Homework Six

Chiayu Tu (Louis Tu)

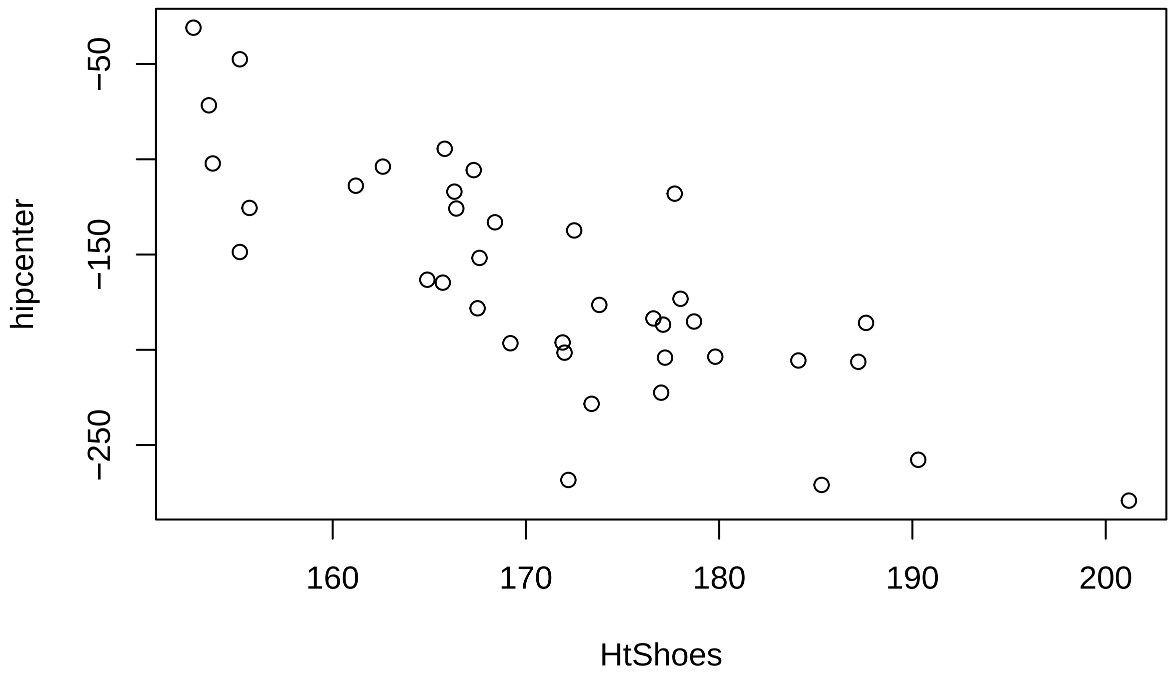
2022-11-25

Question One

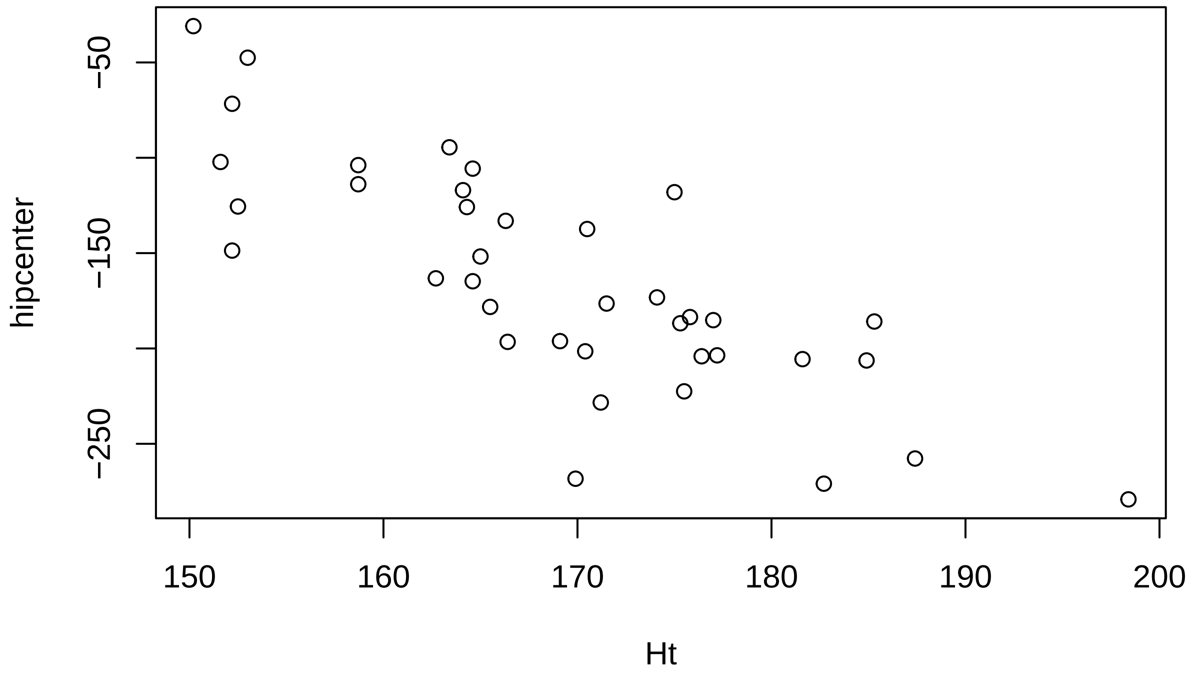
Using the seatpos data, perform a PCR analysis with hipcenter as the response and HtShoes, Ht, Seated, Arm, Thigh and Leg as predictors. Select an appropriate number of components and give an interpretation to those you