HW13

109006206

Set Working Directories & Reading Files

```
setwd("/Users/olivia/Documents/Documents/Study/Semester 6/BACS/HW13")
library(readxl)
cars<-read.table("auto-data.txt", header=FALSE, na.strings = "?")
names(cars) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")
cars_log <- with(cars, data.frame(log(mpg), log(cylinders), log(displacement), log(horsepower),
                                  log(weight), log(acceleration), model_year, origin))
cars_log <- na.omit(cars_log)
```

QUESTION 1

A) Let's analyze the principal components of the four collinear variables

1) Create a new data.frame of the four log-transformed variables with high multicollinearity

```
log_collinear <- with(cars_log, data.frame(log.cylinders., log.displacement.,
                                           log.horsepower., log.weight.))
head(log_collinear)
```

```
##   log.cylinders. log.displacement. log.horsepower. log.weight.
## 1       2.079442          5.726848        4.867534    8.161660
## 2       2.079442          5.857933        5.105945    8.214194
## 3       2.079442          5.762051        5.010635    8.142063
## 4       2.079442          5.717028        5.010635    8.141190
## 5       2.079442          5.710427        4.941642    8.145840
## 6       2.079442          6.061457        5.288267    8.375860
```

2) How much variance of the four variables is explained by their first principal component?

```
eigenvalues <- eigen(cov(log_collinear))
variance_explained <- eigenvalues$values / sum(eigenvalues$values)
variance_explained
```

```
## [1] 0.934616867 0.040246293 0.015893283 0.009243557
```

3) Looking at the values and valence (positiveness/negativeness) of the first principal component's eigenvector, what would you call the information captured by this component?

```r
first <- eigenvalues$vectors
rownames(first) = c("log.cylinders","log.displacement","log.horsepower","log.weight")
colnames(first) = c("PC1","PC2","PC3","PC4")
first
```

```
##                         PC1         PC2        PC3         PC4
## log.cylinders    -0.3944484  0.32615343  0.6895416  0.51241263
## log.displacement -0.7221160  0.36134848 -0.1626248 -0.56703525
## log.horsepower   -0.4322835 -0.87289692  0.2158783 -0.06766477
## log.weight       -0.3689037 -0.03319916 -0.6719242  0.64134686
```

Answer : PC1 captures the variation in the data related to the overall size or magnitude of the variables.

B) Let's revisit our regression analysis on cars_log:

1) Store the scores of the first principal component as a new column of cars_log

```r
cars_pca <- prcomp(log_collinear)
cars_log$scores <- cars_pca$x
head(cars_log$scores)
```

```
##             PC1          PC2         PC3         PC4
## [1,] -0.7962713  0.104715883 -0.12092122 -0.01019809
## [2,] -1.0133713 -0.057768937 -0.11577240 -0.06696767
## [3,] -0.8763230 -0.006825013 -0.15925641 -0.05241126
## [4,] -0.8434885 -0.023065253 -0.16716529 -0.02744146
## [5,] -0.8106129  0.034618895 -0.15022034 -0.01604810
## [6,] -1.2987927 -0.148741022 -0.01340663 -0.09102584
```

2) Regress mpg over the column with PC1 scores (replacing cylinders, displacement, horsepower, and weight), as well as acceleration, model_year and origin

```r
summary(lm(log.mpg.~ cars_log$scores[,"PC1"]+log.acceleration. + model_year
           + factor(origin),data = cars_log))
```

```
##
## Call:
## lm(formula = log.mpg. ~ cars_log$scores[, "PC1"] + log.acceleration. +
```

```
##     model_year + factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53593 -0.06148  0.00149  0.06293  0.50928
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.395518   0.172873   8.073 8.84e-15 ***
## cars_log$scores[, "PC1"]  0.387073   0.014110  27.433  < 2e-16 ***
## log.acceleration.        -0.189830   0.043246  -4.390 1.47e-05 ***
## model_year                0.029244   0.001871  15.628  < 2e-16 ***
## factor(origin)2          -0.010840   0.020738  -0.523    0.601
## factor(origin)3           0.002243   0.020517   0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1239 on 386 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8672
## F-statistic: 511.7 on 5 and 386 DF,  p-value: < 2.2e-16
```

3) Try running the regression again over the same independent variables, but this time with everything standardized.

```r
cars_log_standardized <- data.frame(scale(cars_log))
regression_model_standardized <- lm(log.mpg. ~ cars_log$scores[,"PC1"] + log.acceleration. +
                                      model_year + factor(origin),
                                    data = cars_log_standardized)
summary(regression_model_standardized)
```

```
##
## Call:
## lm(formula = log.mpg. ~ cars_log$scores[, "PC1"] + log.acceleration. +
##     model_year + factor(origin), data = cars_log_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57609 -0.18081  0.00438  0.18506  1.49772
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.004201   0.026912   0.156    0.876
## cars_log$scores[, "PC1"]          1.138319   0.041494  27.433  < 2e-16 ***
```

```
## log.acceleration.              -0.101021   0.023014  -4.390 1.47e-05 ***
## model_year                      0.316814   0.020272  15.628  < 2e-16 ***
## factor(origin)0.525710525810929 -0.031878   0.060987  -0.523    0.601
## factor(origin)1.76714743013553   0.006595   0.060336   0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3644 on 386 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8672
## F-statistic: 511.7 on 5 and 386 DF,  p-value: < 2.2e-16
```

