Why Do We Use R^2 and How?

What is R^2 (R-Square)?

- R^2, or the **coefficient of determination**, is a statistical measure that explains how well the independent variable(s) explain the variability of the dependent variable in a regression model.
- It always lies between 0 and 1.

Why Do We Use R^2?

1. **Goodness of Fit** → It tells us how well our regression line fits the observed data.
   - $R^2 = 0 \rightarrow$ The model explains none of the variability.
   - $R^2 = 1 \rightarrow$ The model explains all of the variability perfectly.
2. **Model Comparison** → Higher R^2 generally indicates a better fit when comparing models (but not always).
3. **Decision Making** → It helps decide whether the regression model is useful for prediction.

How is R^2 Calculated?

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- $SS_{res} = \sum(y_i - \hat{y_i})^2 \rightarrow$ **Residual Sum of Squares** (error not explained by the model).
- $SS_{tot} = \sum(y_i - \bar{y})^2 \rightarrow$ **Total Sum of Squares** (total variability in the data).

Thus, R^2 measures the fraction of total variance in the dependent variable y that is explained by the model.

Example (Intuition)

- Suppose you are predicting **house prices** using **size of the house**.
- If $R^2 = 0.85$, it means 85% of the variability in house prices is explained by house size, and the remaining 15% is due to other factors not included in the model.

Important Note

- A high R^2 does not always mean a good model (it can be misleading with too many predictors $\rightarrow$ overfitting).
- In multiple regression, we often use **Adjusted** R^2 because it penalizes unnecessary predictors.