choose. Add Age and Weight as predictors and repeat the analysis. Use both models to predict the response for predictors taking these values:

Age Weight HtShoes Ht Seated 64.800 263.700 181.080 178.560 91.440 Arm Thigh Leg 35.640 40.950 38.790

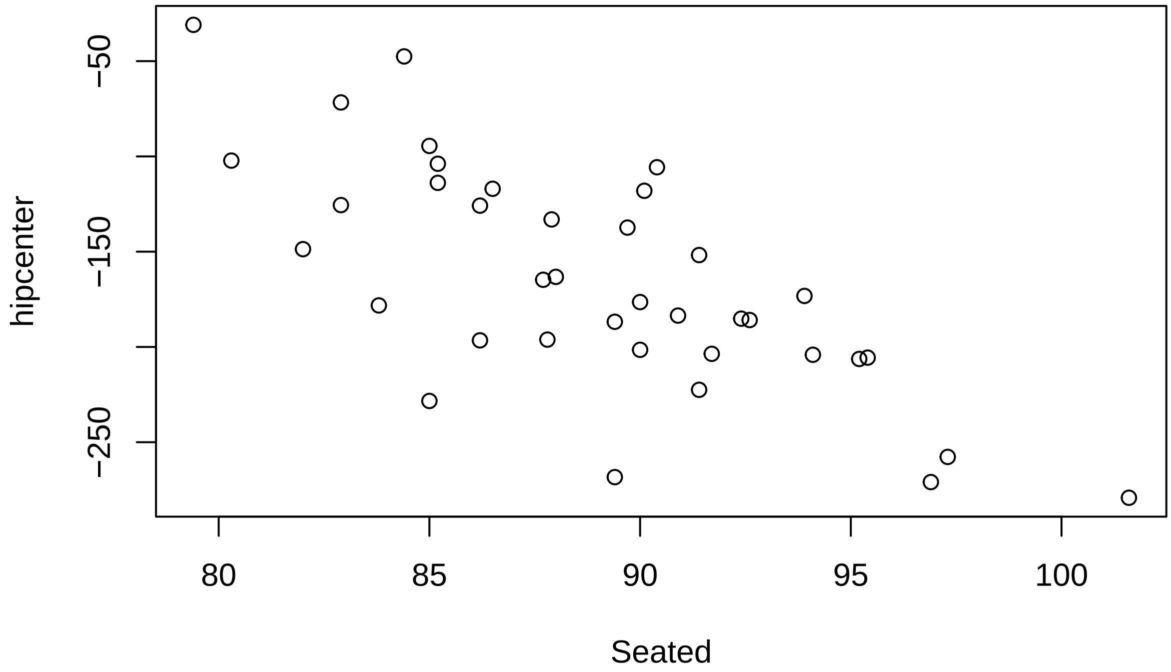
```r
data(seatpos, package = "faraway")
plot(hipcenter ~ HtShoes, seatpos)
```



