HW4

Yu-Xiang, Lin(ID: B06305030)

2020/11/8

library(tidyverse)

## ─ Attaching packages ────────────────── tidyverse 1.3.0 ─

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.0.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## ─ Conflicts ─────────────────── tidyverse\_conflicts() ─  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 3.0-1

# Analyze all the spam data using variable selection methods including (a) forward stepwise regression and backward stepwise regression. (12%) (b) Randomly select two thirds of the data as the training dataset and the remaining one third as the test data. (set.seed(1001)) (i) Use the training data to identify the lambda value that gives minimum cross-validated error for lasso regression, ridge regression, and elastic net regression with different alpha values. (20%) (ii) Apply the lambda with minimum cross-validated error to the test data and evaluate their mean square errors. (20%)

## (a)

p1 <- read.table("/Users/linyuxiang/R/MSA/Data/spam.dat", header = T)  
p1.fwd <- step(lm(Y ~ 1, data=p1),~X.1+X.2+X.3+X.4+X.5+X.6+X.7+X.8+X.9+X.10+X.11+X.12+X.13+X.14+X.15+X.16+X.17+X.18+X.19+X.20+X.21+X.22+X.23+X.24+X.25+X.26+X.27+X.28+X.29+X.30+X.31+X.32+X.33+X.34+X.35+X.36+X.37+X.38+X.39+X.40+X.41+X.42+X.43+X.44+X.45+X.46+X.47+X.48+X.49+X.50+X.51+X.52+X.53+X.54+X.55+X.56+X.57, data=p1, direction="forward")  
fw\_res <- p1.fwd$model %>% names()  
# backward  
p1.lrs <- lm(Y~., data = p1)  
pr.bwd <- step(p1.lrs, direction = "backward")  
bw\_res <- pr.bwd %>% names()

intersect(fw\_res, bw\_res)

## character(0)

setdiff(fw\_res, bw\_res)

## [1] "Y" "X.21" "X.23" "X.7" "X.57" "X.16" "X.52" "X.25" "X.5" "X.53"  
## [11] "X.8" "X.24" "X.6" "X.20" "X.22" "X.42" "X.18" "X.27" "X.46" "X.49"  
## [21] "X.45" "X.19" "X.33" "X.9" "X.26" "X.12" "X.17" "X.37" "X.44" "X.3"   
## [31] "X.4" "X.48" "X.47" "X.43" "X.1" "X.2" "X.35" "X.55" "X.40" "X.30"  
## [41] "X.11" "X.54" "X.38" "X.10" "X.50" "X.56" "X.34" "X.39"

setdiff(bw\_res, fw\_res)

## [1] "coefficients" "residuals" "effects" "rank"   
## [5] "fitted.values" "assign" "qr" "df.residual"   
## [9] "xlevels" "call" "terms" "model"   
## [13] "anova"

* Ans: Forward和Backward方法篩選出來的變數大致相同，有X.21, X.23, X.7, X.57, X.16, X.52, X.25, X.5, X.53, X.8, X.24, X.6, X.20, X.22, X.42, X.18, X.27, X.46, X.49, X.45, X.19, X.33, X.9, X.26, X.12, X.17, X.37, X.44, X.3, X.4, X.48, X.47, X.43, X.1, X.2, X.35, X.40, X.30, X.11, X.54, X.38, X.10, X.50, X.56, X.34, X.39，唯一不同的是Forward方法有選出X.55，Backward方法沒有

## (b)

set.seed(1001)  
indx <- sample(1:nrow(p1) ,nrow(p1)\*.67)  
train.p1 <- p1[indx,]  
test.p1 <- p1[-indx,]  
nrow(train.p1)+nrow(test.p1) == nrow(p1)

## [1] TRUE

df <- data\_frame(alpha = rep(NA,11), MSE = NA)

