STAT452/652 Solution to Assignment 1 - Part 2

Due on Oct 16, 2021

1 Applications

1.1 Question 1

The variables included in the best model of each size and their corresponding BIC values are as follows.

```
library(leaps)

### Prepare data for model fitting
X.mat = model.matrix(Ozone ~ ., data = AQ)
Y = AQ$Ozone

### Fit all-subsets sequence
fit.subsets = regsubsets(x = X.mat, y = Y, intercept = F)

### Extract the models' variables
info.subsets = summary(fit.subsets)
seq.subsets = info.subsets$which
vars.seq.subsets.raw = apply(seq.subsets, 1, function(W){
  vars.list = names(W)[W]
  output = paste0(vars.list, collapse = ", ")
})

### Clean-up the output
library(stringr)
vars.seq.subsets = str_replace_all(vars.seq.subsets.raw, "\\(Intercept\\)", "Intercept")
vars.seq.subsets = str_replace_all(vars.seq.subsets, "TWrat", "TW-Ratio")
vars.seq.subsets = str_replace_all(vars.seq.subsets, "TWcp", "TW-CrossProd")

### Store variables in a data frame and add BICs
BICs.subsets = info.subsets$bic
data.subsets = data.frame(Vars = vars.seq.subsets,
  BIC = signif(BICs.subsets, 4))
print(data.subsets)
```

```
##                                                      Vars    BIC
## 1                                                TW-Ratio -185.2
## 2                                       Solar.R, TW-Ratio -189.1
## 3                               Intercept, Temp, TW-Ratio -204.2
## 4                      Intercept, Solar.R, Temp, TW-Ratio -207.1
## 5            Intercept, Solar.R, Wind, Temp, TW-CrossProd -204.6
## 6 Intercept, Solar.R, Wind, Temp, TW-CrossProd, TW-Ratio -202.9
```

The best model appears to be the one with four predictors: solar radiation, temperature, temperature/wind speed, and an intercept.

1.2 Question 2

```r
### Fit smallest and largest models for stepwise
fit.start = lm(Ozone ~ 1, data = AQ)
fit.end = lm(Ozone ~ ., data = AQ)

### Fit stepwise sequence
fit.step = step(fit.start, list(upper = fit.end), trace = 0)

### Extract chosen predictors
vars.step.raw = names(fit.step$coef)

### Clean-up names
vars.step = str_replace_all(vars.step.raw, "\\(Intercept\\)", "Intercept")
vars.step = str_replace_all(vars.step, "TWrat", "TW-Ratio")
vars.step = str_replace_all(vars.step, "TWcp", "TW-CrossProd")
vars.step = paste0(vars.step, collapse = ", ")
```

The model that is chosen by the default settings in stepwise contains: Intercept, TW-Ratio, Temp, Solar.R.

1.3 Question 3

```r
### Create function which constructs folds for CV
### n is the number of observations, K is the number of folds
get.folds = function(n, K) {
  ### Get the appropriate number of fold labels
  n.fold = ceiling(n / K) # Number of observations per fold (rounded up)
  fold.ids.raw = rep(1:K, times = n.fold)
  fold.ids = fold.ids.raw[1:n]

  ### Shuffle the fold labels
  folds.rand = fold.ids[sample.int(n)]

  return(folds.rand)
}

### Create function to compute MSPEs
get.MSPE = function(Y, Y.hat){
  return(mean((Y - Y.hat)^2))
}
```

The CV MSPE for each fold, as well as the average across all folds, is as follows.

```r
K = 10 #Number of folds

### Create a container to store MSPEs, with one extra space for the average
all.CV.MSPEs = rep(0, times = K+1)
names(all.CV.MSPEs) = c(paste0("Fold ", 1:10), "Average")

### Get CV fold labels
set.seed(2928893)
n = nrow(AQ)
folds = get.folds(n, K)
```

```r
### Perform cross-validation
for (i in 1:K) {
  ### Get training and validation sets
  data.train = AQ[folds != i, ]
  data.valid = AQ[folds == i, ]
  Y.valid = data.valid$Ozone

  ################
  ### Stepwise ###
  ################

  ### Fit minimum, maximum and optimal
  fit.start = lm(Ozone ~ 1, data = data.train)
  fit.end = lm(Ozone ~ ., data = data.train)
  fit.step = step(fit.start, list(upper = fit.end), trace = 0)

  ### Get predictions and MSPE
  pred.step = predict(fit.step, data.valid)
  MSPE.step = get.MSPE(Y.valid, pred.step)
  all.CV.MSPEs[i] = MSPE.step
}

### Compute the average MSPE across all K folds
all.CV.MSPEs["Average"] = mean(all.CV.MSPEs[-(K+1)])

print(all.CV.MSPEs)
```

```
##    Fold 1    Fold 2    Fold 3    Fold 4    Fold 5    Fold 6    Fold 7    Fold 8
##  328.0475  403.9859  577.2344  106.9023  178.0752  369.4887  233.2169  634.9388
##    Fold 9   Fold 10   Average
## 1074.3133  786.7373  469.2940
```