Stat 243 Class Project: Building a Genetic Algorithm Based Variable Selection Algorithm

Joy Hou, Kevin Li, Greta Olesen

December 11, 2014

1 Introduction

Genetic Algorithms are search heursitics that mimic the process of Darwinian natural selection. As Givens and Hoeting (2005) states, candidate solutions to a maximization/minimization problem are considered as biological organisms represented by the genetic code which specifies the model attributes. Genetic algorithms are especially useful in large scale combinatorial and nonlinear optimization when traditional optimization techniques become untractable. Rather than exhaustively searching for the global optimal solution, the Genetic Algorithm utilizes heuristic-based search methods and returns a solution close to the global optimal solution.

Variable selecton problems, when presented with a large set of potential predictors, become increasingly computationally expensive. As a result, practitioners and academics have turned to search heuristics such as the Genetic Algorithm.

The basic genetic algorithm is as follows:

1. Initialize the first generation of models by generating a random population of n chromosomes
2. Evaluate the fitness/performance f(m) of each chromosome/model m in the first generation
3. Create a new population by repeating the following steps until the termination condition is met:
   (a) Select n/2 pairs of parents chromosomes/models from the previous generation according to the fitness (fitter chromosomes have a greater chance to be selected)
   (b) Crossover is carried out with a crossover probability. If crossover was performed, the 2 parent chromosomes are crossed over to produce 2 children. If no crossover was performed, 2 copies of the original chromosomes/models will be kept.
   (c) Mutation is carried out with a mutation probability. When mutation is carried out, the new offspring is mutated at each locus.
   (d) Accept the children/offsprings as the next population
4. If the terminal condition is satisfied, stop, and return the best solution in the current population. Otherwise, return to step 2

The genetic algorithm by no means returns the optimal solution, but users can generally expect an acceptable solution with rapid convergence that often resembles that presented figure below:

2 Code Structure

We utilized functions in R to carry out our Genetic Algorithm. We chose to make use of functions rather than OOP methods because the functions could be put to use in a clear, orderly manner. Consequentially, the algorithm is outlined as follows:

```r
result <- select(X, y, popSize, criterion, type, family, criFun, max_iterations,
                 crossRate, mRate, zeroToOneRatio)
```

Please refer to the help manual for more information about each argument of *select*. The function *select* employs all of the auxiliary functions that are required to carry out the Genetic Algorithm.

The following is a list of brief descriptions of the main auxiliary functions that are utilized in the primary *select* function.

popInitialize function

This function randomly generates the initial population to start the Genetic Algorithm. Its main arguments are the desired population size and appropriate gene length (the number of potential predictors).

The function returns a matrix with the following dimension: *popSize* × *geneLength*. Each row of the matrix represents an initial parent/model. Each column of the matrix represents a variable that could potential be included in the final model.

The function makes sure to check that each individual includes at least one variable as a predictor. If the individual does not include any predictors, it is omitted and regenerated until it fits the criteria.

evalFunction function

The *evalFunction* makes use of parallelization to evaluate each individual/model in the current population based on some criterion.

This is a function called *singleEval* that is implemented inside *evalFunction*. *singleEval* evaluates a single model and returns the criterion value for that model. The criterion can be one of the built-in criterions: AIC or BIC. The user can also input their own function into the argument *criFun* to evaluate the model. The user inputed *criFun* is the function that should be minimized. If specified, the user's function should take in an lm object and return a single criterion value. The following is an example of a function that the user could pass as an argument for *criFun*:

```r
fun <- function(lm_ob){
    adj_r_squared <- summary(lm_ob)$adj.r.squared
    return(-adj_r_squared)
}
```

The function *evalFunction* utilizes the *foreach* function in the *foreach package* package to parallelize the execution of *singleEval*. This optimizes the speed of the algorithm.

Next, the criterion values are ranked from lowest to highest for each individual/model. The probabilities are determined directly from the rank:

$$Pi = \frac{-rank_i}{\sum_{i=1}^{n} N}$$

Where N is the population size (or the total number of individuals).

