hw7.R

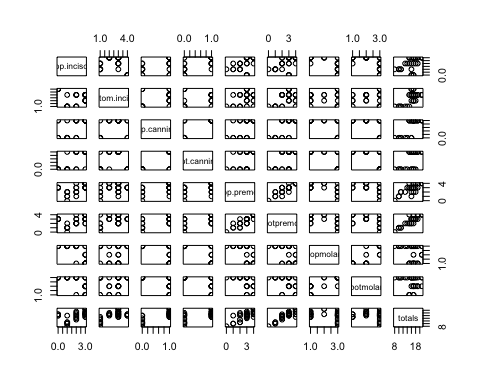
christieburris

Thu Apr 12 20:33:03 2018

#1a. We first output the correlation matrix to look for high values.   
# We observe is high correlation amongst top and bottom sets. For instance, top and bottom cannine (cor=.808)  
# as well as top and bottom premolars (cor=.896). Naturally, the total number of teeth is highly, positively correlated with all other appearances of teeth apart from molars.  
# Molars are negatively correlated with everything besides each other, with the exception of a low positive correlation with bottom incisors.   
  
# Making a plot of this allows us to visualize and interpret these correlations. Correlation with totals seem to reflect a legitimate trend.   
# Top and bottom premolars create a positive lienar trend. Although the data is not presented well for canines, only 3/32 animals only have one of the two teeth.   
# All other animals have either both or none, indicating a high positive correlation.  
  
#1b. We perform PCA to determine the dimensionality of the data. From the scree plot and the cumulative proportions it does not seem like 2 components is adequate.   
  
#1c. The biplot presents the data on the coordinates of the first 2 components. The display is quite skewed to in component 2.   
# This is line with what we discovered from the scree plot. It seems that 2 components do not adequately describe the data.   
  
dataset<-read.csv("~/Desktop/Biometry/hw7/mammals.csv",header=T)  
  
# keep only the numerical data, remove names  
numdata = dataset[,2:10]  
  
# correlation matrix and plot  
round(cor(numdata),3)

## top.incisor bottom.incisor top.cannine bot.cannine  
## top.incisor 1.000 -0.072 0.600 0.743  
## bottom.incisor -0.072 1.000 0.502 0.299  
## top.cannine 0.600 0.502 1.000 0.808  
## bot.cannine 0.743 0.299 0.808 1.000  
## top.premol 0.507 0.384 0.553 0.479  
## botpremol 0.469 0.486 0.660 0.556  
## topmolar -0.697 -0.077 -0.527 -0.653  
## botmolar -0.601 0.018 -0.471 -0.583  
## totals 0.493 0.678 0.765 0.629  
## top.premol botpremol topmolar botmolar totals  
## top.incisor 0.507 0.469 -0.697 -0.601 0.493  
## bottom.incisor 0.384 0.486 -0.077 0.018 0.678  
## top.cannine 0.553 0.660 -0.527 -0.471 0.765  
## bot.cannine 0.479 0.556 -0.653 -0.583 0.629  
## top.premol 1.000 0.896 -0.604 -0.536 0.771  
## botpremol 0.896 1.000 -0.497 -0.414 0.878  
## topmolar -0.604 -0.497 1.000 0.855 -0.307  
## botmolar -0.536 -0.414 0.855 1.000 -0.184  
## totals 0.771 0.878 -0.307 -0.184 1.000

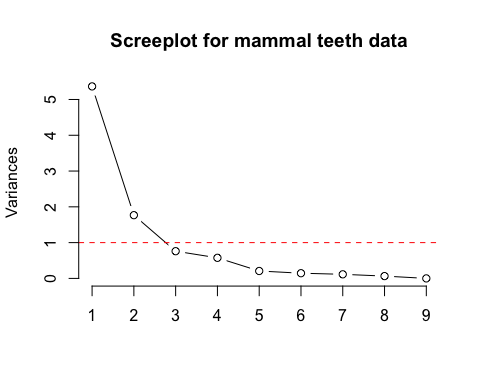
pairs(numdata)



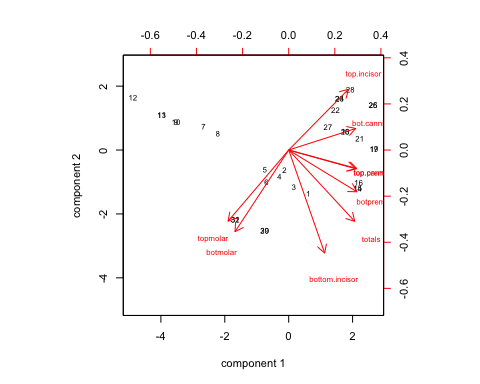
# pca summary, notice cumulative proportions. The majority is picked up by 2 components, but it isn't until 5 components that we pick up 95%.  
pca = prcomp(numdata, center=TRUE,scale=TRUE)  
summary(pca)

