Updated Conclusion

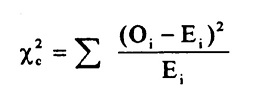
**Summary**: This paper is looking for the relationship between leaks and bugs in the same class. In total, we have two ways to look at the relationship. 1) The relation between number of leaks and number of bugs. Example: If more leaks a class has will cause more bugs in the class. 2) The relation between the existence of leaks and existence of bugs. Example: If a class, which has leaks, is more possible to have bugs. We use regression model to test the first relationship, and use chi-square test to do the second relationship. For the regression test, We use 26018 classes as our data. We have number of bugs and number of leaks for each of class. We make number of bugs as Y axis and number of leaks as X axis and get a scatter plot. Using statistical software to get the most suitable regression to fit the data. And See if we can find a good regression. We use R-square as an indicator to decide if the regression is good or not. We also divide the whole data into different groups based on the number of class a project has, number of revision a project has and based on the definition of maturity. And do regression test for different groups to see if there are some changes. For chi-square test, we have use the same three classification, number of class a project has, number of revision a project has and the definition of maturity. For each classification, we have four numbers, number of leak-bug class, number of leak-nonbug class, number of nonleak-bug class and number of nonleak-nonbug class. We use X^2 as an indicator to decide the confidence level to reject Null hypothesis: There is no relationship between the existence of bugs and existence of leaks. If we reject Null hypothesis, then we accept the alter hypothesis: There exist such relationship. Through all test, we have the conclusion that there is no relationship between number of bugs and number of leaks. However, there exist a relation between existence of bugs and existence of leaks, except some special classification, which we will mention later.

From the leaking query, we have the number of leaking places in each class. By going through each revision's log, we check if it mentions "bug(s)", "fix(es)", "error(s)" and "issue(s)". If yes, then we consider this revision is about fixing bugs. We consider the related classes in this revision had bugs once. Everytime we found bug fixing revision. add 1 to bugs number for each class. Right now, we have leaking number and bugs number for each class. we can start to test our assumption: If there is correlation between bugs and leaks? The first test is regression model test. We have 26,018 ckasses and their bugs number and leaks number. We put leaks number as X-axis, bugs number as Y-axis and get figure 1. We fit the data with regression model and get the equation below:

Y= a+b\*X+ c\*lnX + d\*√ X + e\*X^-1 + f\*X^-2

The R-square value of this model is only 0.03. This value explains how well this model fits the data. 0.03 is too small to be a good model. **So we consider that there is no correlation between number of bugs and number of leaks**.

Right now, we want to figure out if leaks have influence on the appearance of bugs. To figure out this question, we us Chi-square test of independence to check if leaks and bugs are two independent variable, or there exist some influence on each other. Our null hypothesis is (H0): There is no relationship between leak and bug. And Ha: There is relationship between leak and bug. We have 31256 classes in total, and each number for leak, non-leak, bug and non-bug, in table 1.



Oi is the number of 3690, 5460, 8292 and 13810 in the table 1. Ei is the estimated number in table 2 correspondingly. By the formula, we have X^2 is 21.365. We choose the 1 degree of freedom and 0.05 level of P-value = 3.841. 21.365 is much bigger than 3.841, which means our null hypothesis is only less 0.05 possibility correct. So we reject null hypothesis. **We conclude that There is relationship between leaks and bugs.**

Combine two conclusions we get above, we know That:

1. No correlation between number of bugs and number of leaks, which means that more leaks does means more bugs in the class.

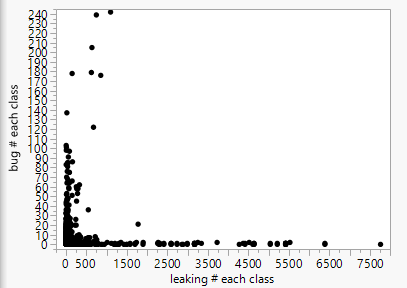
2. There is relationship between leaks and bugs, which means a class with leaks is more likely has bugs compared with the class without leaks.