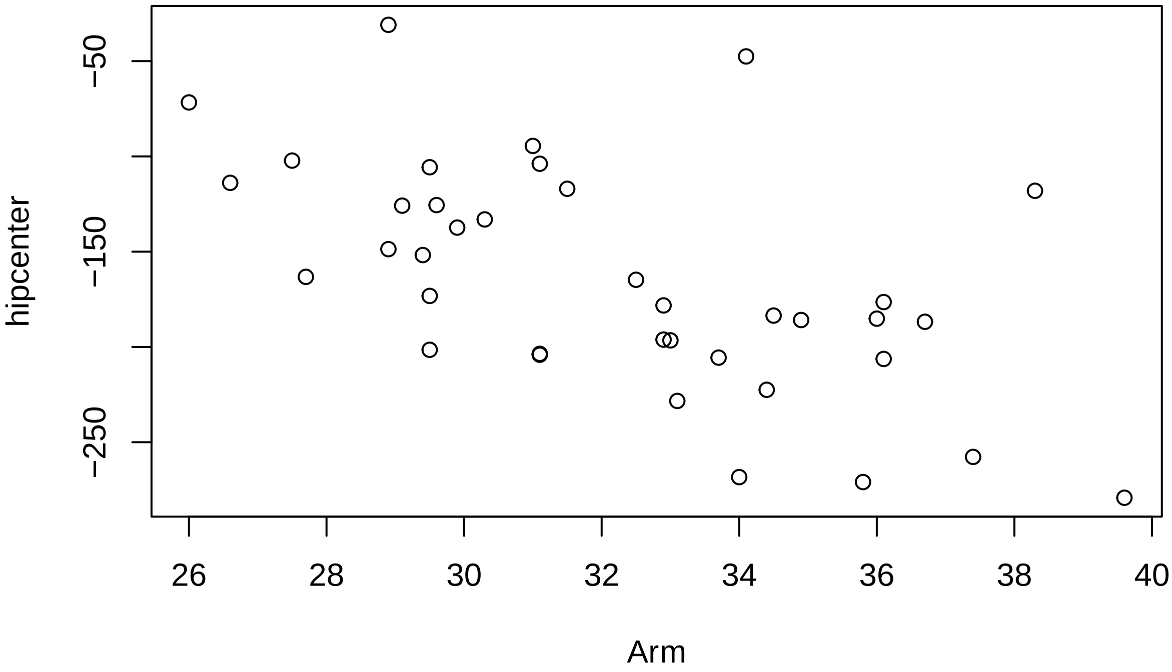
```
plot(hipcenter ~ Ht, seatpos)
```



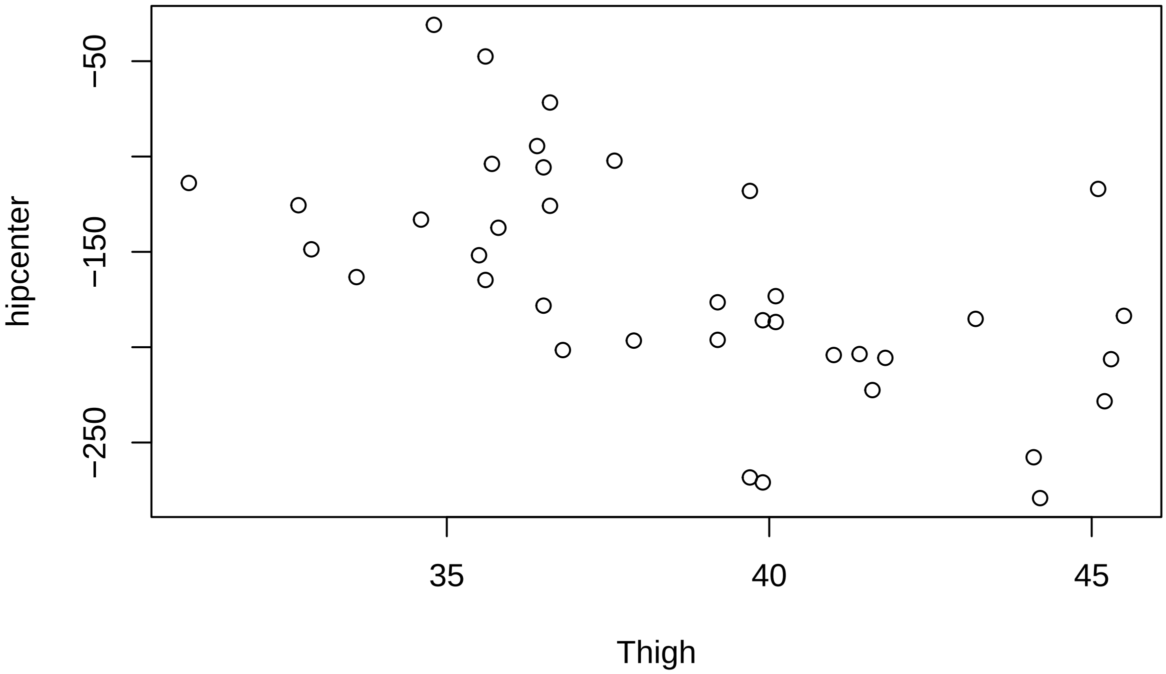
```
plot(hipcenter ~ Seated, seatpos)
```



