**Feature selection** is a process where you automatically select those features in your data that contribute most to the prediction variable. Feature selection reduces overfitting problem, improves accuracy & reduces training time.

P values evaluate how well the sample data support the argument that the null hypothesis is true. In case of feature selection, the null hypothesis is something like *this feature contains no information about the prediction target*, so the higher scores indicates null hypothesis is true that is predator does not contain much information and can be eliminated.

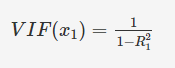
**p-value** is a number between 0 and 1 and interpreted in the following way:

* A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis, implies that a particular feature is significant
* A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis, implies that a particular feature is insignificant
* p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).

**Multicollinearity** refers to a situation where multiple predictor variables are correlated with each other. Since multiple variables are involved.

**VIF** can be a good metrics to look at to tackle multicollinearity between variables. The VIF value can range from 1 to any higher value which indicates that if the particular variable is taken into the model how much it is contributing to the multicollinearity.

For example, there are multiple input variables x1,x2,x3 VIF value of x1 can be calculated as



If the value of R1is high, such as 0.90, implies VIF(x1) = 5.26 (which implies that x1 can be predicted using the other predictors), it is advisable to remove it from the model to avoid making the model unnecessarily complex.

**Eliminate Features** based on the combination of VIF & p-value

**High VIF means we can predict that particular feature using other predators.**

**High p-value (>0.5) means null hypothesis that feature is insignificant is true.**

**First eliminate those which are having high VIF & high p-value**