The function returns a matrix that contains the fitness level of each individual/model, the rank for each model, and the sampling probability for each model.

updateSamp function

The *updateSamp* function selects n/2 pairs from the parents individuals/models from the previous generation according to the fitness level. The sampling probabilities are determined by output from the *evalFunction* from the previous iteration of the algorithm. Fitter individuals/models have a greater chance of being selected.

crossover function

This function performs crossover for one pair of individuals/models depending on a crossover probability. It randomly generates a cutoff value and then crossover is performed. See the following example:

$$Individual_1 = 1\ 0\ 1\ 1\ 0\ 1$$

$$Individual_2 = 0\ 1\ 1\ 0\ 0\ 1$$

$$\text{Random Cutoff} = 2\text{, now we have:}$$

$$Individual_1 = 1\ 0\ 1\ 0\ 0\ 1$$

$$Individual_2 = 0\ 1\ 1\ 1\ 0\ 1$$

mutation function

This function performs mutation on a pair of individuals/models depending on a mutation probability. A mutation is a switch from a 1 to a 0 or a 0 to a 1 in an individual/model (in other words, it's the act of changing the inclusion or exclusion of a variable in a given model).

Mutation can only occur in positions where both parents shared the same value. The following is an example of boxes are the positions in which mutation could occur:

$$Individual_1 = 1\ 0\ \boxed{1}\ 0\ \boxed{0}\ \boxed{1}$$

$$Individual_2 = 0\ 1\ \boxed{1}\ 1\ \boxed{0}\ \boxed{1}$$

Now for each of the following positions, mutations will occur at the user specified *mRate*. For example, if *mRate* = .01 and the last position in $Individual_2$ is selected to be mutated, we will have the following output:

$$Individual_1 = 1\ 0\ 1\ 0\ 0\ 1$$

$$Individual_2 = 0\ 1\ 1\ 1\ 0\ \boxed{0}$$

select function

This is the main function that implements the genetic algorithm. The functions are implemented in the following order:

1. *popInit* to initialize the population
2. *evalFunction* to get the initial sampling probabilities for each model
3. Repeat until convergence, or the maximum number of iterations have been completed:
   (a) *updateSamp*: sample from the
   (b) *crossover*: execute in a loop to iterate over all of the pairs in the population
   (c) *mutation*: execute in a loop to iterate over all of the pairs in the population
   (d) *evalFunction*: execute to get the sampling probabilities for each individual/model

