R-Squared (Coefficient of Determination)

R-squared, also known as the Coefficient of Determination (R^2), is a statistical metric used to evaluate the performance of regression models. It represents the proportion of the variance in the dependent variable (y) that is predictable from the independent variables (X). R-squared provides insight into the goodness-of-fit of a model.

Formula

The formula for R^2 is:

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

Where:

- $SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is the residual sum of squares, representing the total error of the model,
- $SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares, representing the total variance in the actual data,
- y_i is the actual value,
- $\hat{y}_i$ is the predicted value,
- $\bar{y}$ is the mean of the actual values.

Alternatively, in terms of correlation:

$$R^2 = (r_{xy})^2$$

Where, r_{xy} is the Pearson correlation coefficient between the actual and predicted values.

Characteristics of R-Squared

1. **Range of Values:**
   - R^2 ranges from 0 to 1, Where:
     - $R^2 = 1$: Perfect fit (model explains 100% of the variance in y),
     - $R^2 = 0$: The model explains none of the variance in y (equivalent to predicting $\bar{y}$ for all observations),
     - Negative R^2: Indicates that the model performs worse than simply predicting the mean of the actual values.
2. **Interpretation:**
   - An R^2 value of 0.8 means that 80% of the variance in y is explained by the model, and 20% is unexplained.

Adjusted R-Squared

Adjusted R-Squared modifies R^2 to account for the number of predictors in the model. It penalizes the addition of irrelevant predictors and is calculated as:

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Where:

- n is the number of observations,
- p is the number of predictors.

Key Difference:

- Adjusted R^2 decreases if irrelevant predictors are added, whereas R^2 always increases, even with irrelevant predictors.

Advantages

1. **Intuitive Measure of Fit:**
    - Provides a clear indication of how well the model explains the variation in the target variable.
2. **Useful for Comparing Models:**
    - Allows comparison of different models' fit on the same dataset.
3. **Simplicity:**
    - Easily interpretable, making it accessible for non-technical stakeholders.

Disadvantages

1. **Insensitive to Overfitting:**
    - R^2 increases as predictors are added, even if they do not contribute to the model's performance.
2. **Not Always Indicative of Predictive Power:**
    - A high R^2 does not guarantee that the model will generalize well to unseen data.
3. **Scale-Dependent:**
    - R^2 depends on the scale of the target variable, making it unsuitable for comparing models across datasets with different scales.
4. **Cannot Detect Bias:**
    - R^2 does not indicate if the model is biased or has systematic prediction errors.

When to Use R-Squared

1. **Explaining Variance:**
    - When the goal is to quantify how much of the variation in the target variable is explained by the predictors.
2. **Model Comparison:**
    - To compare the goodness-of-fit of different regression models on the same dataset.
3. **Feature Evaluation:**
    - To assess the relevance of features in explaining variance.

Comparison with Other Metrics

1. **R-Squared vs. Adjusted R-Squared:**
    - Adjusted R^2 is better for multiple regression as it accounts for the number of predictors, preventing overfitting.
2. **R-Squared vs. RMSE/MSE/MAE:**
    - R^2 evaluates the proportion of explained variance, while RMSE, MSE, and MAE quantify prediction errors directly.
    - R^2 is more interpretable for understanding overall model fit, whereas error metrics provide direct measures of accuracy.
3. **R-Squared vs. Log-Loss/AUC (for Classification):**
    - R^2 is specific to regression tasks, while metrics like Log-Loss or AUC are used for classification.

Example Calculation

Suppose we have the following actual (y_i) and predicted ($\hat{y}_i$) values:

- Actual: [3, -0.5, 2, 7]
- Predicted: [2.5, 0.0, 2, 8]

1. Calculate $\bar{y}$ (mean of actual values):

$$\bar{y} = \frac{3 + (-0.5) + 2 + 7}{4} = 2.875$$

2. Calculate SS_{tot} (total sum of squares):

$$SS_{tot} = (3 - 2.875)^2 + (-0.5 - 2.875)^2 + (2 - 2.875)^2 + (7 - 2.875)^2 = 29.6875$$

3. Calculate SS_{res} (residual sum of squares):

$$SS_{res} = (3 - 2.5)^2 + (-0.5 - 0)^2 + (2 - 2)^2 + (7 - 8)^2 = 1.5$$

4. Calculate R^2:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{1.5}{29.6875} \approx 0.9495$$

Interpretation

An R^2 value of 0.9495 indicates that the model explains approximately 94.95% of the variance in the target variable, which suggests a very good fit.

Use Cases

1. **Linear Regression:**
   - To measure how well the predictors explain the variance in the dependent variable.
2. **Model Validation:**
   - To evaluate the fit of a regression model before testing its predictive power.
3. **Feature Selection:**
   - To determine the contribution of specific features to the explained variance.

R-squared is an essential metric for evaluating regression models, but it should be used alongside other metrics to ensure a comprehensive understanding of model performance.