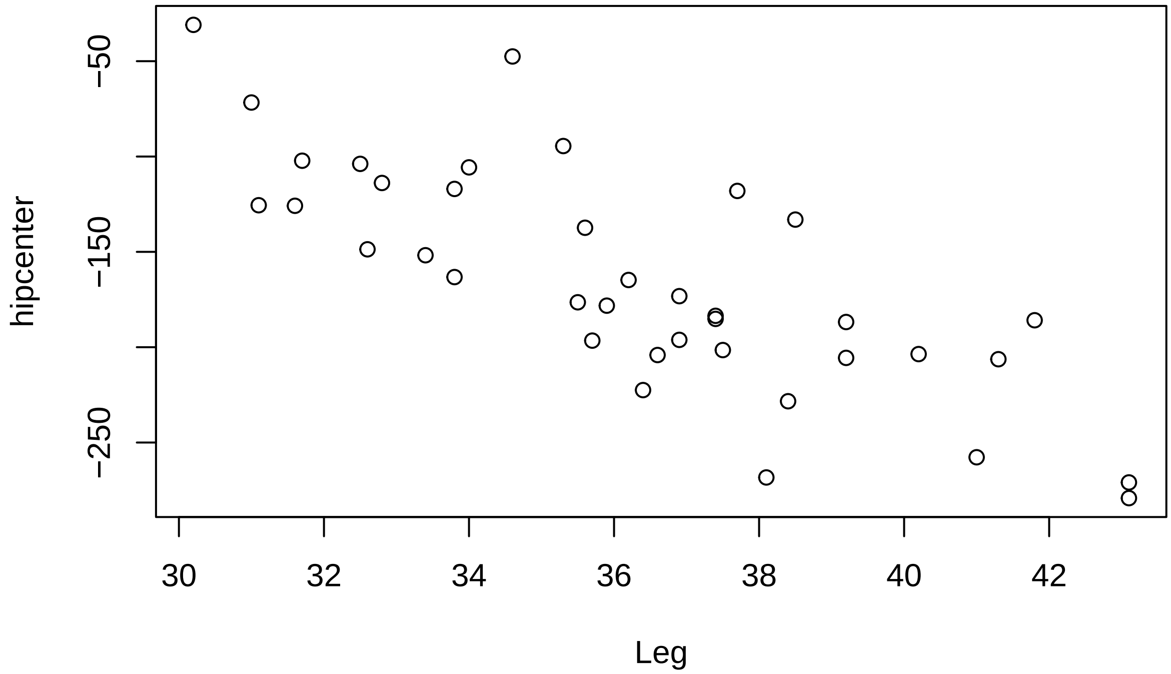
```
plot(hipcenter ~ Arm, seatpos)
```



```
plot(hipcenter ~ Thigh, seatpos)
```



```
plot(hipcenter ~ Leg, seatpos)
```



```r
cseatpos <- seatpos[, c(3,4,5,6,7,8)]
prseatpos <- prcomp(cseatpos)
dim(prseatpos$rotation)
```

```
## [1] 6 6
```

```r
dim(prseatpos$x)
```

```
## [1] 38  6
```

```r
summary(prseatpos)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     17.1573 2.89689 2.11907 1.56412 1.22502 0.46218
## Proportion of Variance  0.9453 0.02695 0.01442 0.00786 0.00482 0.00069
## Cumulative Proportion   0.9453 0.97222 0.98664 0.99450 0.99931 1.00000
```

```r
round(prseatpos$rotation[, 1], 2)
```

```
## HtShoes      Ht  Seated     Arm   Thigh     Leg 
##   -0.65   -0.65   -0.27   -0.15   -0.17   -0.18
```

```r
prseatposc <- prcomp(cseatpos, scale = TRUE)
summary(prseatposc)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.2240 0.7082 0.58575 0.39551 0.22554 0.04149
## Proportion of Variance 0.8244 0.0836 0.05718 0.02607 0.00848 0.00029
## Cumulative Proportion  0.8244 0.9080 0.96516 0.99124 0.99971 1.00000
```

```r
round(prseatposc$rotation[, 1], 2)
```

```
## HtShoes      Ht  Seated     Arm   Thigh     Leg
##   -0.44   -0.44   -0.41   -0.37   -0.36   -0.42
```