## Warning: `data\_frame()` is deprecated as of tibble 1.1.0.  
## Please use `tibble()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

for (i in 0:10){  
cv\_fit <- cv.glmnet(train.p1[,-58] %>% as.matrix, train.p1[,58] %>% as.matrix, type.measure="mse", alpha=i/10,family="gaussian")  
  
fit <- glmnet(train.p1[,-58] %>% as.matrix, train.p1[,58] %>% as.matrix, family="binomial", alpha=1, lambda = cv\_fit$lambda.min)  
  
fit\_pred <- predict(fit, newx = test.p1[,-58] %>% as.matrix)  
# MSE lasso  
df$MSE[i+1] <- sum((fit\_pred - test.p1$Y)^2) / nrow(fit\_pred)  
df$alpha[i+1] <- (i/10)  
message(i)  
}

## 0

## 1

## 2

## 3

## 4

## 5

## 6

## 7

## 8

## 9

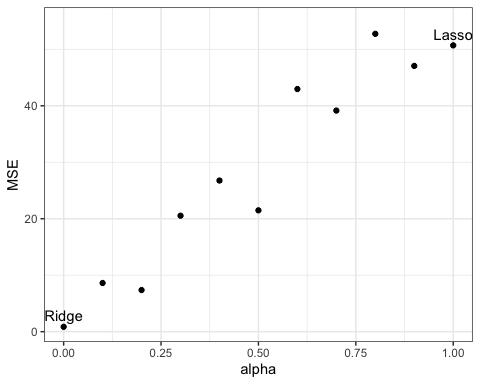
## 10

df$label[1] <- "Ridge"

## Warning: Unknown or uninitialised column: `label`.

df$label[2:10] <- NA  
df$label[11] <- "Lasso"  
df %>%   
 ggplot(aes(x = alpha, y = MSE))+  
 geom\_point()+  
 geom\_text(aes(label = label, y = MSE+2))+  
 theme\_bw()

## Warning: Removed 9 rows containing missing values (geom\_text).



* Ans: 從圖中可以看出選擇Alpha = 0，也就是使用Ridge Regression時MSE為最低，當選擇的Alpha值上升，也就是Lasso方法的比例提升時，模型MSE數值會提高，在使用Lasso Regression，即Alpha = 1時，MSE為最高。由此得知該資料可能較適合使用Ridge Regression進行變數篩選。

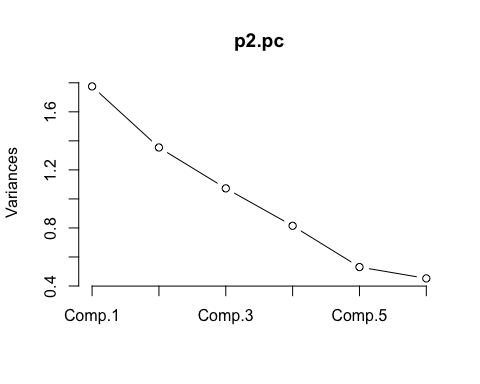
# Problem 2

Carry out a principal component analysis on the engineer data as follows. Ignore groups and use a correlation matrix based on all 40 observations. (48%)(pilots.dat) Table. Comparison of Six Tests on Engineer Apprentices and Pilots

p2 <- read.table("/Users/linyuxiang/R/MSA/Data/pilots.dat", header = F)  
  
p2 <- p2[,-1]  
eigen(cor(p2))

## eigen() decomposition  
## $values  
## [1] 1.7751277 1.3544159 1.0726505 0.8147958 0.5306128 0.4523973  
##   
## $vectors  
## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] -0.40239072 0.3964661 0.4617841 -0.3928149 -0.2103062 0.5187674  
## [2,] 0.09715877 0.7472294 -0.1752970 -0.1315611 -0.2801896 -0.5528697  
## [3,] -0.38541311 -0.2181560 -0.4329575 -0.7177525 0.2585104 -0.1855163  
## [4,] -0.54333623 -0.3144601 -0.1065065 0.2453920 -0.7066663 -0.1869825  
## [5,] 0.31188931 -0.3559400 0.6268314 -0.3992852 -0.2012981 -0.4279773  
## [6,] -0.53629229 0.1062657 0.4053555 0.3058981 0.5201339 -0.4155385

p2.pc <- princomp(p2,cor=TRUE)  
plot(p2.pc, type = "l")



* Ans: 從圖中可以看出第六個Component幾乎沒有變異，此資料若要進一步分析，最後可能只會選擇五個主成分。