How important is this new column relative to other columns?

```
cor_matrix <- cor(cars_log_standardized)
pc1_correlation <- cor_matrix[9,]
pc1_correlation
```

```
##          log.mpg.    log.cylinders. log.displacement.   log.horsepower.
##      8.830302e-01     -9.542881e-01     -9.912446e-01     -9.205490e-01
##       log.weight. log.acceleration.        model_year            origin
##     -9.592987e-01      5.636235e-01      3.497284e-01      6.325888e-01
##        scores.PC1        scores.PC2        scores.PC3        scores.PC4
##      1.000000e+00      2.686578e-16     -1.830919e-15     -4.510778e-15
```

QUESTION 2

A) How much variance did each extracted factor explain?

```
security <- read_excel("security_questions.xlsx",sheet = "data")
pca2 <- prcomp(security,scale. = TRUE)
summary(pca2)$importance[2,]
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
## 0.51728 0.08869 0.06386 0.04233 0.03751 0.03398 0.02794 0.02602 0.02511 0.02140
##    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
## 0.01972 0.01674 0.01624 0.01456 0.01303 0.01280 0.01160 0.01120
```

B) How many dimensions would you retain, according to the two criteria we discussed? (Eigenvalue >= 1 and Scree Plot – can you show the screeplot with eigenvalue=1 threshhold?)

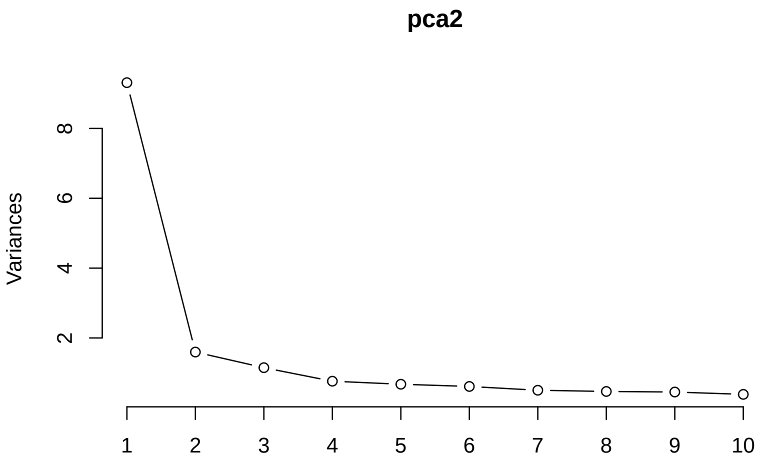
```
eigenvalues2 <- eigen(cor(security))
eigenvalues2$values
```

```
##  [1] 9.3109533 1.5963320 1.1495582 0.7619759 0.6751412 0.6116636 0.5029855
##  [8] 0.4682788 0.4519711 0.3851964 0.3548816 0.3013071 0.2922773 0.2621437
## [15] 0.2345788 0.2304642 0.2087471 0.2015441
```

```
num_dimensions_eigenvalue <- sum(eigenvalues2$values >= 1)
num_dimensions_eigenvalue
```

```
## [1] 3
```

```
screeplot(pca2, type="lines")
```



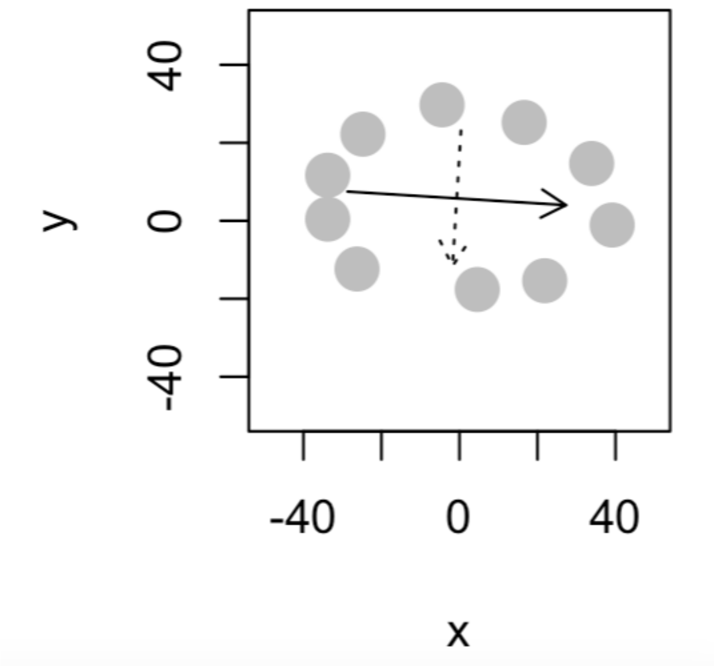
Answer : From the above, I would retain 3 dimensions

C) Can you interpret what any of the principal components mean? Try guessing the meaning of the first two or three PCs looking at the PC-vs-variable matrix

Answer : In general, a principal component is a linear combination of the original variables in a dataset. It is calculated in such a way that it captures the maximum amount of variation or information present in the data. I think the first principal component captures the most significant source of variation in the data.

QUESTION 3

A) Create an oval shaped scatter plot of points that stretches in two directions



B) Can you create a scatterplot whose principal component vectors do NOT seem to match the major directions of variance?