Exercises5stsci4520

Nick Gembs

2/24/2023

Statistical Computing - Exercises 05 - Variable Selection

In these exercises, you will implement a regression variable selection procedure called forward selection. You will apply your variable selection procedure to the Ames, Iowa housing dataset, which can be read into R with (after you've installed the modeldata package):

```
#install.packages("modeldata")
data("ames", package = "modeldata")
```

Forward selection starts with a model with only an intercept, and then picks the covariate that, when added to the model, minimizes an *information criterion*, such as the Akaike Information Criterion (AIC). After that covariate is added, it attempts to add the covariate from the remaining set of covariates that when added to the model, makes the information criterion smallest. This process is continued until none of the remaining covariates decrease the information criterion.

1. Look up the definition of the AIC (Wikipedia is fine), and write a function that takes in an object of class `lm` (the result of running the `lm` function to fit a linear model), and returns the AIC value. There is a function in R called `logLik` that might be helpful.

```
lmtest <- lm(ames$Lot_Area ~ ames$Year_Built)

AICfunc = function (model, k=2) {
  # AIC is -2*logLik(lmtest)+2*df, but there is already a built in AIC function
  AIC(model, k=k)
}

AICfunc(lmtest)
```

```
## [1] 60894.81
```

2. Write a function that takes in a string indicating the response variable, and a vector of strings indicating covariate values, and returns a string containing a model formula, such as `"Sale_Price ~ Gr_Liv_Area + Total_Bsmt_SF"`

```
variables <- function (response, covariates) {
  str = paste(response, "~")
  if(is.null(covariates)){
    str = paste(str, "NULL")
    return(str)
```

```
  } else {
  for (i in 1:(length(covariates)-1)) {
    str = paste(str, covariates[i])
    str = paste(str, " + ")
  }
  end = tail(covariates, n=1)
  str = paste(str,end)
  return(str)
}
}
variables("Sale_Price", c("GR_Liv_Area", "Total_Bsmt_SF"))
```

```
## [1] "Sale_Price ~ GR_Liv_Area  +  Total_Bsmt_SF"
```

```
variables("Sale_Price", NULL)
```

```
## [1] "Sale_Price ~ NULL"
```

3. Write a function that takes in a string response, two vectors of string covariates, a dataset, and returns the covariate from the second set that decreases the AIC most when added to a model containing the first set of covariates. If none decrease the AIC, the function should return NULL or a length-zero vector.

```
AICdec <- function (response, startingcovs, addedcovs, dataset){


  initAIC = AICfunc(lm(variables(response,startingcovs), data = dataset))
  mincov = NULL

  for (elem in addedcovs){
    newAIC = AICfunc(lm(variables(response,c(startingcovs,elem)),
                        data = dataset))
    if (newAIC < initAIC) {
      mincov = elem
      initAIC = newAIC
    }
  }

  return(mincov)
}


AICdec("Lot_Area", c("Lot_Frontage", "Year_Built"),c("Lot_Config","Neighborhood"
                                          ,"Overall_Cond","First_Flr_SF"),ames)
```

```
## [1] "Neighborhood"
```

4. Use these functions to write a function implementing the forward selection algorithm on the Ames housing data, using both Sale_Price and log Sale_Price as the response. Your forward selection function should take in a response string, a character vector of candidate covariates (in case you don't want to test all the covariates, or want to try some interactions), a data frame, and return the vector of selected covariates. Your function should check whether the candidate covariates are available in the data frame, and return an error with an informative error message if not.

```r
ames$logSale_Price = log(ames$Sale_Price)

forwardselection <- function (response, covariates, dataset){

  tryCatch({
    finalcovs <- AICdec(response,NULL, covariates, dataset = dataset)
    covariates = covariates[!(covariates %in% finalcovs)]
    lastadded <- finalcovs

  while (!is.null(lastadded)) {
    lastadded <- AICdec(response,finalcovs, covariates, dataset = dataset)
    covariates = covariates[!(covariates %in% finalcovs)]
    if (!is.null(lastadded)) {finalcovs = c(finalcovs, lastadded)}
  }

  cat("Forward Selection Covariants:\n\n")
  print(finalcovs)
  cat("\n\n")
  } ,error = function(e){
    message("Not all covariates exist in the dataframe:")
    message("Output error message:")
    message(e)
    stop()
  })


  logresponse = paste("log",response, sep="")

  tryCatch({
    finalcovs <- AICdec(logresponse,NULL, covariates, dataset = dataset)
    covariates = covariates[!(covariates %in% finalcovs)]
    lastadded <- finalcovs

  while (!is.null(lastadded)) {
    lastadded <- AICdec(logresponse,finalcovs, covariates, dataset = dataset)
    covariates = covariates[!(covariates %in% finalcovs)]
    if (!is.null(lastadded)) {finalcovs = c(finalcovs, lastadded)}
  }

  cat("Forward Selection log response Covariants:\n\n")
  print(finalcovs)
  return()
  } ,error = function(e){
    message("Not all covariates exist in the dataframe:")
    message("Output error message:")
    message(e)
    stop()
  })


}
```

```r
forwardselection("Sale_Price",covariates = (names(ames)[c(-72,-75)]),
                 dataset = ames)
```

```
## Forward Selection Covariants:
##
##  [1] "Neighborhood"   "Gr_Liv_Area"    "Bsmt_Exposure"  "MS_SubClass"
##  [5] "Year_Built"     "Overall_Cond"   "Roof_Matl"      "Misc_Feature"
##  [9] "Total_Bsmt_SF"  "Bsmt_Unf_SF"    "Sale_Condition" "Garage_Area"
## [13] "Condition_2"    "Mas_Vnr_Area"   "Pool_QC"        "Exterior_1st"
## [17] "BsmtFin_Type_1" "Functional"     "Fireplaces"     "Condition_1"
## [21] "Garage_Finish"  "Bedroom_AbvGr"  "Land_Contour"   "Screen_Porch"
## [25] "Mas_Vnr_Type"   "BsmtFin_Type_2" "Lot_Area"       "Roof_Style"
## [29] "Year_Remod_Add" "Second_Flr_SF"  "Land_Slope"     "Latitude"
## [33] "TotRms_AbvGrd"  "Kitchen_AbvGr"  "Garage_Cars"    "Street"
## [37] "First_Flr_SF"   "Bldg_Type"      "Heating_QC"     "Central_Air"
## [41] "BsmtFin_SF_2"   "Wood_Deck_SF"   "Half_Bath"      "Full_Bath"
##
##
## Forward Selection log response Covariants:
##
##  [1] "Garage_Type"        "Foundation"         "MS_Zoning"
##  [4] "Open_Porch_SF"      "Bsmt_Full_Bath"     "House_Style"
##  [7] "Sale_Type"          "Lot_Shape"          "Lot_Frontage"
## [10] "Paved_Drive"        "Exter_Cond"         "Electrical"
## [13] "Lot_Config"         "Exterior_2nd"       "Bsmt_Cond"
## [16] "Fence"              "Garage_Cond"        "BsmtFin_SF_1"
## [19] "Heating"            "Misc_Val"           "Bsmt_Half_Bath"
## [22] "Three_season_porch" "Enclosed_Porch"     "Pool_Area"

## NULL
```