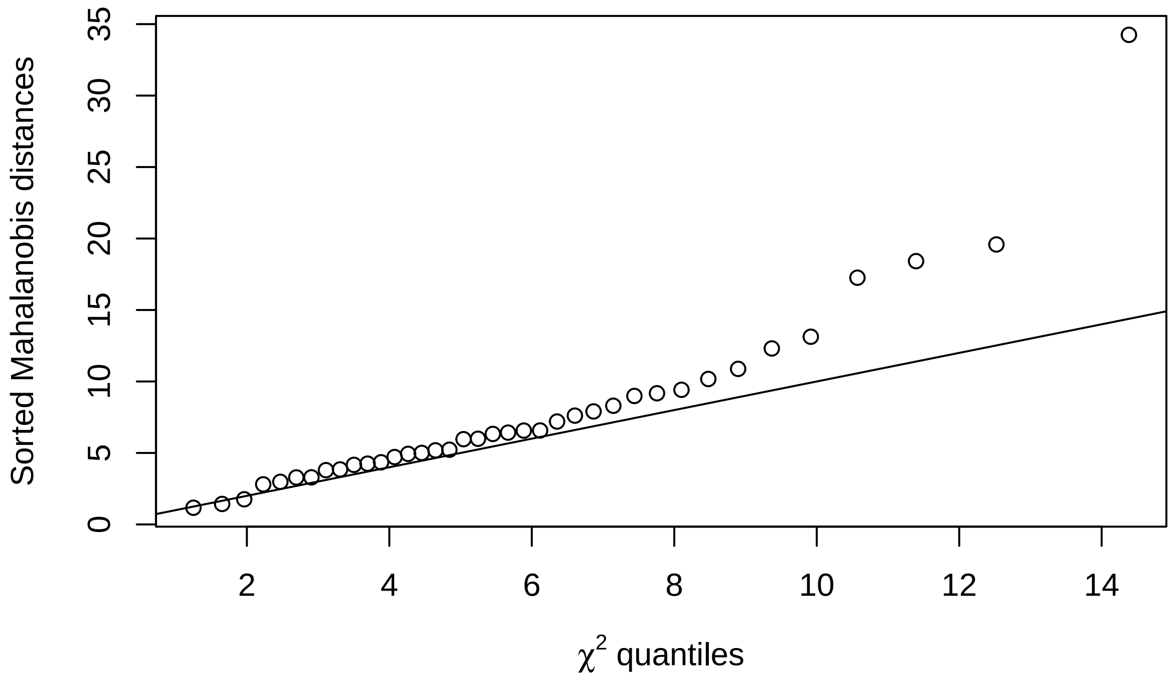
```r
require(MASS)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

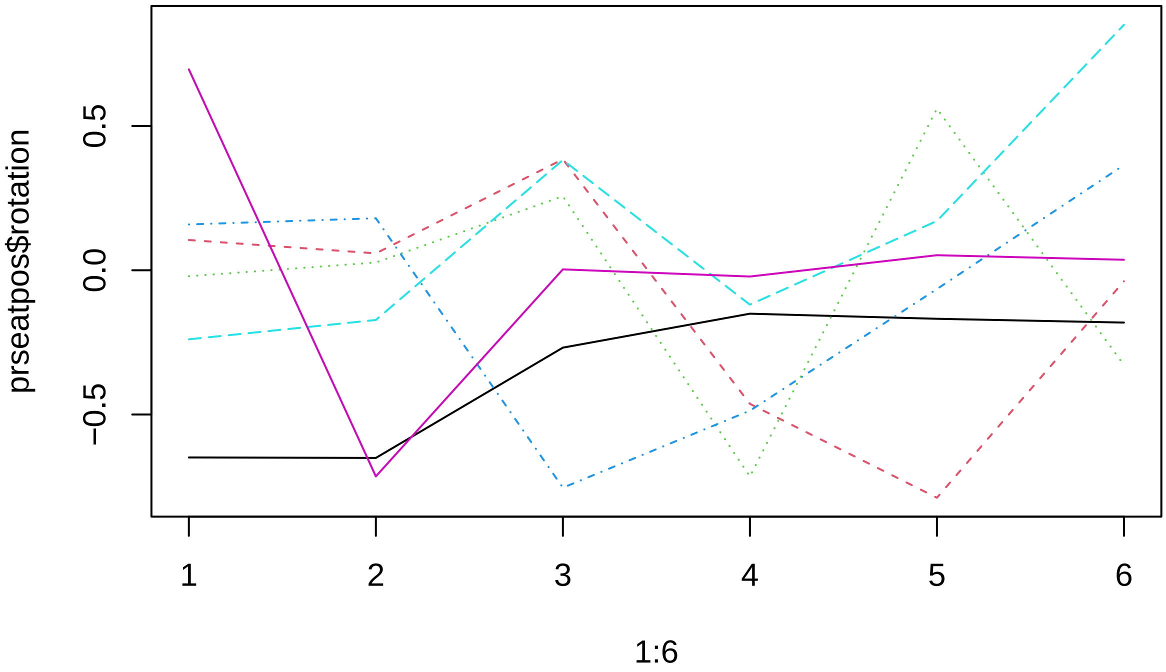
```r
robseatpos <- cov.rob(cseatpos)
md <- mahalanobis(cseatpos, center = robseatpos$center, cov = robseatpos$cov)
n <- nrow(cseatpos);p <- ncol(cseatpos)
plot(qchisq(1:n/(n + 1), p),
     sort(md),
     xlab = expression(paste(chi^2," quantiles")),
     ylab="Sorted Mahalanobis distances")
abline(0, 1)
```



