# Linear Regression

---

## 1. Introduction to Linear Regression

- In many real-world situations, there's a **dependent variable** Y whose value depends on **independent variables** $x_1, x_2, \dots, x_r$.
- **Ideal linear relation**:

  $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$

  where:

  - $\beta_0, \beta_1, \dots, \beta_r$ = regression coefficients (unknown constants).
- **Reality**:

  There will always be some **random error eee**.

  So the real model is:

  $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$

  where e has mean 0.
- ➡️ This is called the **Linear Regression Equation**.

---

## 2. Estimating the Regression Line (Least Squares)

- Suppose you observe pairs $(x_i, Y_i)$ for i = 1, ..., n.
- Want to find estimates A and B for $\alpha$ and $\beta$ that minimize the **Sum of Squared Errors (SS)**:

  $SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$
- **Goal**:

  Find A, B that minimize SS → called the **Method of Least Squares**.

### Finding A and B (Normal Equations)

To minimize SS, take partial derivatives and set them to zero:

$\frac{\partial SS}{\partial A} = 0 \quad \text{and} \quad \frac{\partial SS}{\partial B} = 0$

This gives two **normal equations**:

$\sum Y_i = nA + B \sum x_i$

$\sum x_i Y_i = A \sum x_i + B \sum x_i^2$

From these, solving gives:

### ➡️ Estimators:

$$B = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2} \text{ and } A = \bar{Y} - B\bar{x}$$

where:

- $\bar{x}$ = mean of $x_i$
- $\bar{Y}$ = mean of $Y_i$

✅ So you first find B (the slope), then A (the intercept).

---

# 3. Final Form of Estimated Regression Line

The fitted line is:

$\hat{Y} = A + Bx$

This is the **estimated relationship** between x and y.

---

# 4. Distribution of Estimators A and B

## Assumptions for Distribution

To talk about their distribution, we need some assumptions:

- The errors $e_i$ (random deviations) are:
  - **Independent** (errors at different points don't affect each other),
  - **Normally distributed** (bell curve shaped),
  - **Mean 0**, **Variance** $\sigma^2$ (constant across all observations).

That is:

$e_i \sim N(0, \sigma^2)$

Thus:

$Y_i = \alpha + \beta x_i + e_i \quad \Rightarrow \quad Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

✅ This is called the **classical linear regression model** assumptions.

# What Happens to B?

From the least squares formula, B is calculated as:

$$B = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2}$$

Notice:

- B is a **linear combination** of the $Y_i$'s.
- Since the $Y_i$'s are **normal**, any linear combination of normal variables is **also normal**.

Thus, B is **normally distributed**.

## Mean and Variance of B

- **Mean of B**:

  $E[B] = \beta$

  ( B is an **unbiased** estimator of the true slope $\beta$).

- **Variance of B**:

  $\mathrm{Var}(B) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$

  (Variance depends on how spread out the $x_i$ values are — more spread gives smaller variance.)

# What Happens to A?

From the formulas:

$A = \bar{Y} - B\bar{x}$

where $\bar{Y}$ is the mean of $Y_i$.

Since B is normal and $\bar{Y}$ is normal, A (being a combination of them) is also **normally distributed**.

- **Mean of A**:

  $E[A] = \alpha$

  (✅ A is an **unbiased** estimator of the true intercept $\alpha$).

- **Variance of A**:

  $\mathrm{Var}(A) = \sigma^2 \left( \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right)$

---

# 5. Residuals and Estimating $\sigma^2$

- **Residual** = The difference between the **actual observed value** $Y_i$ and the **predicted value** $\hat{Y}_i$ from your regression line.

Mathematically:

$$\text{Residual}_i = Y_i - (A + Bx_i)$$

where:

- $Y_i$ = actual value,
- $A + Bx_i$ = predicted value based on your regression line.

✅ Residuals tell you **how much your line is "off"** at each data point.

## Sum of Squares of Residuals (SSR)

- To measure the **overall error** across all points, we **square each residual** and **add them up**:

$SSR = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$

- This SSR gives the **total squared error** between your data and your fitted line.

✅ Think of it like "how badly" your line misses the data — **smaller SSR = better fit**.

## Why Estimate $\sigma^2$?

- $\sigma^2$ represents the **true variance** of the errors $e_i$ in your model.

- **In reality**, we don't know $\sigma^2$, because we don't know the true errors.
- So, we **estimate** it using the residuals from the fitted model.

## How to Estimate $\sigma^2$?

From theory:

- It can be shown that:

$\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$
(That is, it follows a **Chi-squared distribution** with n−2 degrees of freedom.)

Because we lose **2 degrees of freedom**:

- 1 for estimating the intercept A,
- 1 for estimating the slope B.

Thus, the **unbiased estimator** of $\sigma^2$ is:
$\hat{\sigma}^2 = \frac{SSE}{n-2}$
✅ So you:

- Calculate the residuals,
- Find SSE,
- Divide SSE by n−2 to get $\hat{\sigma}^2$.

---

| Symbol | Full Form | Formula | Use |
|---|---|---|---|
| SSR | Sum of Squares due to Regression | $\sum(\hat{y}_i - \bar{y})^2$ | Variation **explained** by regression |
| SSE | Sum of Squares of Errors (Residuals) | $\sum(y_i - \hat{y}_i)^2$ | Variation **not explained** by regression |
| SST | Total Sum of Squares | $\sum(y_i - \bar{y})^2$ | Total variation in data |
| $\sigma^2$ | Estimate of error variance (MSE) | $\frac{SSE}{n-2}$ | Used for inference (e.g., standard errors, confidence intervals) |