Empirical Study

1. Statistical tests:

We are studying and building a bug prediction model, with dependent variable being presence of bug ( if bug>0 : 1, else : 0) , hence we need a statistical test for binary variables.

We are using logistic regression as we have only two outputs, i.e. binary.

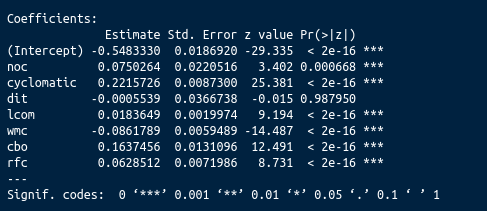
2. Variables with high correaltion: \*\*(format these and reffer to the heatplots and enter individual pairs and their correlations)\*\*

* cbo vs [rfc,lcom,dit]
* rfc vs dit

e.g. pairs (cbo vs rfc : corr)

\*\*( add regression plots only for the ones we mention here and for any one of the software)\*\*

3.Regression with all independent variables in the model:

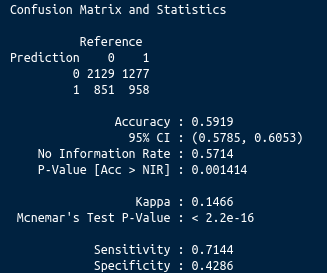
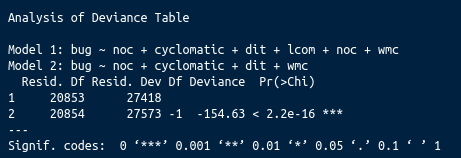


We use P-values to make conclusions in significance testing. More specifically, we compare the P-value to a significance level  to make conclusions about our hypotheses.

The signficance Codes above, display levels of signficance. We accept the variables with \*\*\* as they execute a P-value between (0.000, 0.001).

Hence we can discard ‘**dit**’ as it’s significance is **0.98**, which shows that it does not contribute much to our model.

4. Hypothesis:



Null hypothesis : Ck metrics does not contribute to bug prediction.

Alternative hypothesis : Ck metrics contributes to bug prediction.

* The basic model without hyper-parameter tuning has a 60% prediction accuracy, with specificity of 42% and sensitivity of 71%, shows that the model is not overfitted.
* After eliminating insignificant variables we see that the model 2 is significant compared to model 1.

We reject the null hypothesis.