Implementation: Does class size matter? An in-depth assessment of the effect of class size in software defect prediction

CS22S504 Rajrupa Chattaraj

February 20, 2023

1 Research questions and Methods implemented

The goal of the work[Tahir et al.(2021)]is to understand indirect effect of class size in defect prediction models across both count and binary models. To achieve this goal, we have investigated the following key research questions:

RQ1: If the class size has mediation effect on the relationships between object oriented metrics and the number of defects(count model)? In order to answer the first research question we have calculated direct effect, indirect effect and total effect by considering number of defects as the dependent variable, single OO metric as independent variable and class size as mediator and the below steps were followed:

direct path- Linear regression model is being considered where number of defects is derived by single OO metric and p value is being estimated.

indirect path- Linear regression model is being considered where the mediator variable(class size) is derived by single OO metric

total path- For this we have built poisson regression model where number of defects is explained by the mediator variable, i.e the class size and a sinle OO metric. Here the p value is being estimated by comparing this model with a null model.

RQ2: If the class size has mediation effect on the relationships between object oriented metrics and defect-proneness(binary model)? In order to answer the second research question we have calculated direct effect, indirect effect and total effect, same as the count model but by considering defect proneness as the dependent variable, single OO metric as independent variable and class size as mediator and the below steps were followed:

direct path- Logistic regression model is being considered where defect-proneness is derived by single OO metric and p value is being estimated.

indirect path- Linear regression model is being considered where the mediator variable(class size) is derived by single OO metric total path- For this again we have built poisson regression model where defect-proneness is explained by the mediator variable, i.e the class size and a sinle OO metric

Along with these for mediation analysis we have employed bootstaping mediation analysis with 5000 bootstrap resamples and set the confidence intervals to 95%.

RQ3: If the class size has moderation effect on the relationships between object oriented metrics and the number of defects(count model)?

For moderation analysis we have calculated interaction term by multiplying a single OO metric and the moderation variable(class size). In count model number of bugs is taken as dependent variable and p value is being estimated by fitting it into into poisson regression model. Also same implementation is being done using multivariate predictor.

RQ4: If the class size has moderation effect on the relation- ships between object oriented metrics and the number of defects(binary model)? Here moderation analysis is done using same manner as count model but by taking defect-proneness as dependent variable. In this we have used binomial regression model and multivariate predictor model is also taken into consideration.

2 Tools, Techniques and Dataset used

Code implementation is developed in Rstudio. For RQ1, RQ2, RQ3 and RQ4 the implementation is provided into Median_analysis_count_model.R,

Median_analysis_Binary_model.R, Moderation_analysis_Count_model.R and Moderation_analysis_Binary_model.R files respectively. Also scatter plots between the OO metrics and number of defects is provided in Plots.R file. Libraries used- Mediation, ggplot2,ggsignif and ggpubr Regression models- Linear regression, Logistic regression and Poisson regression We have used the same JURE dataset[Jureczko and Madeyski(2010)]

(dataset.csv) as the original paper[Tahir et al.(2021)] with a little modification(for example Fan-in and Fan-out parameters are being added and Check column is added for null model).

3 Result analysis

3.1 Scatter plots between OO metrics and Number of defects

We have analysed the scatter plots and P-value graphs between the object oriented metrics and Number of defects for all 6 metrics. For example, in figure 1 we have shown the relation between Fan-in metric and Number of defects and in 2 we have shown the relation between CBO metric and Number of defects.

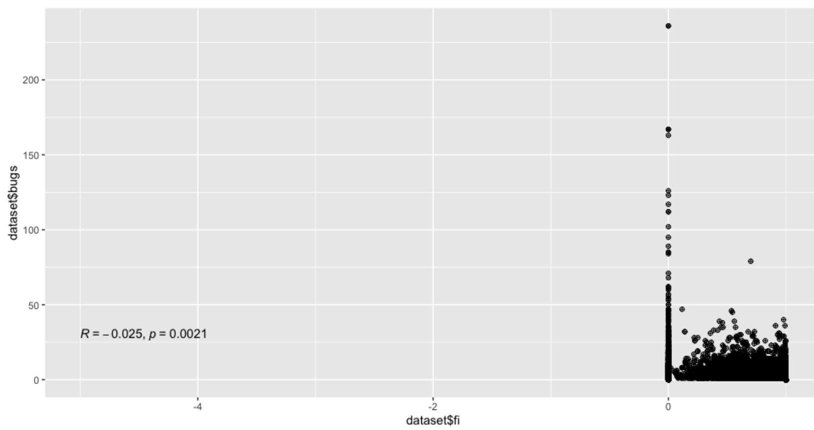

Figure 1: Scatter plot, X axis= Fan-in, Y-axis= Number of defects

3.2 Tables for showing P values

The results from the mediation and moderation analysis are shown in tabular form for both count and binary models. P value is significant when it's less than 0.05 and it's provided as bold characters.

