# R² Score (Coefficient of Determination) in Detail

The R² score, also called the coefficient of determination, measures how well a regression model explains the variability of the target variable. It tells us how much of the variation in the dependent variable (y\_actual) is captured by the independent variable (X) using the given model.

## Why Do We Need R² Score?

Imagine you're predicting house prices using features like square footage, number of bedrooms, location, etc.

If R² = 0.85, your model explains 85% of the variability in house prices.

If R² = 0.2, your model only explains 20% of the variability, meaning it’s not very good.

If R² = 0, your model is no better than simply taking the average of all house prices.

Thus, higher R² means a better fit, but it doesn’t always mean the best model.

## Formula for R² Score

R² = 1 − (SS\_residual / SS\_total)

where:

SS\_residual = ∑(y\_actual − y\_predicted)² → Sum of squared residuals (errors)

SS\_total = ∑(y\_actual − y\_mean)² → Total variance in actual values

## Step-by-Step Calculation:

1️⃣ Compute the mean of the actual values:

y\_mean = ∑ y\_actual / n

2️⃣ Compute SS\_total:

SS\_total = ∑(y\_actual − y\_mean)²

3️⃣ Compute SS\_residual:

SS\_residual = ∑(y\_actual − y\_predicted)²

4️⃣ Compute R² using the formula:

R² = 1 − (SS\_residual / SS\_total)

## How to Interpret R² Score?

R² Value Interpretation

1.0 → Perfect model (explains 100% of variance) ✅

0.8 - 0.9 → Strong predictive power 👍

0.5 - 0.7 → Moderate predictive power 🤔

0.0 - 0.4 → Weak model 😕

< 0 → Worse than a simple mean prediction ❌

## R² Score in Python (Example)

Let's calculate R² Score using sklearn:

from sklearn.metrics import r2\_score  
  
# Actual vs Predicted values  
y\_actual = [3, -0.5, 2, 7]   
y\_pred = [2.5, 0.0, 2, 8]  
  
# Calculate R² Score  
r2 = r2\_score(y\_actual, y\_pred)  
print("R² Score:", r2)

## When R² Might Be Misleading

1️⃣ Adding more features always increases R² (even if features are useless).

2️⃣ Nonlinear relationships: If data follows a curved pattern, a linear model may give low R², but a better model exists.

3️⃣ Presence of outliers: Outliers can distort the R² value.

✅ Solution: Use Adjusted R² when comparing models with different features.

## Adjusted R² (For Multiple Features)

Why Adjusted R²?

R² always increases when adding more features, even if they don't improve the model.

Adjusted R² fixes this by penalizing unnecessary features.

Formula for Adjusted R²:

R²\_adjusted = 1 − [(1 − R²) × (n−1) / (n−p−1)]

where:

n = number of samples

p = number of features

## When to Use R² Score?

✅ Good for:

- Evaluating linear regression models.

- Comparing models with the same dataset.

- Understanding how well the model explains the data.

❌ Avoid when:

- Working with nonlinear data (R² assumes linearity).

- Comparing models with different datasets (Adjusted R² is better).

- Data has outliers (use robust error metrics like MAE or RMSE).

## Summary: R² vs Adjusted R²

|  |  |  |
| --- | --- | --- |
| Metric | When to Use | Limitations |
| R² | Quick model evaluation | Can be misleading if too many features are added |
| Adjusted R² | When comparing models with different features | Slightly harder to interpret |

## Final Takeaway

🚀 If R² is high (~0.8+), your model is good.

🚀 If R² is low (<0.4), the model needs improvement.

🚀 If R² is negative, it's worse than a random guess!

🚀 Use Adjusted R² for multiple features.