Adjusted R- Squared

Adjusted R- squared is a modified version of R- squared (R^2) that adjusted for the number of predictor variable in a multiple regression model. It provides a more accurate measure of the goodness-of-fit of a model by considering the model's complexity.

In a multiple regression model, R-squared (R^2) measured the propotion of variance in the response variable that is explained by the predictor variables. However, R-Squared always increase as stays the same with the addition of new predictor variables, regardless of whether those variables contributes ~~variable~~ valuable information to the model. This can lead to overfitting, where a model becomes too complex and starts capturing noise in the data instead of the underlying relationships.

Adjusted R- squared accounts for the number of predictor variable in the model and the sample size, penalizing the model for adding unnecessary complexity. Adjusted R- squared can decrease when an irrelevant predictor variable is added to the model, making a better metric for comparing models with different numbers of predictor variables.

The formula for adjusted R-squared is:

$$\boxed{\text{Adjusted } R^2 = \left[1 - \frac{(1-R^2) * (n-1)}{(n-k-1)}\right]}$$

Where:

- R^2 is the R-squared of the model
- n is the number of observations in the dataset
- k is the number of predictor variables in the model.

By using adjusted R-squared, you can more accurately assess the goodness-of-fit a model and choose the optimal set of predictor variable for your analysis.

Which One Should be used?

The choice between using R-squared and adjusted R-squared on the content and the goals of your analysis. Here are some guidlines to help you decide which one to use:

1. **Model Comparison :** If you're comparing models with different numbers of predictor variables, it's to use adjusted R-squared. This is because adjusted R-squared takes into account the complexity of the models, penalizing models that include irrelevant predictor variables. R-squared, on the other hand, can be misleading in this context, as it trends to increase with the addition of more predictor variables, even if they don't contribute valuable information to the ~~squa~~model.

2. ~~Ihh~~ **Model Interpretation :-** If you're interested in understanding the propotion of variance in the response variable that can be explained by the predictor variables in the model, R-squared can be useful metric. However, keep in mind that R-squared does not provide information about the significance or relevance of individual predictor variables. It's also important to remember that a high-R-squared value does not

necessarily imply causation or a good predictive model.

3. **Model Selection and Overfitting** : When building a model and selecting predicting ~~varieties~~ variables, it's important to guard against overfitting. In this context, adjusted R-squared can be a helpful metric, as it accounts for the number of predictor variables and penalizes the model for unnecessary complexity. By using adjusted R-squared, you can avoid including irrelevant predictor variables that might lead to overfitting.

In summary, adjusted R-squared is generally more suitable when comparing models with different number of predictor variables or when you're concerned about overfitting. R-squared can be useful for understanding the overall explanatory power of the models but it should be interpreted with caution, especially in cases with ~~many~~ many predictor variables or potential multicollinearity.

T- Statistic

Performing a t-test for a simple linear regression including the Intercept term and using the p-value approch, involves the following steps:

1. State the null and alternative hypothesis for the slope and intercept coefficient:

For the slope coefficient ($\beta 1$):

- Null Hypothesis (H0): $\beta 1 = 0$ (no relationships beth the predictor variable (X) and the response variable (Y)).

- Alternative hypothesis (H1): $\beta 1 \neq 0$ (a relationship exists between the predictor variable and the response variable)

For the Intercept coefficient ($\beta 0$):

- Null Hypothesis (H0): $\beta 0 = 0$ (the regression line passes through the origin)

- Alternative hypothesis (H1): $\beta 0 \neq 0$ (the regression

2. Estimate the slope and intercept coefficient (b_0 and b_1):

Using the sample data, calculate the slope (b1) and intercept (b0) coefficients for the regression model.

3. Calculate the standard errors for the slope and intercept coefficient (SE (b0) and SE(b1)):

Compute the standard errors of the slope and intercept coefficient using the following formula:

$b_1 \to \beta_1$

$$SE(b_1) = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{(n-2)\,\Sigma (x_i - \bar{x})^2}}$$

$b_0 \to \beta_0$

$$SE(b_0) = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma (x_i - \bar{x})^2}\right]}$$

4. Compute the t-statistic for the slope and intercept coefficients:

Calculate the t-statistics for the slope and intercept coefficient using the following formula:-

$$t\text{-value } b_0 = \frac{b_0 - 0}{SE(b_0)}$$

Simple LR

x | y

$\beta_1 = 0$

$\beta_1 \neq 0$

estimate (slope)
(sample mean)

population (slope)
(mean)

$$t\text{-statistic} = \frac{\beta_1 - 0}{SE(\beta_1)}$$

$$SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x})^2}}$$

$$t\text{-statistic} = \frac{\beta_0 - 0}{SE(\beta_0)}$$

$$SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}$$

* Find P-value using t-statistic and degree of freedom (no. of feature or colmn) in regression

* Significance level generally = 0.05

* if P-value < 0.05
  Now, reject null Hypothesis ($\beta_1 = 0$)

* if P-value > 0.05
  Not reject null hypothesis ($\beta_1 = 0$)

Confidence Intervals Of Coeffients

1. **Estimate the slope and intercept coefficients (b_0 and b_1)** : Using the sample data, calculate the slope (b_1) and intercept (b_0) coefficients for the regression model.

2. **Calculate the standard errors for the slope and intercept coefficients ($SE(b_0)$ and $SE(b_1)$)**:

$$\boxed{SE(b_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)\sum (x_i - \bar{x}_i)^2}}}$$

$$\boxed{SE(b_0) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-2)} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}}$$

3. **Determine the degrees of freedom** : In a sample linear regression, the degrees of freedom (df) is equal to the number of observation (n) minus the number of estimated parameter (2: the intercept and the slope coefficient).

$df = n-2$

4. **Find the critical t-value :** Look up the critical t-value from the t- distribution table or use a statistical calculator based on the chosen confidence level (eg., 95%) and the degrees of the freedom calculated in step 3.

5. **Calculate the confidence intervals for the slope and intercept coefficients :**

$$\boxed{CI_{b_0} = b_0 \pm \text{t-value} * SE(b_0)}$$

find with significance value (0.05) and degree of freedom.

$$\boxed{CI_{b_1} = b_1 \pm \text{t-value} * SE(b_1)}$$

These confidence intervals ~~expost~~ represent the range within which the true population regression coefficients are likely to fall with a specified level of confidence (eg. 95%)