P values for Mediation analysis- Count model			
OO metric	Direct effect	Indirect effect	Total effect
RFC	0.9986	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
WMC	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
CBO	0.1213	0.1184	**2.2e-16 < 0.001**
LCOM	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-in	**1.061e-16< 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-out	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**

P values for Mediation analysis- Binary model			
OO metric	Direct effect	Indirect effect	Total effect
RFC	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
WMC	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
CBO	**7.7e-16 < 0.001**	0.1184	**2.2e-16 < 0.001**
LCOM	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-in	0.1859	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-out	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**

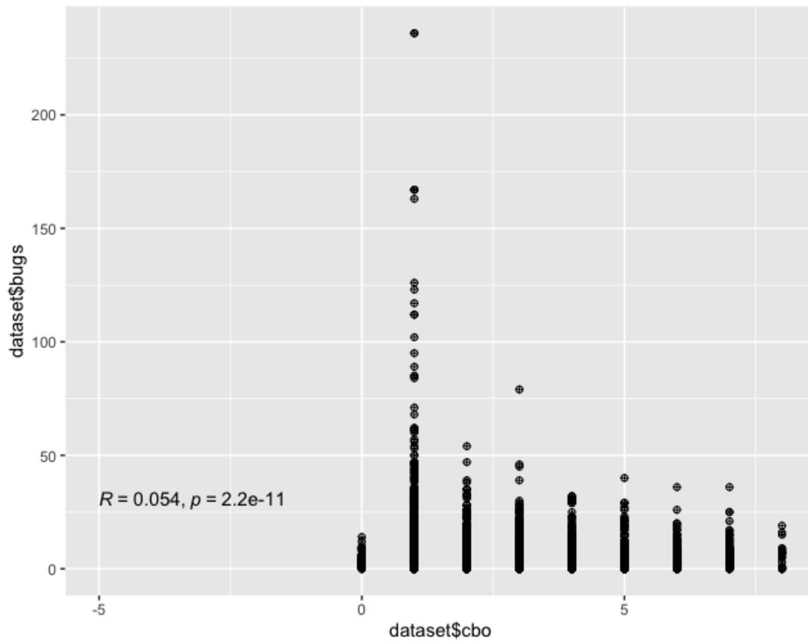

Figure 2: Scatter plot, X axis= CBO Y-axis= Number of defects

P values for Moderation analysis		
OO metric	Count model	Binary model
RFC	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
WMC	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
CBO	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
LCOM	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-in	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**
Fan-out	**2.2e-16 < 0.001**	**2.2e-16 < 0.001**

As per our mediation analysis results, except for CBO metric all other metrics are showing significant results, while in moderation analysis all the metrics are reflecting significant outcome. The result indicates that the class size has indirect effect on defect prediction models, but the effect is not trivial for every scenario. Hence as the assigned paper[Jureczko and Madeyski(2010)] suggested, we should at least consider the effect of class size in defect prediction models and check the effect for given dataset.

4 Acknowledgements

I would like to thank Dr. Sridhar Chimalakonda and all the reviewers of this work for providing us with the opportunity to understand the empirical study and their valuable inputs on this implementation work.

References

[Jureczko and Madeyski(2010)] Marian Jureczko and Lech Madeyski. 2010. Towards identifying software project clusters with regard to defect prediction. In *Proceedings of the 6th international conference on predictive models in software engineering*. 1–10.

[Tahir et al.(2021)] Amjed Tahir, Kwabena E Bennin, Xun Xiao, and Stephen G MacDonell. 2021. Does class size matter? An in-depth assessment of the effect of class size in software defect prediction. *Empirical Software Engineering* 26 (2021), 1–38.