{y} = 9/7 + 4/7x$.

Example 2: Calculation of the SSE i

The SSE represents the overall measure of prediction error. Below is the calculation for 10 competitors using $\hat{y} = 6 + 2x$.

Subject	**Time (x)**	**Distance (y)**	**Predicted ($\hat{y}$)**	**Residual** $(y - \hat{y})$	$(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	-1	1
5	4	14	14	0	0
6	5	15	16	-1	1
7	6	20	18	2	4
8	7	18	20	-2	4
9	8	22	22	0	0
10	9	25	24	1	1

Table 3: Data from Table 8.3

Sum of Squares Error (SSE):

$$SSE = \sum (y - \hat{y})^2 = 0 + 1 + 0 + 1 + 0 + 1 + 4 + 4 + 0 + 1 = 12.$$

- **Sum of Squared Error** $SSE = \sum(y - \hat{y})^2$.
- **Sum of Squared Total** $SST = \sum(y - \bar{y})^2$
- **Sum of Squared Regression** $SSR = SST - SSE$.
- We also call the constants $b_0, \ b_1$ as the regression coefficients.

The **Coefficient of Determination (r^2)** measures the proportion of variability in y explained by the regression.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

- $SST = SSR + SSE$.
- r^2 ranges from 0 to 1.
- Values near 1 indicate an extremely good fit.

Example 3: Projected Score

Example

Suppose, in a T20 match between India and South Africa, the progress of runs scored in India innings are given as follows:

Over	4	8	12	16
Run	33	68	115	150

(a) Find the esyimated linear regression line to the above data.

(b) What is the projected score at the end of the India innings?

(c) What are various types of errors in this estimation?

(d) How good were the above data fit by the regression line? Explain using the coefficient of determination.

Standard Error of the Estimate (s): Correlation Coefficient (r)

The s statistic measures the ”typical” residual size (precision).

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n - m - 1}}$$

The **Pearson correlation coefficient** (r) measures the strength and direction of the linear relationship.

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{\sum x^2 - (\sum x)^2/n}\sqrt{\sum y^2 - (\sum y)^2/n}} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

- Range: $[-1, 1]$.
- Positive r: y increases as x increases.
- Negative r: y decreases as x increases.
- $r = \pm\sqrt{r^2}$ (sign depends on the slope b_1).