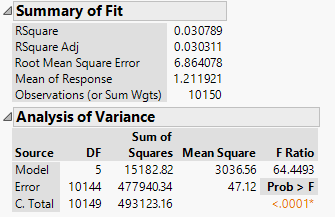
|  |  |  |  |
| --- | --- | --- | --- |
|  | leak | Non-leak | sum |
| bug | 3690 | 5464 | 9154 |
| Non-bug | 8292 | 13810 | 22102 |
| sum | 11982 | 19274 | 31256 |

Table 1

|  |  |  |
| --- | --- | --- |
|  | estimated leak | estimated nonleak |
| estimated bug | 3509.189532 | 5644.810468 |
| estimated nonbug | 8472.810468 | 13629.18953 |

Table 2

Figure 1



**Detail Analysis:**

To look for more detail relationship between leaks and bugs, we analyze the relationship depending on different classification.

Firstly, we divide our class data into different groups based on the scale of the project the class belongs to. We have 6 groups, number of class is less than 6, 12, 25, 50, 100 and bigger than 100. For each group, we did regression analysis for each of the groups. The corresponding R-square value is in table 3, And the regression plot is in appendix. We also did the chi-square test, where X^2 is in table 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | For regression test depending on class classification | | | | | |
|  | class#<6 | class#<12 | class#<25 | class#<50 | class#<100 | class#>100 |
| R-square | 0.005898 | 0.03808 | 0.001054 | 0.01928 | 0.001958 | 0.0237 |

Table 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | For chi-square test depending on class classification | | | | | |
|  | class#<6 | class#<12 | class#<25 | class#<50 | class#<100 | class#>100 |
| X^2 | 2.5109 | 6.8175 | 1.4227 | 2.2381 | 6.0087 | 17.6405 |

Table 4

Analysis: Because the R-square stands for how well the regression model fit our data, The R-squares are too small, which means there is not strong relation between number of leaks and number of bugs. As for the chi-square test, the group class#<12, class#<100 and class#>100 have a large X^2 (bigger than 0.05 level of P-value = 3.841). Then in these group, There is a strong confidence(95%) to prove the relationship between the existence of leaks and bugs. On the contrary, we don’t have a strong confidence to prove that there is a strong relationship between existence of leaks and bugs. However, these these three numbers are bigger than the 0.25 level p-value 1.32. So we have a moderate confidence(75%) to prove the relationship.

Secondly, we divide our class data into different groups based on the scale of revisions the project has. We have 6 groups, number of revision is less than 6, 12, 25, 50, 100 and bigger than 100. And then we did the regression (table 5) and chi-square (table 6) test on these groups.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | For regression test depending on revision classification | | | | | |
|  | revision#<6 | revision#<12 | revision#<25 | revision#<50 | revision#<100 | revision#>100 |
| R-square | 0.002513 | 0.04194 | 0.001453 | 0.0249 | 0.01877 | 0.02703 |

Table 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | For chi-square test depending on revision classification | | | | | |
|  | revision#<6 | revision#<12 | revision#<25 | revision#<50 | revision#<100 | revision#>100 |
| X^2 | 3.4899 | 4.7648 | 23.3354 | 0.1184 | 122.8997 | 6.1123 |

Table 6

Analysis: For the same reason, the small R-square tells us there is no strong relationship between number of leaks and number of bugs. Under chi-square test, for the groups revision#<12, revision#<25, revision#<100 and revision#>100, we have strong(95%) confidence to prove the relationship between the existence of leaks and existence of bugs. But for group revision#<6, we have 90%(p-value 2.71) confidence to prove the relationship. For group revision#<50, we have a only 25%(p-value 0.102) confidence to prove the relationship.

Thirdly, We divide our class data into different groups based on the definition of mature project (the project repository has more than 10 classes, more than 1000 statements and more than 50 commits). Regression result is in table 7 and chi-square test is in table 8:

|  |  |  |  |
| --- | --- | --- | --- |
|  | For regression test depending on mature classification | | |
|  | mature | non-mature | All |
| R-square | 0.02604 | 0.01392 | 0.01874 |

Table 7

|  |  |  |  |
| --- | --- | --- | --- |
|  | For chi-square test depending on mature classification | | |
|  | mature | non-mature | All |
| X^2 | 40.3618 | 19.5166 | 16.751 |

Table 8

Analysis: The small R-square explains the weak relationship between number of bugs and number of leaks. And for chi-square test, the large X^2 also proves that we have a strong confidence(95%) to prove that there is a relationship between the existence of leaks and existence of bugs.

**Appendix**