**Linear Regression**

Linear Regression is a method used to find a linear relationship between independent variable(s) (X) and a dependent variable (Y).

Formula

Y=b0​+b1​X1​+b2​X2​+...+bn​Xn​+ϵ

| **Symbol** | **Meaning** | **Analogy** |
| --- | --- | --- |
| **Y** | Dependent variable (the value you’re predicting) | e.g., Life Expectancy |
| **X** | Independent variable (the input) | e.g., GDP |
| **b₀** | Intercept → value of Y when X = 0 | Like “c” in y = mx + c |
| **b₁** | Slope → how much Y changes when X increases by 1 unit | Like “m” in y = mx + c |
| **ε** | Error term → difference between actual and predicted Y | The leftover or noise |

B0 = Intercept(c)

B1 = the slope(m)

| Type | Description |
| --- | --- |
| Simple Linear Regression | One independent variable |
| Multiple Linear Regression | More than one independent variable |
| Polynomial Linear Regression | Uses higher powers of X (like X², X³) to model curved relationships |

y=mX+c

m = coefficient(s) → how much Y changes when X changes

c = intercept → value of Y when X = 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**MSE (Mean Squared Error):** Measures average of squared prediction errors — lower is better.

**RMSE (Root Mean Squared Error):** Average error in actual units — shows how far predictions are from real values.

**MAE (Mean Absolute Error):** Average absolute difference between predicted and actual values.

**R² (R-Square):** Shows how much of the variation in the target is explained by the model.

**Adjusted R²:** R² adjusted for number of predictors — penalizes adding useless variables.

Adjusted R square = 1 - (1-R2) \* (N-1) / N - p -1

### **Where:**

| **Symbol** | **Meaning** |
| --- | --- |
| **R²** | Normal R-squared value |
| **N** | Total number of observations (rows) |
| **p** | Number of independent variables (features/predictors) |

## **💬 Why we need Adjusted R²**

You said it perfectly —  
 **R² always increases when you add more independent variables**,  
 even if those variables are **not actually useful** in prediction.

👉 So, R² can **mislead you** — it’ll look like your model is improving,  
 but in reality, the model might just be becoming more complex and overfitted.

## **⚙️ How Adjusted R² fixes that**

* Adjusted R² **penalizes unnecessary variables** (via the term *p*).
* If you add a new variable that **actually helps**, Adjusted R² will **increase**.
* If you add a variable that **doesn’t help**, Adjusted R² will **decrease**.

This makes it a **better, fairer measure** of model performance when multiple predictors are used.

| **Variable** | **R²** | **Adjusted R²** |
| --- | --- | --- |
| 1 Predictor | 0.75 | 0.74 |
| 3 Predictors | 0.80 | 0.77 |
| 6 Predictors | 0.82 | 0.76 |

👉 Notice: even though R² keeps rising,

Adjusted R² starts dropping because not all predictors are useful.

## **💡 In short:**

| **Metric** | **Meaning** | **Problem / Solution** |
| --- | --- | --- |
| **R²** | % of variation in Y explained by X | Always increases with new variables |
| **Adjusted R²** | Penalizes unnecessary predictors | Only increases if the new variable adds real value |

### **🧩 Case 1 – Few useful variables**

| **Model** | **Independent Variables (p)** | **What they represent** | **R²** | **Adjusted R²** |
| --- | --- | --- | --- | --- |
| Model 1 | 1 → engine\_size | Bigger engine → higher price | 0.70 | 0.69 |
| Model 2 | 2 → engine\_size, mileage | Mileage also affects price | 0.82 | 0.81 |
| Model 3 | 3 → engine\_size, mileage, brand\_rating | Brand also important | 0.88 | 0.87 |

✅ Here every new p (variable) **adds real information**,  
 so **R² ↑** and **Adjusted R² ↑**.  
 The model truly got better.

### **🧩 Case 2 – Adding useless variables**

Now you start adding random columns like the color of the dashboard,  
 number of cup holders, or serial number.

| **Model** | **Independent Variables (p)** | **Are they useful?** | **R²** | **Adjusted R²** |
| --- | --- | --- | --- | --- |
| Model 4 | + dashboard\_color | ❌ No relation to price | 0.885 | 0.86 |
| Model 5 | + cup\_holders, serial\_number | ❌ Still no relation | 0.89 | 0.84 |

😬 R² **keeps increasing a little** (because adding any variable can always fit the data a tiny bit more),  
 but **Adjusted R² drops** — it’s punishing you for adding useless features.

When we add more independent variables (**p increases**),  
 the denominator N - p - 1 **decreases**.  
 As a result, the fraction value **increases**.

Since this entire fraction is **subtracted from 1**,  
 the overall Adjusted R² **decreases** — unless the new variable actually improves R² a lot.

This means Adjusted R² **penalizes** the model for adding too many variables that don’t truly help in prediction.

Case 2 = If p decreases denominator N - p - 1 increases and the fraction gets smaller, hence R square goes up.

### **🎯 In short:**

| **Change in p** | **Effect on Denominator** | **Effect on Fraction** | **Effect on Adjusted R²** |
| --- | --- | --- | --- |
| **p increases** | Denominator ↓ | Fraction ↑ | Adjusted R² ↓ |
| **p decreases** | Denominator ↑ | Fraction ↓ | Adjusted R² ↑ |