annand\_final\_project\_pcr

Joseph Annand

2023-12-16

## Load Libraries

library(ISLR2)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:ISLR2':  
##   
## Boston

library(pls)

## Warning: package 'pls' was built under R version 4.3.2

##   
## Attaching package: 'pls'

## The following object is masked from 'package:stats':  
##   
## loadings

View(Credit)

For the Principal Components regression problem, the Credit data set that is part of the ISLR2 package is used. The focus of this problem is to predict the average credit card balance of a credit holder based on a variety of characteristics. The data set includes 400 records and 10 different features that are not the customer’s average balance. these features include Income, Credit limit, Credit rating, number of cards, Age, Education, home ownership status, student status, marital status, and region. The data set is split in half to form a training data set and test data set.

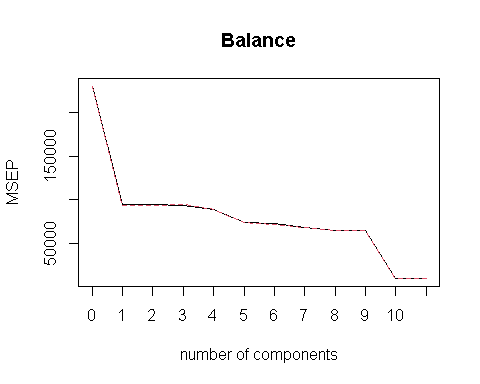
The PCR model is trained using the training data set. Each predictor is standardized prior to generating the principal components. A ten-fold cross-validation error is computed for each possible number of principal components.

set.seed(1)  
train <- sample(1:nrow(Credit), nrow(Credit) / 2)  
test <- (-train)

set.seed(2)  
  
pcr.fit <- pcr(Balance ~ ., data = Credit, subset = train, scale = T,  
 validation = "CV")  
summary(pcr.fit)

## Data: X dimension: 200 11   
## Y dimension: 200 1  
## Fit method: svdpc  
## Number of components considered: 11  
##   
## VALIDATION: RMSEP  
## Cross-validated using 10 random segments.  
## (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps  
## CV 479.6 307.2 307.0 306.9 299.3 273.2 270.3  
## adjCV 479.6 306.7 306.7 307.2 298.2 272.8 268.1  
## 7 comps 8 comps 9 comps 10 comps 11 comps  
## CV 262.2 255.5 254.9 102.8 101.4  
## adjCV 261.8 255.0 254.3 102.4 101.1  
##   
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps  
## X 25.58 40.56 51.56 61.42 70.87 79.34 87.20 94.74  
## Balance 59.81 59.96 60.18 63.82 69.74 70.97 71.72 74.04  
## 9 comps 10 comps 11 comps  
## X 98.19 99.98 100.00  
## Balance 74.33 95.89 96.04

validationplot(pcr.fit, val.type = "MSEP")  
axis(side=1, at=seq(1,20,by=1))



pcr.pred <- predict(pcr.fit, Credit[test, ], ncomp = 11)  
pcr.mse <- mean((pcr.pred - Credit$Balance[test])^2)  
pcr.mse

## [1] 10691.12

sqrt(pcr.mse)

## [1] 103.3979

The validation plot shows that the number of components that minimizes the mean squared error is 11. When the number of components is set equal to 11, the PCR model yields a test MSE of 12275.55, which means that on average, the model is off by $110.80 in its prediction of a customer’s average credit balance.

lm.fit <- lm(Balance ~ ., data = Credit, subset = train)  
lm.pred <- predict(lm.fit, Credit[test, ])  
lm.mse <- mean((lm.pred - Credit$Balance[test])^2)  
lm.mse

## [1] 10691.12

Fitting the data to a least squares regression demonstrates that the PCR model does not improve the test MSE compared to an ordinary multiple linear regression model.