**Title**

Improving computational efficiency in identifying parsimonious statistical models

**Introduction**

Stepwise regression is a method of continually adding or removing predictive variables from a regression model to find its most parsimonious fit. This is done by calculating its AIC each step and finding the lowest value. The lower the value, the more parsimonious of a model it is. StepAIC is an algorithm coded for the R statistical environment that can do this automatically.

Many factors determine how long StepAIC will take to run. These include how big of a dataset is being used, does it contain squared terms, and what interaction level is being used. The more complex of a model passed into StepAIC the longer it will take. When a model is sufficiently complex, it can take many hours, or even days to produce results. To address this issue, we developed a new algorithm.

We present Greedy, another stepwise regression algorithm, but one that can produce essentially equal parsimonious models as StepAIC, but considerably faster. The reason StepAIC takes so long is that it considers models cubic to the amount of predictor variables. This contrasts with Greedy, which considers models quadratic to the amount of predictor variables. This allows us to have a dramatically lower run time while producing models with very similar AIC values to the ones produced by StepAIC.

**Materials and Methods**

This algorithm was developed in the R language using the R Studio IDE. All runtimes were gathered in R Studio on a PC with the following specs:

OS: Windows 10 64-bit x64-based processor

CPU: Intel i5-4690k @ 3.50GHz

RAM: 16 GB

GPU: NVIDIA GeForce GTX 770

Nine datasets were used to test the algorithms. Each dataset was obtained from the UCI Machine Learning Repository in the format of CSV files. The datasets ranged from 396 to 39645 observations. Each dataset contained quantitative predictor variables and all but four contained categorical predictor variables.

For each dataset, a group of linear models or generalized linear models were used. The models were various combinations of predictor variables and squared quantitative predictor variables. For each model, the model was tested using either no interactions, 2nd order interactions, 3rd order interactions, or 4th order interactions. Once the model was loaded into the global environment, it was passed into each algorithm with the time to produce results and AIC value of the best possible model returned recorded into an Excel spreadsheet.

**Results**

Testing of Greedy shows that for the 78 models tested, Greedy was faster in 60 of them, StepAIC was faster in 13, and there were 5 times where the runtimes were the same. StepAIC had faster times in only 4 out of the 9 datasets. There were only 2 instances of StepAIC performing better where the runtime for each algorithm exceeded one second. The most time saved by using Greedy was about 4.2 hours. When StepAIC was faster, the most it was ever faster by was 3.3 seconds.

For these results to be significant, it needs to be shown that the model produced by Greedy was as good, or nearly as good as the model produced by StepAIC. The most the AIC value ever differed by between the two algorithms across the 9 datasets was 3.279%. The average AIC difference across the 9 datasets was .206%.

**Conclusion**

Although the Greedy algorithm is still being developed, the preliminary results are promising. Greedy consistently outperforms StepAIC, and even when it doesn’t, it’s only by a matter of seconds.

**Acknowledgements**

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.