## Importance of components%s:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.3164 1.3294 0.87202 0.75847 0.45515 0.38124  
## Proportion of Variance 0.5962 0.1964 0.08449 0.06392 0.02302 0.01615  
## Cumulative Proportion 0.5962 0.7925 0.87702 0.94094 0.96396 0.98011  
## PC7 PC8 PC9  
## Standard deviation 0.33800 0.2545 1.802e-16  
## Proportion of Variance 0.01269 0.0072 0.000e+00  
## Cumulative Proportion 0.99280 1.0000 1.000e+00

#screeplot with horizontal line. 3 components necessary to drop below 1.  
screeplot(pca,type='l',main='Screeplot for mammal teeth data')   
abline(1,0,col='red',lty=2)



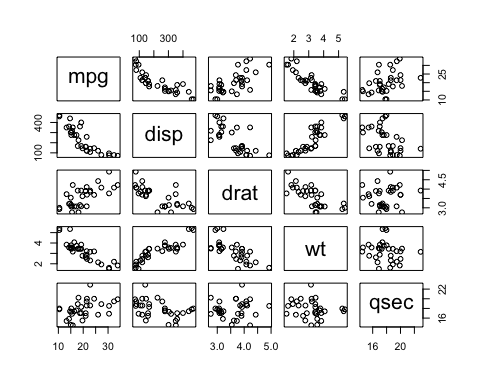
# create biplot, notice that component 2 still has a lot of spread  
par(mfrow=c(1,1),cex=0.7)  
biplot(pca, scale=0, xlab='component 1', ylab='component 2')



#2a. Of the continuous variables (mgp, disp, drat, wt, qsec), we notice high negative correlation between mpg and disp/wt.  
# This makes sense as heavier cars tend to get fewer mpg and cars with more power tend to be heavier.   
# drat is also negatively correlated with wt and disp, but not as highly as mpg.   
  
#2b. PCA indicates 2 components are sufficient to characterize the data. The biplot indicates that one axis contains mostly information on wt, mpg, drat, and disp.   
# The other axis is primarily determined by qsec. As such, two components are sufficient to summarize the data. One might interpret these components as 1) performance and 2) power  
  
attach(mtcars)  
cars=as.matrix(mtcars[,1:11])  
  
# pull continuous variables  
cars.cont=cars[,c(1,3,5:7)]  
  
# correlation matrix and plot  
round(cor(cars.cont),3)

## mpg disp drat wt qsec  
## mpg 1.000 -0.848 0.681 -0.868 0.419  
## disp -0.848 1.000 -0.710 0.888 -0.434  
## drat 0.681 -0.710 1.000 -0.712 0.091  
## wt -0.868 0.888 -0.712 1.000 -0.175  
## qsec 0.419 -0.434 0.091 -0.175 1.000

pairs(cars.cont)



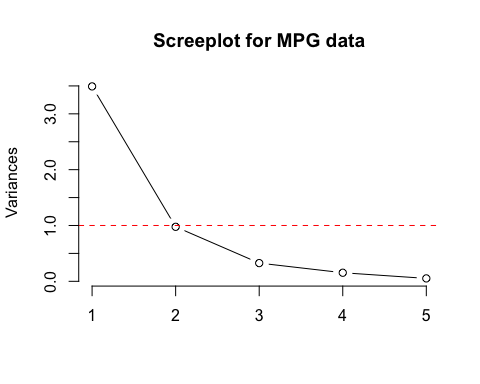
# PCA summary, notice cumulative proportions. The majority is picked up by 2 components, but it isn't until 5 components that we pick up 95%.  
cars.pca = prcomp(cars.cont, center=TRUE,scale=TRUE)  
summary(cars.pca)

## Importance of components%s:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.8685 0.9880 0.5714 0.39202 0.22852  
## Proportion of Variance 0.6983 0.1952 0.0653 0.03074 0.01044  
## Cumulative Proportion 0.6983 0.8935 0.9588 0.98956 1.00000

# extract data from PCA  
eigenvectors = cars.pca$loadings  
eigenvalues = cars.pca$sdev\*cars.pca$sdev  
prop\_var\_exp = eigenvalues/sum(eigenvalues)  
cumulative = cumsum(prop\_var\_exp)  
  
# nice table of eigenvalues, the proportion of the expression from each component and the cumulative contribution of components.   
results = as.data.frame(cbind(eigenvalues, prop\_var\_exp, cumulative))  
round(results,3)

## eigenvalues prop\_var\_exp cumulative  
## 1 3.491 0.698 0.698  
## 2 0.976 0.195 0.894  
## 3 0.326 0.065 0.959  
## 4 0.154 0.031 0.990  
## 5 0.052 0.010 1.000

screeplot(cars.pca,type='l',main='Screeplot for MPG data') #screeplot with dots and lines  
abline(1,0,col='red',lty=2)



par(mfrow=c(1,1),cex=0.7)  
biplot(cars.pca, scale=0, xlab='component 1', ylab='component 2',xlim=c(-4,4),ylim=c(-4,4))

