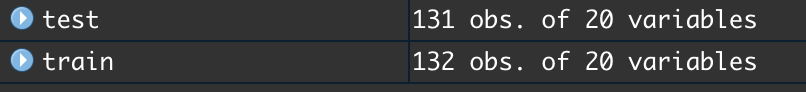
Assignment 4

1. **Export the data set and clean the data**
   1. **Omit the missing value(s)**
   2. **Split the data into train set and test set using the 50-50 split approach**
   3. **Re-express the variables as numeric**

A close up of a sign

Description automatically generated



A close up of text on a black background

Description automatically generated

1. **Summarize the traning data set**
   1. **Compare the mean of predictors. Comments.**

The range of the means is 0.5 to 2510. The means are very different which means that the predictors are used different measurments. In orther words, predictors measured in different unit.

A screenshot of a cell phone

Description automatically generated

* 1. **Compare the variance of the predictors. Comments.**

The range of the means 0.247 to 392970. The variances are very different which means that the predictors are used different measuements.

A screenshot of a cell phone

Description automatically generated

1. **Perform the PCA using the train set.**

A close up of text on a black background

Description automatically generatedA screenshot of a cell phone

Description automatically generated A close up of text on a black background

Description automatically generated

* 1. **How many PCs are enough to explain the majority of the variability of the Hitter data set? And, explain how to choose the appropriate number of PCs.** 
     + **PEV**

A screenshot of a social media post with text and a black background

Description automatically generated

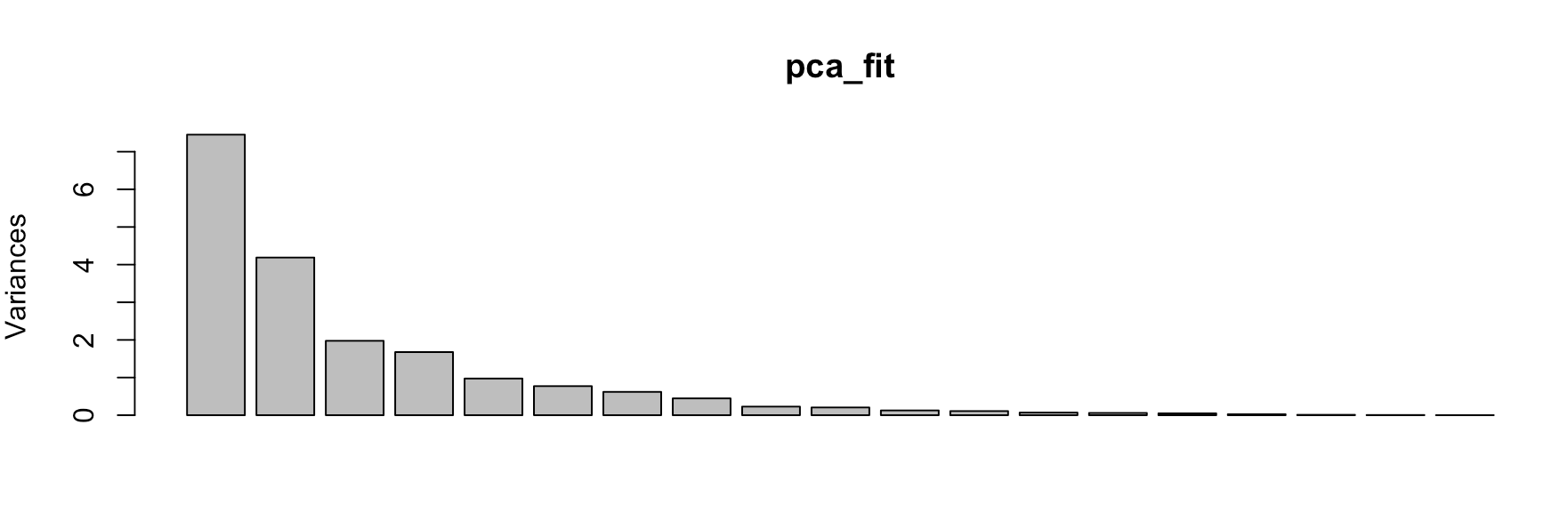
* + - **PCA Summary**

A screenshot of a cell phone

Description automatically generated

* + - **Screeplot(s)**

A black and silver text

Description automatically generatedA close up of a device

Description automatically generated

* 1. **See the result of the “pca\_fit$rotation” result. Which predictors are on the fitst principal components? What predictors are on the second principal components?**

AtBat, Hits, Runs, RBI, Walks, HumRun, PutOuts, new leagueN, LeagueN are in the fist principal components, Errors, Assits, DivisionW, CHmRun, CRBI, CRuns,Chits, CWalks, CAtBat and Year are in the second principal components.

* 1. **Using the first 2 PCs, create a biplot.**

A close up of a map

Description automatically generated

1. **Perform the PCR using the train set.** 
   1. **How many PCs are enough to explain the majority of the variability of the Hitter data set? And, explain how to choose the appropriate number of PCs.** 
      * **RMSEP (smaller is better)**

A picture containing racquetball

Description automatically generated

* + - **MSEP (smaller is better)**

A picture containing racquetball

Description automatically generated

* + - **R-Squared value (larger is better)**

A screenshot of a social media post

Description automatically generated

* 1. **Using the number of PCs from the PCR results in part 1), evaluate its test set performance based on the cross-validation.**



* 1. **Fit PCR on the full data set using the the number of PCs identified by cross-validation in part 2).**

A close up of text on a black background

Description automatically generated

1. **Perform the PLS using the train set.** 
   1. **How many PCs are enough to explain the majority of the variability of the Hitter data set? And, explain how to choose the appropriate number of PCs.**

2 PCs are enough to explain the majority of the variability of the Hitter data set. PC 1,2 are appropriate because they have smaller RMSEP,MSEP and bigger Rsquare. 3rd or more component are considerable bisas- variance trade-off

* + - **RMSEP (smaller is better)**

A close up of a map

Description automatically generated

* + - **MSEP (smaller is better)**

A close up of a map

Description automatically generated

* + - **R-Squared value (larger is better)**

A close up of a womans face

Description automatically generated

* 1. **Using the number of PCs from the PLS results in part 1), evaluate its test set performance based on the cross-validation.**

A screenshot of a cell phone

Description automatically generated

Nocomp = 3, mse = 151753.3 (BEST if you consider the smallest mse)

Nocomp = 2, mse = 146297.9 (BEST if you consider the model variance)

* 1. **Fit PLS on the full data set using the the number of PCs identified by cross-validation in part 2).**

A screenshot of a cell phone

Description automatically generated

With 2 compoeents, explain the variance 51% of the pradictors (Xs) and 45% of the outcome variable(y).

1. **Which method performed better? Justify your answer.**

PCR (BEST if you consider the smallest mse)

PLS (BEST if you consider the model variance)

1. **Which method do you think tends to have lower bias (smaller MSE values)?**

PCR method tends to have lower bias. Smaller mse values

1. **Which method do you think tends to have lower variance ?**

PLS method tends to have lower variance.