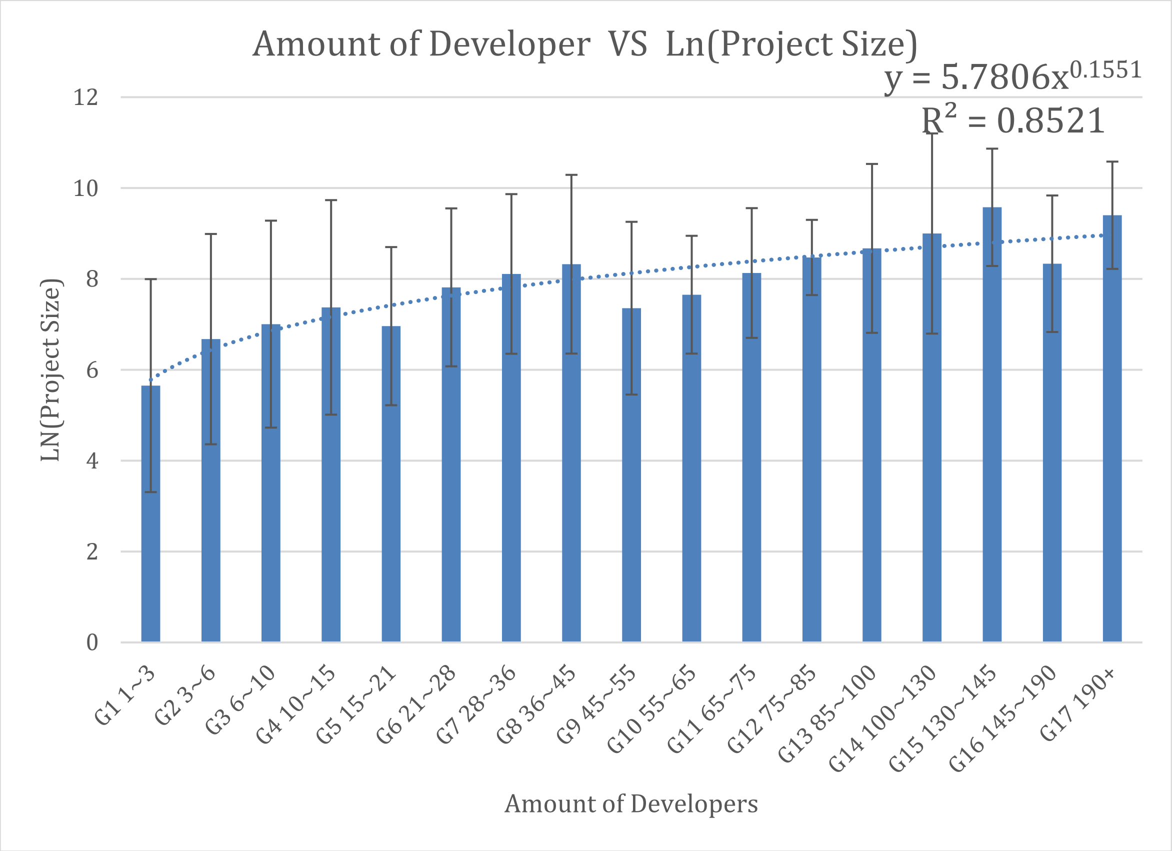
VANT 149 Research Project Proposal: Amount of Developers Influenced Size of OSS Projects

Kewei Li and Yuhang Zeng

Vantage College, University of British Columbia

Data Interpretation

A summarized data is attached below, with 17 groups of developer data corresponding to the average of natural logarithm of project sizes and trend line was generated by excel.



The model can be used in the following steps:

1. Estimate the line of code you probably need => (a).
2. Many style guides for computer programming define the maximum character per line of source code as 80 characters. Considering not all lines in the code can fulfill this maximum limit, the amount of characters per line can be estimate as 60% of 80 character (80*0.6=48). And 1 kilobyte (kB) = 1024 characters = 1024/48 lines of code = 21.3 lines of code.
3. a/21.3 would give the approximately size of project in unit of kilobytes (kB) => (b).

4. Use natural logarithm on the size of project (b): Ln(b) =>(y).
5. Plug (y) into the model: $y=5.7806\ x\text{^}(0.1551)$.
6. Figure out the (x). which integer parts indicated which group it would fall into among all 17 groups in the model, and the fractional part will indicate the exact place it would at in this indicated group. E.g. if the x=8.3, the integer part indicates the group: 8 (36-45 developers); and the fractional part indicates the accurate number: 36+0.3*(45-36) =38.7. (Approximately 39 developers would need for this project).

The results indicated that with the increase of amount of developers, the size of repositories increased and by trying match with different kind of trend line, the model fit a power trend line the best ($y=5.7806\ x\text{^}\ 0.1551$). Which full filled the expectation of this study and also proved the hypothesis.

According to the raw data which used to generate this model, the model is good at estimate the project size between 788kB and 4500kB or in terms of amount of developers, between 3 to 85 developers. Since the data projects that were too small might been highly separate, and due to the limit of time and resource, some of the large repositories were abandoned in this study, which makes there exist not enough data of large projects to conclude a solid result.

The results can be improved by analyzing more research sample especially more large repositories, which could fill in the leak of large repositories in the current data set of this study. As for future studies, different time range may be used and more samples may be testified to gain other interesting models.

This model is currently still unfinished, since the sample size we used are still not big enough to finished the estimation.