3 Testing

Auxilary Functions

We have rigorously tested each auxiliary function and checked that they return the desired results. The following shows the test functions for each auxilary function:

```
testInitial <- function(){
  pop <- popInitialize(popSize = 100, zeroToOneRatio = 1, geneLength = 10)
  cat("1. The popInitialize() function returns an initial generation of models with 1\nindicating an in
  cat("Here is what one model inside the generation looks like :\n")
  print(pop[1,])
  cat("The zero-to-one ratio in the model is approximately 1 to 1 as specified in default.\n")
}
testInitial()
```

```
## 1. The popInitialize() function returns an initial generation of models with 1
## indicating an included variable and 0 indicating an excluded variable.
## Here is what one model inside the generation looks like :
##  [1] 0 0 0 1 1 0 0 0 1 0
## The zero-to-one ratio in the model is approximately 1 to 1 as specified in default.
```

```
testSingleEval <- function(){
  X <- mtcars[,2:11]
  y <- mtcars[,1]
  singleGene <- sample(c(0,1),dim(X)[2],replace=T)
  criValue <- singleEval(singleGene,X,y,"lm","BIC",NULL,"gaussian")
  cat("2. The singleEval() function returns the value of the evaluation criterion for one\nspecified mo
  cat("For example, the BIC here is:\n")
  cat(criValue)
  cat("\n")
}
testSingleEval()
```

```
## 2. The singleEval() function returns the value of the evaluation criterion for one
## specified model.
## For example, the BIC here is:
## 169.6
```

```
testEval <- function(){
  X <- mtcars[,2:11]
  y <- mtcars[,1]
  currentGenePool <- popInitialize(popSize = 100, geneLength = dim(X)[2],zeroToOneRatio = 1)
  criterion <- evalFunction(X,y,currentGenePool = currentGenePool,popSize = 100)
  cat("3. The evalFunction() returns the AIC rank and sampling probability of each of \nthe models in th
  print(criterion[,1])
  cat("\n")
}
testEval()

## 3. The evalFunction() returns the AIC rank and sampling probability of each of
## the models in the generation, for example:
##           AIC         ranks samplingProbs
##      154.3274        2.0000        0.0196
```

```
testUpdate <- function(){
  X <- mtcars[,2:11]
  y <- mtcars[,1]
  weights <-rep(1,100)
  currentGenePool <- popInitialize(popSize = 100, geneLength = dim(X)[2],zeroToOneRatio = 1)
  newGenePool <- updateSamp(currentGenePool,popSize = 100, weights = weights)
  cat("4. The updateSamp function updates the population according to the specified weights.\n")
  cat("Here is what one model in the new generation looks like:\n")
  print(newGenePool[1,])
}
testUpdate()

## 4. The updateSamp function updates the population according to the specified weights.
## Here is what one model in the new generation looks like:
##  [1] 0 1 1 0 0 1 0 1 1 1
```

```
testCrossOver <- function(){
  set.seed(1)
  v1 <- rep(1,10)
  v2 <- rep(0,10)
  geneLength <- length(v1)
  child <- crossover(v1,v2,geneLength,1)
  cat("5. The crossover() function returns two models generated by cross over:\n")
  cat("The genes before crossover:\n")
  print(rbind(v1,v2))
  cat("The genes after crossover:\n")
  print(child)
  cat("\n")
}
testCrossOver()

## 5. The crossover() function returns two models generated by cross over:
## The genes before crossover:
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## v1    1    1    1    1    1    1    1    1    1     1
## v2    0    0    0    0    0    0    0    0    0     0
## The genes after crossover:
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## new1    1    1    1    1    0    0    0    0    0     0
## new2    0    0    0    0    1    1    1    1    1     1
```

```
testMutation <- function(){
  v1 <- sample(c(0,1),10,replace=T)
  v2 <- sample(c(1,0),10,replace=T)
  geneLength <- length(v1)
  child <- mutation(v1,v2,1)

  cat("6. The mutation() function returns two models generated from mutation:\n")
  cat("The genes before mutation:\n")
  print(rbind(v1,v2))
  cat("The genes after mutation:\n")
  print(child)
  cat("\n")
}
testMutation()

## 6. The mutation() function returns two models generated from mutation:
## The genes before mutation:
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## v1    1    1    0    1    1    1    1    0    0     0
## v2    0    1    0    1    0    0    1    0    0     1
## The genes after mutation:
##        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## v1Copy    1    0    1    0    1    1    0    1    1     0
## v2Copy    0    0    1    0    0    0    0    1    1     1
```

```
testBest <- function(){
  X <- mtcars[,2:11]
  y <- mtcars[,1]
  currentPool <- popInitialize(popSize = 100, geneLength = dim(X)[2], zeroToOneRatio = 1)
  best(X, y, pool = currentPool, popSize = 100, type = "lm", criterion = "AIC")
}
testBest()

##
## Call:
## lm(formula = formula, data = cbind(y, X[, index2]))
##
## Coefficients:
## (Intercept)         disp           hp           wt         qsec
##     14.3619       0.0112      -0.0212      -4.0843       1.0069
##          am
##      3.4705
```

Main Function: select

To test the overall function, we first tested it against stepwise regression result. We did variable selection on the ”mtcars” dataset using our function and stepwise regression. The results are shown below:

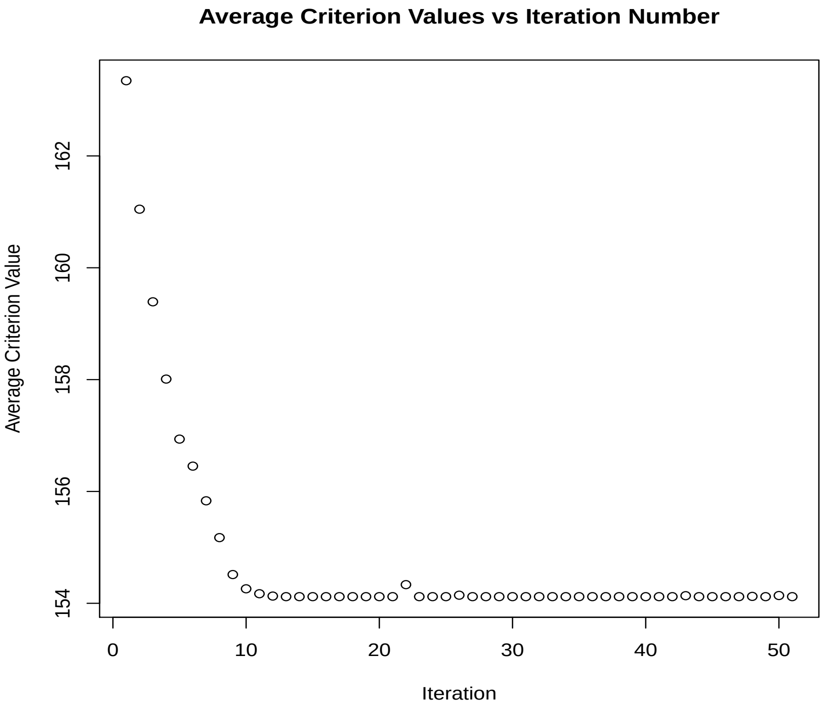
```r
testStepwise = function(){
  ##### Implement our function on the mtcars dataset ######
  ##### Using stepwise regression on the same dataset and compare the results #####
  X <- mtcars[,2:11]
  y <- mtcars[,1]

  cat("Testing select() function on mtcars dataset ... \n")
  cat("Our function running ...")
  set.seed(2)
  result <- select(X, y, popSize = 100, max_iterations = 500, crossRate = 0.95, mRate = 0.0001)
  cat("Now we implement stepwise regression on the dataset.")
  fullModel <- lm( mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)
  stepResult <- step(fullModel, direction = "both", trace = FALSE)
  cat("The stepwise regression has picked the following model:")
  print(summary(stepResult))
  cat("The AIC value for this model is:",AIC(stepResult),"\n")
  cat("\n")
  cat("Our function has chosen the following model:")
  print(summary(result))
  cat("The AIC value for our model is:",unlist(AIC(result)),"\n")
  cat("\n")

  if((abs(AIC(result)-AIC(stepResult))) < 10)
    cat("The model our function chose is close to the one that stepwise regression chose. Test succeede
  else
    cat("The model our function chose is not close to the one that stepwise regression chose. Test faile
}
testStepwise()
```

```
## Testing select() function on mtcars dataset ...
## Our function running ...
```

```
## Warning:  executing %dopar% sequentially:  no parallel backend registered
```



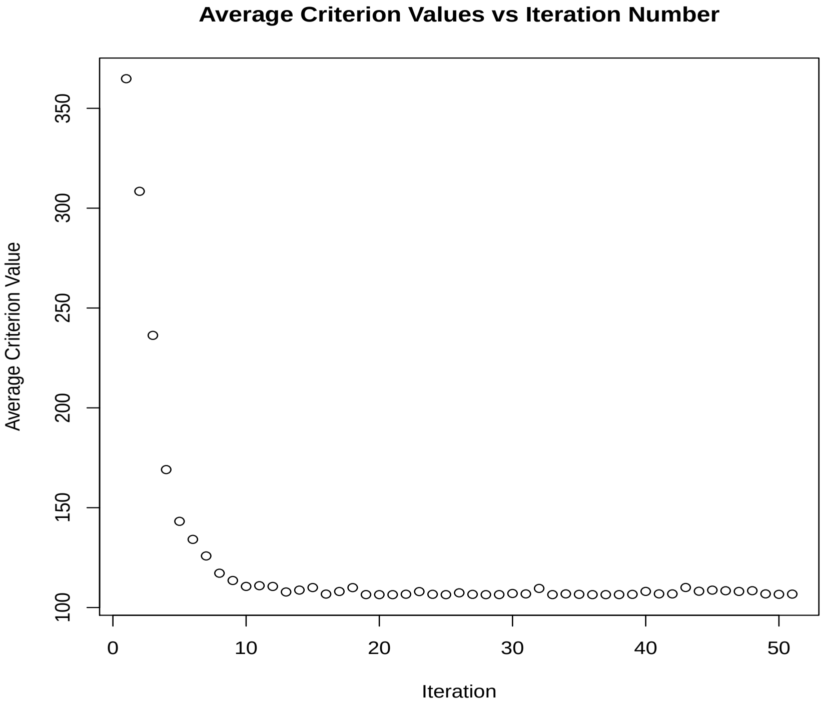
```
## Now we implement stepwise regression on the dataset.The stepwise regression has picked the following
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## am             2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
##
## The AIC value for this model is: 154.1
##
## Our function has chosen the following model:
## Call:
## lm(formula = formula, data = cbind(y, X[, index2]))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## am             2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
##
## The AIC value for our model is: 154.1
##
## The model our function chose is close to the one that stepwise regression chose. Test succeeded.
```

