Comments on "Researcher Bias: The Use of Machine Learning in Software Defect Prediction"

Chakkrit Tantithamthavorn[1], Shane McIntosh[2], Ahmed E. Hassan[3], and Kenichi Matsumoto[4]

[1]**Nara Institute of Science and Technology, Japan.**
[2]**McGill University, Canada**
[3]**Queen's University, Canada**
[4]**Nara Institute of Science and Technology, Japan.**

ABSTRACT

Shepperd et al. (2014) find that the reported performance of a defect prediction model shares a strong relationship with the group of researchers who construct the models. In this paper, we perform an alternative investigation of Shepperd et al. (2014)'s data. We observe that (a) researcher group shares a strong association with the dataset and metric families that are used to build a model; (b) the strong association among the explanatory variables introduces a large amount of interference when interpreting the impact of the researcher group on model performance; and (c) after mitigating the interference, we find that the researcher group has a smaller impact than the metric family. These observations lead us to conclude that the relationship between the researcher group and the performance of a defect prediction model may have more to do with the tendency of researchers to reuse experimental components (e.g., datasets and metrics). We recommend that researchers experiment with a broader selection of datasets and metrics to combat potential bias in their results.

Keywords: Software Quality Assurance, Software Defect Prediction, Machine Learning

1 INTRODUCTION

Shepperd et al. (2014) study the extent to which the researcher group that performs a defect prediction study associates with the reported performance of defect prediction models. Through a meta-analysis of 42 primary studies, they find that the reported performance of a defect prediction model shares a strong relationship with the group of researchers who construct the models.

In this paper, we perform an alternative investigation of Shepperd *et al.*'s data. More specifically, we set out to investigate (1) the strength of the association among the explanatory variables, e.g., research group and metric family (Section 2); (2) the interference that these associations introduce when interpreting the impact that the explanatory variables have on the outcome (Section 3); and (3) the impact that the explanatory variables have on the outcome after we mitigate the interference introduced by strongly associated explanatory variables (Section 4).

2 THE PRESENCE OF COLLINEARITY

We suspect that researcher groups are likely to reuse experimental components (e.g., datasets, metrics, and classifiers) in several studies. This tendency to reuse experimental components would introduce a strong association among the explanatory variables of Shepperd et al. (2014). To investigate our suspicion, we set out to measure the strength of the association between each pair of the explanatory variables that are used by Shepperd et al. (2014), i.e., ResearcherGroup, DatasetFamily, MetricFamily, and ClassifierFamily.

Approach. Since the explanatory variables are categorical, we first use a Pearson χ^2 test (Agresti, 1996) to check whether a statistically significant association exists between each pair of explanatory variables ($\alpha = 0.05$). Then, we compute Cramer's V (Cramér, 1999) to quantify the strength of the association between each pair of two categorical variables. The value of Cramer's V ranges between

Table 1. The association among explanatory variables.

Pair	Cramer's V	Magnitude
ResearcherGroup & MetricFamily	0.65^{***}	Strong
ResearcherGroup & DatasetFamily	0.56^{***}	Relatively strong
MetricFamily & DatasetFamily	0.55^{***}	Relatively strong
ResearcherGroup & ClassifierFamily	0.54^{***}	Relatively strong
DatasetFamily & ClassifierFamily	0.34^{***}	Moderate
MetricFamily & ClassifierFamily	0.21^{***}	Moderate

Statistical significance of the Pearson χ^2 test:
$\circ p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

0 (no association) and 1 (strongest association). We use the convention of Rea and Parker (2014) for describing the magnitude of an association. To compute the Pearson's χ^2 and Cramer's V values, we use the implementation provided by the `assocstats` function of the `vcd` R package (Meyer et al., 2015).

Results. **Researcher group shares a strong association with the dataset and metrics that are used.** Table 1 shows the Cramer's V values and the p-value of the Pearson χ^2 test for each pair of explanatory variables. The Cramer's V values indicate that researcher group shares a strong association with the dataset and metrics that are used. Indeed, we find that 13 of the 23 researcher groups (57%) only experiment with one dataset family, where 9 of them only use one NASA dataset, which contains only one family of software metrics (i.e., static metrics). Moreover, 39% of researcher groups only use the static metric family of the NASA dataset in several studies. The strong association among researcher groups, dataset, and metrics confirms our suspicion that researchers often reuse experimental components.

