Regression Measure (Metric) - R^2

Definition: R^2 is a statistical measure that represents the proportion of the variance for in dependent variable that explained by an independent variable or variable in a regression model.

Formula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where, RSS = Residual Sum of Squares
TSS = Total Sum of Square.

Steps for Calculation

x (Independent Variable)	Y (Dependent Variable)
1	2
2	3
3	5
4	4
5	6

Calculate the mean of the dependent variable (Y):

$$\bar{Y} = \frac{2+3+5+4+6}{5} = 4$$

Perform a simple linear regression to find the line of best fit:

The equation of the line is typically

$$Y = a + bX$$

For simplicity, assume we've already performed the regression analysis and found the best fit line equation to be:

$$\bar{Y} = 1.4 + 0.8X$$

Calculate the predicted values ($\hat{Y}$) using the regression equation:

$$\hat{Y}_1 = 1.4 + 0.8 \times 1 = 2.2$$
$$\hat{Y}_2 = 1.4 + 0.8 \times 2 = 3.0$$
$$\hat{Y}_3 = 1.4 + 0.8 \times 3 = 3.8$$
$$\hat{Y}_4 = 1.4 + 0.8 \times 4 = 4.6$$
$$\hat{Y}_5 = 1.4 + 0.8 \times 5 = 5.4$$

Calculate the Total Sum of Square (TSS):

$$TSS = \Sigma(y_i - \bar{y})^2$$

$$TSS = (2-4)^2 + (3-4)^2 + (5-4)^2 + (4-4)^2 + (6-4)^2$$
$$= 4 + 1 + 1 + 0 + 4 = 10$$

Calculate the Residual Sum of Square (RSS):

$$RSS = \Sigma(y_i - \hat{y_i})^2$$

$$RSS = (2-2.2)^2 + (3-3.0)^2 + (5-3.8)^2 + (4-4.6)^2 + (6-5.4)^2$$
$$= (-0.2)^2 + (0)^2 + (1.2)^2 + (-0.6)^2 + (0.6)^2$$
$$= 0.04 + 1.44 + 0.36 + 0.36$$
$$= 2.2$$

Calculate the R^2 value.

$$R^2 = 1 - \frac{RSS}{TSS}$$
$$= 1 - \frac{2.2}{10}$$
$$= 1 - 0.22$$
$$= 0.78$$

Interpretation

An R^2 value of 0.78 means that 78% of the variance in the dependent variable (Y) can be explained by the independent variable (X) using the regression model.

Regression Measure (Metric) – Adjusted R^2

Defination: Adjusted R-square (Adjusted R^2) adjust the R-square value for the no. of prediction in the model. This adjustment is important because R-square can be artificially high when more predicters are added to the model, even if those predictors do not improve the model's explantory power.

Formula of for Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-R)^2 . (n-1)}{n-k-1}\right)$$

where,

- R^2 is the coefficient of the determination.
- n is the no. of observation.
- K is the no. of predictors (independent variables)

Steps for calculation

X (independent Variable)	Y (Dependent Variable)
1	2
2	3
3	5
4	4
5	6

we calculated,
$R^2 = 0.78$
$n = 5$
$K = 1$

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1-0.78)(5-1)}{5-1-1}\right)$$

$$= 1 - \left(\frac{(0.22) . 4}{3}\right)$$

$$= 1 - \left(\frac{0.88}{3}\right)$$

$$= 1 - 0.2933 = 0.7067$$

An Adjusted R^2 value of approximately 0.71 means that around 71% of the variance in the dependant variable (Y) can be explained by the independent variable (X) when adjusted for the number of predictors in the model.

Difference between R^2 and Adjusted R^2

1. Penalizes for Adding Non-significant Predictors

R-square (R^2) always increases or stays the same when more predictors are added to a model, regardless of wheather the predictors are significant. This can lead to a misleadingly high R^2 value for models with many predictors that don't actually improve the model's explanatory power.

Adjusted R^2, on the other hand, adjusts for the no. of predictors in the model. It penalizes the addition of non-significant predictors by decreasing unless the new predictor improves the model more than would be expected by chance. This ensures that the model is only rewarded for adding predictors that provide genuine explantery power.

2. Provides a More Accurate Measure of Model Fit

Adjusted R^2 gives a more accurate measure of how well the model fits the data by taking into account the complexity of the model. It is especially useful when comparing models with different numbers of ~~pr~~ predictors. A higher adjusted R^2 indicates a better model, balancing goodness of fit with model complexity.

③ Helps in Model selection:

- When building regression models, one often needs to compare models with different no. of predictors to determine which model is best. Adjusted R^2 provides a more reliable metric for comparision because it accounts for the no. of predictors. A model with a higher adjusted R^2 is generally preferred as it suggests a better fit while avoiding over fitting