Another test of the overall function is done through a simulated dataset. We generated an outcome variable based on the first 5 variables in the "mtcars" data set. Then we used the 5 variables together with 6 other "noise" variables as the variable pools for our function to select from. The results are shown below:

```
testSim1 <- function(){
  ##### Simulate outcome variable based on 5 predicting variables #####
  ##### Throw in 6 more "noise" variables and use our function to select the predictor variables #####
  X <- mtcars[,1:11]
  n <- dim(mtcars)[1]
  set.seed(1)
  error <- matrix(rnorm(n),nrow = n)
  y <- 1*X[,1] + 2*X[,2] + 3*X[,3] + 4*X[,4] + 5*X[,5] + error

  cat("Testing our function on the simulated dataset ...\n")
  cat("Function is running ...\n")
  set.seed(1)
  result <- select (X, y, popSize = 200, max_iteration = 200, criterion = "BIC", zeroToOneRatio = 1, cr
  cat("Our function has chosen the following model:")
  print(summary(result))
  cat("The BIC value for our model is:",unlist(BIC(result)),"\n")
  cat("The true model has the mpg, cyl, disp, hp and drat as the independent variables.\n")
  cat("Using the current seed, our function has picked out all of the 5 relevant variables\nbut include
}
testSim1()
```

```
## Testing our function on the simulated dataset ...
## Function is running ...
```



```
## Our function has chosen the following model:
## Call:
## lm(formula = formula, data = cbind(y, X[, index2]))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.886 -0.503  0.104  0.488  1.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20296    2.87015    0.07    0.944
## mpg          1.07368    0.06441   16.67  4.7e-15 ***
## cyl          1.73786    0.27023    6.43  9.8e-07 ***
## disp         3.00703    0.00393  765.29  < 2e-16 ***
## hp           3.99140    0.00660  604.97  < 2e-16 ***
## drat         4.62214    0.54096    8.54  6.9e-09 ***
## carb         0.37456    0.20097    1.86    0.074 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.937 on 25 degrees of freedom
## Multiple R-squared:     1,Adjusted R-squared:     1
## F-statistic: 2.18e+06 on 6 and 25 DF,  p-value: <2e-16
##
## The BIC value for our model is: 106.5
## The true model has the mpg, cyl, disp, hp and drat as the independent variables.
## Using the current seed, our function has picked out all of the 5 relevant variables
## but included 1 additional irrelevant variable. The performance of our function is decent. Test succe
```

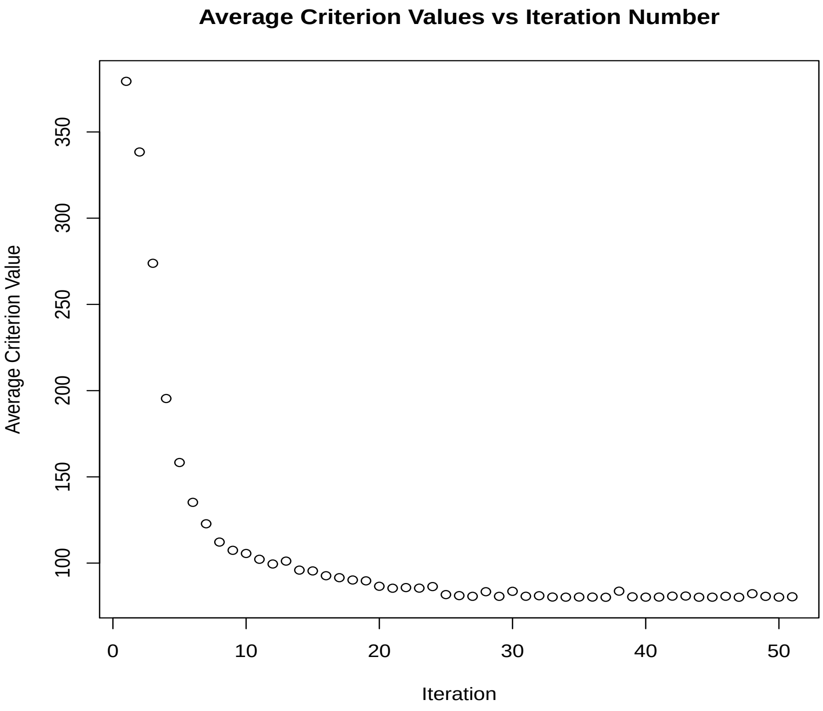
Finally, we further added 30 ”noise” variables to the variable pool. So we now have 5 predictor variables