3 THE INTERFERENCE OF COLLINEARITY

The strong association among explanatory variables that we observe in Section 2 may introduce interference when one studies the impact that these explanatory variables have on the outcome (Grewal et al., 2004; Tu et al., 2005). Furthermore, this interference among variables may cause impact analyses, such as ANOVA, to report spurious relationships that are dependent on the ordering of variables in the model formula. Indeed, ANOVA is a hierarchical model that first attributes as much variance as it can to the first variable before attributing residual variance to the second variable (R. Clifford, 1978). If two variables share a strong association, the variable that appear first in the model formula will have the brunt of the variance associated with it. Hence, we set out to investigate the interference that is introduced by the strong association among explanatory variables.

Approach. To investigate this interference, we use a bootstrap analysis approach, which leverages aspects of statistical inference (Efron and Tibshirani, 1993). We first draw a bootstrap sample of size N that is randomly drawn with replacement from an original dataset that is also of size N. We train a linear regression model with the data of the bootstrap sample using the implementation provided by the `lm` function of the `stats` R package (R Core Team, 2013). For each bootstrap sample, we train models with all of the 24 possible ordering of the explanatory variables (e.g., ANOVA(ResearcherGroup, DatasetFamily, MetricFamily, ClassifierFamily) versus ANOVA(DatasetFamily, ResearcherGroup, MetricFamily, ClassifierFamily)). Following the prior study (Shepperd et al., 2014), we compute the partial η^2 values (Richardson, 2011), which describe the proportion of the total variability that is attributed to an explanatory variable for each of the 24 models. We use the implementation provided by the `etasq` function of the `heplots` R package (Friendly, 2015). We repeat the experiment 1,000 times for each of the 24 models to produce a distribution of the partial η^2 values for each explanatory variable.

Results. **The strong association among the explanatory variables introduces a large amount of interference when interpreting the impact that researcher group has on model performance.** Figure 1 shows the distributions of the partial η^2 values for each explanatory variable when it appears at each position in the model formula. The results show that researcher group and dataset family tend to have the largest impact when they appear in earlier positions in the model formula, indicating that the impact that explanatory variables have on the outcome depends on the ordering of variables in the model formula. Moreover, we observe that researcher group and dataset family tend to have comparable partial η^2 values, indicating that the strong association introduces a large amount of interference. In particular, a model

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1260v1 | CC-BY 4.0 Open Access | rec: 22 Jul 2015, publ: 22 Jul 2015

