Adjusted R-Squared

Adjusted R-squared is a modified version of R-squared (R^2) that accounts for the number of predictors in a regression model. Unlike R^2, which always increases with the addition of more predictors (regardless of their relevance), Adjusted R-squared penalizes the inclusion of irrelevant variables. This makes it a more reliable measure of model fit, especially for models with multiple predictors.

Formula

The formula for Adjusted R-squared is:

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

Where:

- R^2 is the standard coefficient of determination,
- n is the number of observations (data points),
- p is the number of predictors (independent variables).

Key Concepts

1. **Adjusts for Number of Predictors:**
   - Adjusted R^2 introduces a penalty for adding more predictors, ensuring that only significant variables improve the score.
2. **Comparison Across Models:**
   - Helps compare models with different numbers of predictors to avoid overfitting.
3. **Ranges:**
   - Like R^2, Adjusted R^2 can range from 0 to 1, with higher values indicating a better fit. It can also be negative if the model performs worse than a horizontal line at the mean ($\bar{y}$).

Characteristics

1. **Penalization:**
   - Adjusted R^2 decreases if a new predictor does not significantly improve the model fit.
   - It increases only if the predictor reduces the residual sum of squares (SS_{res}) more than expected by chance.

2. **Handles Overfitting:**
   - Unlike R^2, Adjusted R^2 discourages overfitting by ensuring irrelevant variables do not artificially inflate the score.
3. **Dependence on Sample Size:**
   - For small datasets, Adjusted R^2 may over-penalize, potentially undervaluing the importance of predictors.

Advantages

1. **Accounts for Model Complexity:**
   - Adjusted R^2 balances model complexity and goodness-of-fit, favoring parsimonious models with fewer but meaningful predictors.
2. **Useful for Feature Selection:**
   - Helps identify and retain only those predictors that significantly improve model performance.
3. **Model Comparison:**
   - Enables comparison of regression models with different numbers of predictors on the same dataset.

Disadvantages

1. **Interpretation Complexity:**
   - Adjusted R^2 is less intuitive to interpret than R^2.
2. **Dependence on Sample Size:**
   - With small sample sizes, Adjusted R^2 may penalize predictors too heavily.
3. **Not for Non-Linear Models:**
   - Adjusted R^2 is primarily used for linear regression models and may not capture the complexity of non-linear relationships.

When to Use Adjusted R-Squared

1. **Multiple Regression Models:**
   - When the model includes many predictors, and you want to account for their relevance.
2. **Feature Selection:**
   - During feature selection processes to identify significant variables and avoid overfitting.
3. **Comparing Models:**
   - To compare regression models with varying numbers of predictors to identify the best fit.

Comparison with R-Squared

Aspect	R-Squared (R^2)	Adjusted R-Squared (R^2_{adj})
Effect of Predictors	Always increases as predictors are added	Penalizes for irrelevant predictors
Overfitting	Does not account for overfitting	Accounts for overfitting
Purpose	Measures variance explained by the model	Measures variance explained, adjusted for predictors

---

Example Calculation

Suppose we have a regression model with the following characteristics:

- $R^2 = 0.8$,
- $n = 100$ (number of observations),
- $p = 5$ (number of predictors).

The Adjusted R^2 is calculated as:

$$R^2_{adj} = 1 - \frac{(1 - 0.8)(100 - 1)}{100 - 5 - 1}$$

$$R^2_{adj} = 1 - \frac{(0.2)(99)}{94} = 1 - 0.2106 \approx 0.7894$$

Interpretation

An Adjusted R^2 of 0.7894 indicates that approximately 78.94% of the variance in the dependent variable is explained by the predictors, accounting for their number. This score reflects a slightly reduced value compared to R^2, penalizing the complexity of the model.

Use Cases

1. **Regression Models:**
   - Evaluating and improving multiple linear regression models.
2. **Feature Selection:**

- Identifying the optimal number of predictors during model development.

3. **Model Validation:**
   - Ensuring the model generalizes well to unseen data by preventing overfitting.

Adjusted R^2 is an essential metric for regression analysis, providing a balanced measure of model performance while discouraging overfitting. It is particularly useful when dealing with complex models involving multiple predictors.