R squared and P value

1. What is R squared?

R squared referred sometimes as coefficient of determination is a measure to calculate how much percentage of the total variation in dependent variable (Y) is described by the line

1.1. Calculation

Intuitively, total variation in (Y) can be calculated by using an estimate of the center point $\bar{Y}$.

To calculate total variation (TV):

1. Find the mean $\bar{Y}$of Y.
2. Sum the squared distances from each data point to $\bar{Y}$

Sum of Squared errors (SSE) is a value that explains how much variance in the data is not explained by the line

The proportion of SSE and TV explains how much percentage of the total variation in Y is **not** described by the line

Thus, $1 - \frac{SSE}{TV}$ explains what we are looking for

This formula is also called as R^2

If the line is a good fit, it will have low SSE and high R^2 value.

Similarly for high SSE, R^2 value will be low.

2. What is P value?

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

3. Backward Elimination

Starting with a model that uses all the independent variables and iteratively eliminating independent variables (based on p-value or r^2 value) till we reach parsimony is backward elimination.

Parsimony or a parsimonious model in this context is defined as a model that accomplishes desired level of prediction with as few independent variables as possible.