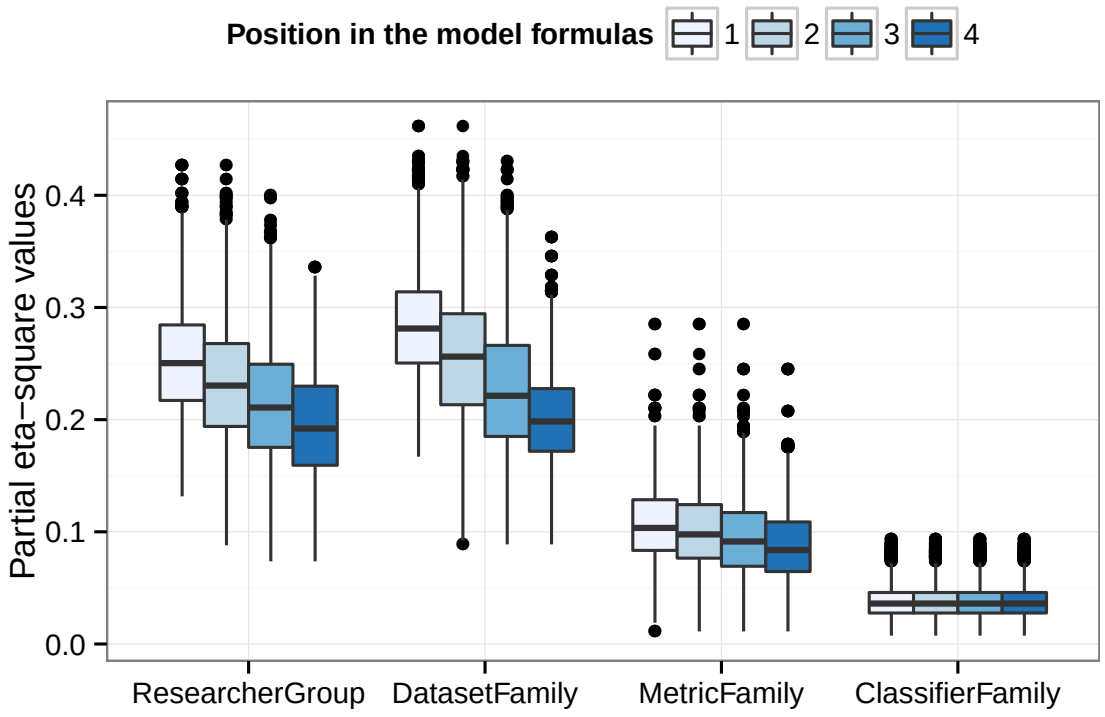

Figure 1. Distribution of the partial η^2 values for each explanatory variable when it appears at each position in the model formula.

with DatasetFamily in the first position and ResearcherGroup in the second position in the model formula would lead us to believe that DatasetFamily plays a bigger role than the ResearcherGroup. However, once we flip the positions of DatasetFamily and ResearcherGroup, we would reach a different conclusion — ResearcherGroup plays a bigger role in the model. On the other hand, the moderate association between the metric and classifier families introduces much less interference.

4 MITIGATING THE COLLINEARITY

The prior sections show that the strong association among the explanatory variables (Section 2) introduces a large amount of interference when we interpret the impact that the explanatory variables have on the outcome (Section 3). In this section, we set out to investigate the impact that the explanatory variables have on the reported performance after mitigating the interference introduced by strongly associated explanatory variables. Our earlier analysis indicates that we cannot have all three explanatory variables in the same model. Instead, we should only include two of these three variables in our models. Hence, we opt to focus on one studied dataset in order to control the dataset family metric by holding it constant.

Approach. To mitigate the interference, we first select the Eclipse dataset family, which is the second-most popular dataset family in the studied dataset. We choose the Eclipse dataset family instead of selecting the NASA dataset family because Section 2 reveals that 39% of the researcher groups only use the static metric family of the NASA dataset in several studies. Overall, the Eclipse dataset family contains 6 metrics families, which are used by 7 researcher groups who fit defect prediction model using 6 classifier families. While controlling the dataset family metric, only researcher group, metric family, and classifier family will be included in the model formulas. Since the dataset family metric is now a constant, it is excluded from our model formula.

Since Table 1 shows that researcher group shares a strong association with the metric family, we build two different linear regression models by removing one of the two strongly associated variables, i.e., one model uses the researcher group and classifier family variables, while another model uses the metric family and classifier family variables.

To confirm the absence of the collinearity in the models, we perform a redundancy analysis (Harrell Jr., 2002) in order to detect redundant variables prior to constructing the models. We use the implementation

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1260v1 | CC-BY 4.0 Open Access | rec: 22 Jul 2015, publ: 22 Jul 2015

Table 2. Partial η^2 values of the ANOVA analysis with respect to the Eclipse dataset family.

	ResearcherGroup Model	**MetricFamily Model**
Adjusted R^2	0.20	0.28
AIC	-80	-95
	Partial η^2	Partial η^2
Researcher Group	0.124 (medium)	†
Metric Family	†	0.201 (large)
Classifier Family	0.120 (medium)	0.096 (medium)

†Strong association variables

provided by the `redun` function in the `rms` R package (Harrell Jr., 2015).

To assess the fit of our models, we compute the adjusted R^2 and the Akaike Information Criterion (AIC) (Akaike, 1974). The adjusted R^2 measures the amount of variability, while AIC measures the goodness-of-fit based on information entropy. In general, higher adjusted R^2 and lower AIC values correspond to a better fit of the model to the underlying data.

Finally, we perform an ANOVA analysis and compute the partial η^2 values. As suggested by Richardson (2011) and Mittas and Angelis (2013), we use the convention of Cohen (1988) for describing the effect size of the partial η^2 values — values below 0.01, 0.06, and 0.14 describe small, medium, and large effect sizes, respectively.

