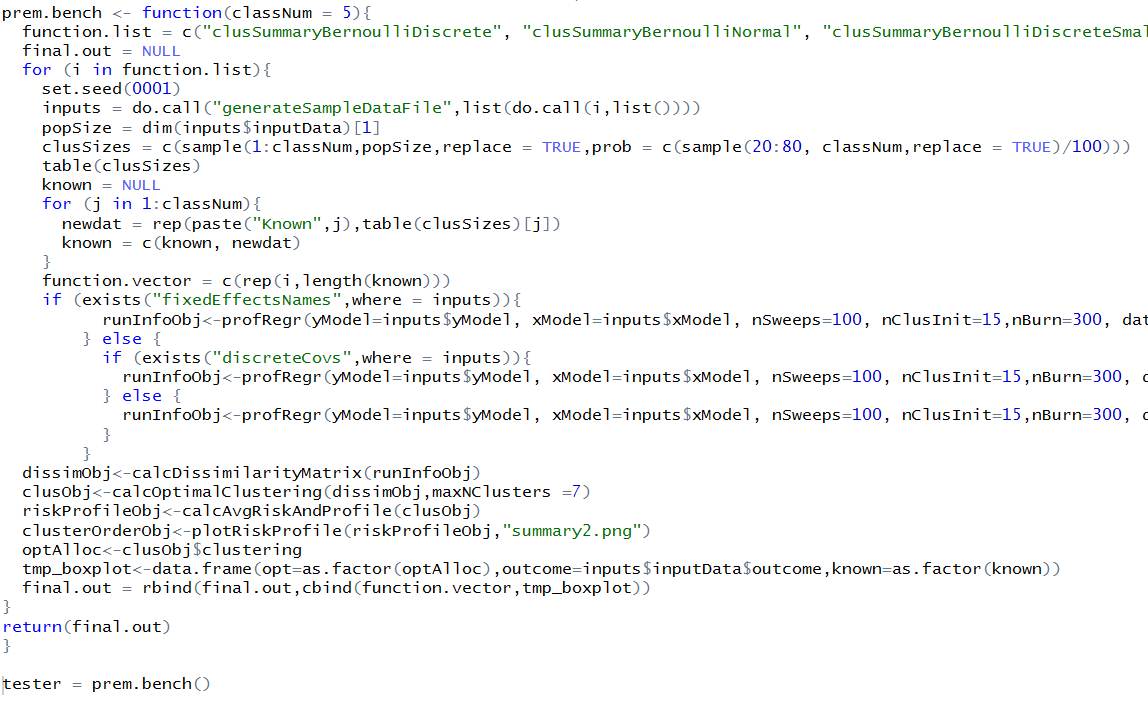
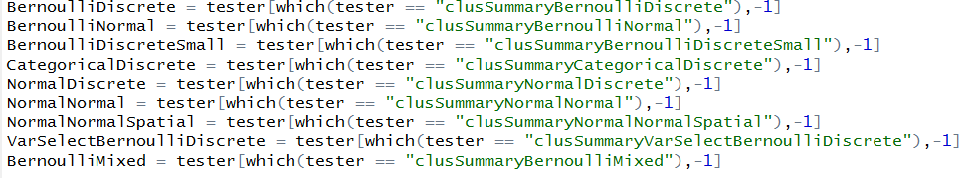
Benchmark Function for PReMiuM R Package.

# Description

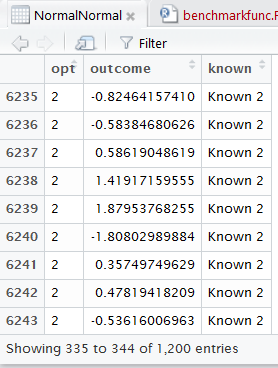
The *prem.bench()* function for R uses multiple functions built into the PReMiuM R package to run regressions on several simulated data sets from varying distributions, find the optimal clustering of the data and compare it to a generate known truth clustering. PReMiuM offers 15 different combinations of distributions to generate data from such as: normal, discrete, Bernoulli, etc. The first named distribution is the outcome type, the second named distribution is the covariate type. The benchmarking function generates sample data from 9 of 15 (see “Future Additions”) possible combinations of distributions offered by the PReMiuM package. After running a profile regression and finding the optimal clustering for each combination, the benchmarking function gives the final output. The final output consists of 4 columns: name distributions, optimal cluster, outcome value, known truth cluster.



Above: the entire *prem.bench()* function and a line to run it, saving the output as *tester*. The output can be left like this where each distributions output will be listed one after another in the 4 columns. However using the following set of code, the user can break up the final output into separate data frames based on the distribution. The code also removes the column of names to avoid redundancy.



The results of separating each data set can then be saved for comparison when benchmarking. Each row will contain an index, the optimal cluster, the outcome value, and the known truth cluster for each entry, as seen below:



# Uses

When benchmarking PReMiuMon high performance computers (HPCs), this function will allow the user to run varying distributions through the PReMiuM software and receive the output in an organized manner for comparison between runs. It can be used for testing accuracy when trying to optimize functions inside the PReMiuM package, as well as the overall performance in time/memory of the functions. The user can currently customize the total number of clusters in the known truth. The default number of known truth clusters is 3.

# Future Additions

Currently, the sample data sets for each distribution in PReMiuM are fixed in population size. The next step is to implement the randomized known truth cluster method used in *prem.bench()* into the generation of the sample data. This would allow the user to customize the population size of each data set, since larger data sets will give better use to benchmarking on HPCs. Additionally, the number of clusters of the generated data is fixed. If the number of clusters can be set as well, then confusion matrices can be implemented as an additional output. The missing 6 distributions are due to additional parameters for some of the distributions, this should be fixed with some additional code in the function and should be possible. Implementing the separation of the output data into the function, rather than an additional line of code could provide useful if trying to run the benchmark function multiple times and store the output sequentially.