Lab3

GroupB:Kyla Hayworth, Lyn Peterson, Yilin Li

2019/9/24

```r
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")

group_B_process <- function(training_data) {

  # create transformed data columns
  training_data = mutate(training_data,NumIllegsr = sqrt(NumIlleg))

}


group_B_fit <- function(training_data) {

  # process data first
  training_data = group_B_process(training_data)

  # run lm() to fit your model.
  lm(ViolentCrimesPerPop ~ racePctWhite + PctKids2Par +NumIllegsr, data = training_data)

}

group_B_MSE <- function(model, data) {

  # process the data first
  data = group_B_process(data)

  # find true values and predicted values
  p = predict(model, data)
  true_values = data$ViolentCrimesPerPop

  # return the MSE value
  mean((p - true_values)^2)
}

group_B_automated_fit <- function(data){

  # delete columns with '?' in it
  new_data = data[,sapply(data, is.numeric)]

  # create two subsets with models using forward selection and backward selection seperately.
  forward = regsubsets(ViolentCrimesPerPop ~ ., data = new_data, nvmax = 25, method = "forward")
  backward = regsubsets(ViolentCrimesPerPop ~ ., data = new_data, nvmax = 25, method = "backward")

  # select out the best model with the lowest BIC
  minBIC_forward = min(summary(forward)$bic)
  minBIC_backward = min(summary(backward)$bic)
  if(minBIC_backward < minBIC_forward){
    index = which.min(summary(backward)$bic)
    model = coef(backward,index)
  } else {
    index = which.min(summary(forward)$bic)
    model = coef(forward,index)
  }

  # resulting model is below:
  #(Intercept)          state   racePctWhite       pctUrban      PctEmploy MalePctDivorce   FemalePctDi
  #0.200527854   -0.001300984   -0.188668931    0.054425195   -0.157910910    0.354996125   -0.04147146
  #PctIlleg   PctHousOccup      NumStreet
  #0.375654821   -0.085362549    0.197275651


  # fit this model with lm()
  lm(ViolentCrimesPerPop ~ state + racePctWhite + pctUrban + PctEmploy + MalePctDivorce + FemalePctDiv
}
```