Results. **When we mitigate the interference of strongly associated explanatory variables, we find that the researcher group has a smaller impact than the metric family with respect to the Eclipse dataset family.** Table 2 shows partial η^2 values of the ANOVA analysis with respect to the Eclipse dataset family. The results show that the MetricFamily model, which achieves a higher adjusted R^2 and a lower AIC, tends to represent the underlying data better than the ResearcherGroup model. The redundancy analysis also confirms that there are no redundant variables in the MetricFamily model. The ANOVA analysis of the MetricFamily model shows that the choice of metrics that are used to build defect prediction models tends to have a large impact on the reported performance with respect to the Eclipse dataset family. Moreover, since the interference has been mitigated, the ANOVA results still hold when the explanatory variables are reordered.

5 CONCLUSIONS

The prior work of Shepperd et al. (2014) suggests that the reported performance of a defect prediction model shares a strong relationship with the researcher group who conducted the study. In this paper, we investigate (1) the strength of the association among the explanatory variables of Shepperd et al. (2014)'s study; (2) the interference that these associations introduce when interpreting the impact of the explanatory variables on the outcome; and (3) the impact that the explanatory variables have on the outcome after we mitigate the interference introduced by strongly associated explanatory variables. We make the following observations:

- Researcher group shares a strong association with the dataset and metrics families that are used in building models, suggesting that researchers should experiment with a broader selection of datasets and metrics.
- The strong association among explanatory variables introduces a large amount of interference when interpreting the impact that researcher group has on the reported model performance, suggesting that researchers should carefully mitigate collinearity issues prior to analysis.
- After mitigating the interference, we find that the researcher group has a smaller impact than metric family with respect to the Eclipse dataset family, suggesting that researchers should carefully examine the choice of metrics when building defect prediction models.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1260v1 | CC-BY 4.0 Open Access | rec: 22 Jul 2015, publ: 22 Jul 2015

These observations lead us to conclude that the relationship between researcher groups and the performance of a defect prediction model may have more to do with the tendency of researchers to reuse experimental components (e.g., datasets and metrics).

ACKNOWLEDGMENTS

We greatly appreciate that Shepperd et al. (2014) have share both their dataset and experimental scripts online, which provided us with the means to conduct this study. In the same spirit, we also provide access to our experimental scripts.[1]

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis*, volume 135. Wiley New York.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic press.

Cramér, H. (1999). *Mathematical methods of statistics*, volume 9. Princeton university press.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Springer US, Boston, MA.

Friendly, M. (2015). heplots: Visualizing hypothesis tests in multivariate linear models. `http://CRAN.R-project.org/package=heplots`.

Grewal, R., Cote, J. a., and Baumgartner, H. (2004). Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing. *Marketing Science*, 23(4):519–529.

Harrell Jr., F. E. (2002). *Regression Modeling Strategies*. Springer, 1st edition.

Harrell Jr., F. E. (2015). rms: Regression modeling strategies. `http://CRAN.R-project.org/package=rms`.

Meyer, D., Zeileis, A., Hornik, K., Gerber, F., and Friendly, M. (2015). vcd: Visualizing categorical data. `http://CRAN.R-project.org/package=vcd`.

Mittas, N. and Angelis, L. (2013). Ranking and Clustering Software Cost Estimation Models through a Multiple Comparisons Algorithm. *IEEE Transactions on Software Engineering*, 39(4):537–551.

R. Clifford, B. (1978). Tests of Hypotheses for Unbalanced Factorial Designs Under Various Regression/Coding Method Combinations. *Educational and Psychological Measurement*, (38):621–631.

R Core Team (2013). R: A language and environment for statistical computing. `http://www.R-project.org/`.

Rea, L. M. and Parker, R. A. (2014). *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2):135–147.

Shepperd, M., Bowes, D., and Hall, T. (2014). Researcher Bias : The Use of Machine Learning in Software Defect Prediction. *IEEE Transactions on Software Engineering*, 40(6):603–616.

Tu, Y.-K., Kellett, M., Clerehugh, V., and Gilthorpe, M. S. (2005). Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *British dental journal*, 199(7):457–461.

[1] `http://sailhome.cs.queensu.ca/replication/researcher_bias_comments/`