Regression Model Evaluation: A Deep Dive into R-Values and Fit

R-Squared (R^2) – Coefficient of Determination

R-Squared (R^2) is a statistical measure that explains how well the independent variable(s) predict the dependent variable in a regression model. It represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).

Formula:

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

Where:

- SS_{res} (Residual Sum of Squares) = $\sum(Y_i - \hat{Y}_i)^2$ → Measures unexplained variance (errors).
- SS_{tot} (Total Sum of Squares) = $\sum(Y_i - \bar{Y})^2$ → Measures total variance in Y.
- Y_i = Actual values, $\hat{Y}i$ = Predicted values, $\bar{Y}$ = Mean of actual values.

Key Properties of R^2:

1. **Ranges from 0 to 1:**
   - $R^2 = 1$ → Perfect model (100% variance explained).
   - $R^2 = 0$ → Model does not explain variance.
   - Can be **negative** if the model performs worse than just predicting the mean $\bar{Y}$
2. **Higher R^2 Means Better Fit:**

- A higher value indicates that the model explains more variability in the data.

3. **Does Not Detect Overfitting:**
   - A high R^2 does **not** guarantee a good model, as it does not consider the number of predictors.

Why Use R^2?

✅ **Easy Interpretation** → It shows how much of the variation in Y is explained by X.

✅ **Compares Different Models** → Helps compare the goodness-of-fit for different regression models.

✅ **Works Well for Simple Linear Regression** → Provides a clear measure of how well the independent variable explains the dependent variable.

When NOT to Use R^2?

❌ **For Non-Linear Models** → R^2 assumes a linear relationship.

❌ **When You Need Adjusted R^2 for Multiple Predictors** → Adjusted R^2 accounts for extra predictors.

Adjusted R-Squared (R^2_{adj})

Adjusted R-Squared is an improved version of **R-Squared (R^2)** that accounts for the number of predictors in a regression model. It adjusts for overfitting by penalizing the addition of unnecessary independent variables.

Formula:

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(N - 1)}{N - k - 1} \right)$$

where:

- N = Total number of observations (data points)
- k = Number of independent variables (predictors)
- R^2 = Regular R-Squared value

Key Differences Between R^2 and Adjusted R^2

Feature	R-Squared (R^2)	Adjusted R-Squared (R^2_{adj})
Formula Accounts for Predictors?	❌ No, increases with more variables	✅ Yes, penalizes unnecessary variables
Overfitting Risk?	✅ Higher risk	❌ Lower risk
Value Can Decrease?	❌ No, always increases with more predictors	✅ Yes, if an added variable does not improve the model

Why Use Adjusted R^2?

✅ **Prevents Overfitting** → Avoids misleading high R^2 values when unnecessary variables are added.

✅ **Better for Multiple Regression** → More reliable when working with multiple predictors.

✅ **Compares Models Fairly** → Helps choose the best model without bias toward complexity.

When NOT to Use Adjusted R^2?

❌ **For Simple Linear Regression** → Adjusted R^2 is unnecessary when there's only one predictor.

❌ **If Model Selection Uses Other Criteria (AIC/BIC)** → Some methods use different evaluation metrics for model selection.

What is Multiple R?

Multiple R (also known as the **multiple correlation coefficient**) is a statistical measure used in **multiple regression analysis** to indicate how well the independent variables collectively predict the dependent variable.

Formula for Multiple R

In multiple regression, the equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Multiple R is the correlation between the observed values of Y and the predicted values ($\hat{Y}$) from the regression model:

$$R = \text{Corr}(Y, \hat{Y})$$

Interpretation of Multiple R

- **R = 1** → Perfect positive correlation (model predicts perfectly).
- **R = 0** → No correlation (model predictions have no relationship with actual values).
- **R = -1** → Perfect negative correlation (very rare in regression).

Higher R values indicate that the independent variables together explain a strong relationship with the dependent variable.

Difference Between Multiple R and R^2

Metric	Definition	Interpretation
Multiple R	Correlation between actual Y and predicted $\hat{Y}$	Measures the strength of the relationship between dependent and independent variables
R^2 (R-Squared)	Proportion of variance in Y explained by independent variables	Shows how well the model explains the variability in the data

📌 **Note:**

- **Multiple R** is just the **square root of R^2** (but it does not indicate the direction of the relationship).
- Unlike R^2, **Multiple R doesn't measure the proportion of variance explained.**

Explaining Multiple R and R-Squared in Simple Words

1️⃣ Multiple R (Correlation Between Actual and Predicted Values)

Think of **Multiple R** as a measure of **how strong the relationship** is between the independent variables (inputs) and the dependent variable (output).

🔷 **If Multiple R is close to 1**, it means the model's predictions are very close to the actual values.

🔷 **If Multiple R is close to 0**, it means the model is bad at predicting the target.

💡 **Example:**

Imagine you're trying to predict someone's weight based on height and age. If **Multiple R = 0.9**, it means height and age **strongly** predict weight. But if **Multiple R = 0.2**, height and age are **weak predictors** of weight.

2️⃣ R-Squared (How Well the Model Explains the Data)

R-squared (R^2) tells you **how much of the variation in the target variable is explained by the model**.

🔷 **If R^2 = 0.8** → The model explains **80% of the variations** in the data.

🔷 **If R^2 = 0.3** → The model explains **only 30%**, so it's not very reliable.

💡 **Example:**

- If we are predicting house prices and R^2 **= 0.85**, it means **85% of the price variation is explained by factors like location, size, etc.**
- If R^2 **= 0.2**, it means **80% of price changes are due to unknown or missing factors**.

Key Difference

Metric	Meaning	Good Value?
Multiple R	Measures how strong the overall relationship is between independent and dependent variables	Closer to **1** is better
R-Squared (R^2)	Measures how well the model explains the variation in the target variable	Higher R^2 (closer to 1) is better

✅ **Multiple R is just the correlation, while R-Squared tells how much of the data the model explains.**