and 36 ”noise” variables in the variable pool.
We tested our function on this dataset.

```
testSim2 <- function(){
  ##### Simulate outcome variable based on 5 predicting variables #####
  ##### Throw in 36 more "noise" variables and use our function to select the predictor variables #####
  set.seed(2)
  X1 <- mtcars[,1:11]
  n <- dim(mtcars)[1]
  X2 <- as.data.frame(matrix(sample(0:100, 20*n,replace = T),nrow = n))
  X <- cbind(X1,X2)
  error <- rnorm(n)
  y <- 1*X[,1] + 2*X[,2] + 3*X[,3] + 4*X[,4] + 5*X[,5] + error

  cat("Testing our function on the simulated dataset ...\n")
  cat("Function is running ...\n")
  set.seed(1)
  result <- select (X, y, popSize = 200, max_iteration = 200, criterion = "BIC", zeroToOneRatio = 1, cr
  cat("Our function has chosen the following model:")
  print(summary(result))
  cat("The BIC value for our model is:",unlist(BIC(result)),"\n")
  cat("The true model has the mpg, cyl, disp, hp and drat as the independent variables.\n")
  cat("Using the current seed, our function has picked out all of the 5 relevant\nvariables but include
}
testSim2()

## Testing our function on the simulated dataset ...
## Function is running ...
```



```
## Our function has chosen the following model:
## Call:
## lm(formula = formula, data = cbind(y, X[, index2]))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8758 -0.3374 -0.0324  0.2914  1.2551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.77721    1.82253    0.98    0.340
## mpg          1.06587    0.04135   25.78  < 2e-16 ***
## cyl          1.63794    0.18558    8.83  7.6e-09 ***
## disp         3.00452    0.00253 1187.50  < 2e-16 ***
## hp           4.00169    0.00413  968.57  < 2e-16 ***
## drat         4.60487    0.34325   13.42  2.3e-12 ***
## carb         0.23443    0.12736    1.84    0.079 .
## V3          -0.01265    0.00435   -2.91    0.008 **
## V9          -0.00645    0.00363   -1.78    0.089 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.581 on 23 degrees of freedom
## Multiple R-squared:     1,Adjusted R-squared:     1
## F-statistic: 4.25e+06 on 8 and 23 DF,  p-value: <2e-16
##
## The BIC value for our model is: 80.16
## The true model has the mpg, cyl, disp, hp and drat as the independent variables.
## Using the current seed, our function has picked out all of the 5 relevant
## variables but included 1 additional irrelevant variable.
## The performance of our function is decent. Test succeeded.
```

4 Contributions

Everyone in the group contributed relatively equally to each aspect of the project. Some efforts were more focused as follows:

1. Code Writing
   (a) best(): Kevin
   (b) crossover(): Greta
   (c) evalFunction(): Kevin (First Iteration), Joy (Second Iteration), Kevin (Third Iteration), Greta (Final Iteration)
   (d) mutation(): Greta
   (e) popInitialize(): Kevin
   (f) select(): Kevin (First Iteration), Greta (Second Iteration)
   (g) singleEval(): Greta
   (h) updateSamp(): Greta
2. Code Testing

   (a) test(): Joy

3. Documentation
   (a) Introduction: Kevin
   (b) Code Structure: Greta
   (c) Testing: Joy, Kevin, Greta
   (d) Contributions: Kevin