```r
lmodpcr <- lm(seatpos$hipcenter ~ prseatpos$x[,1])
summary(lmodpcr)
```

```
##
## Call:
## lm(formula = seatpos$hipcenter ~ prseatpos$x[, 1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.009 -29.349   3.694  19.930  73.502
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -164.8849     5.9017 -27.939  < 2e-16 ***
## prseatpos$x[, 1]    2.7770     0.3486   7.966 1.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.38 on 36 degrees of freedom
## Multiple R-squared:  0.638,  Adjusted R-squared:  0.628
## F-statistic: 63.46 on 1 and 36 DF,  p-value: 1.853e-09
```

```r
matplot(1:6, prseatpos$rotation, type = "l")
```



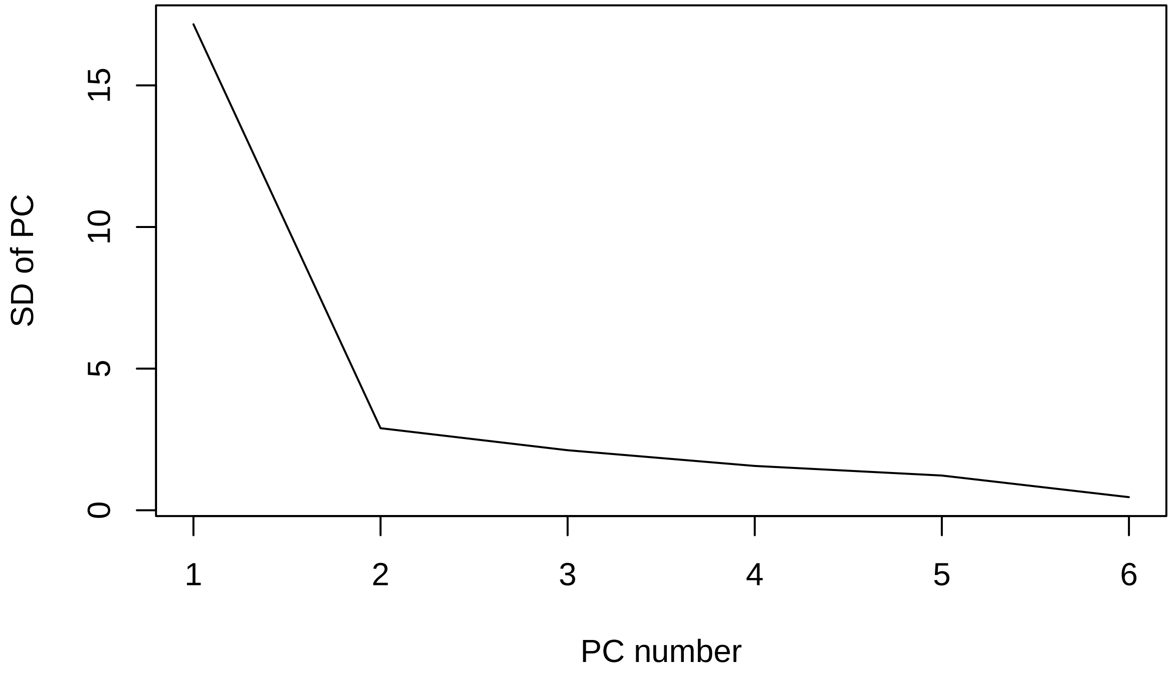
```
require(pls)
```

```
## Loading required package: pls

## Warning: package 'pls' was built under R version 4.1.3

##
## Attaching package: 'pls'

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

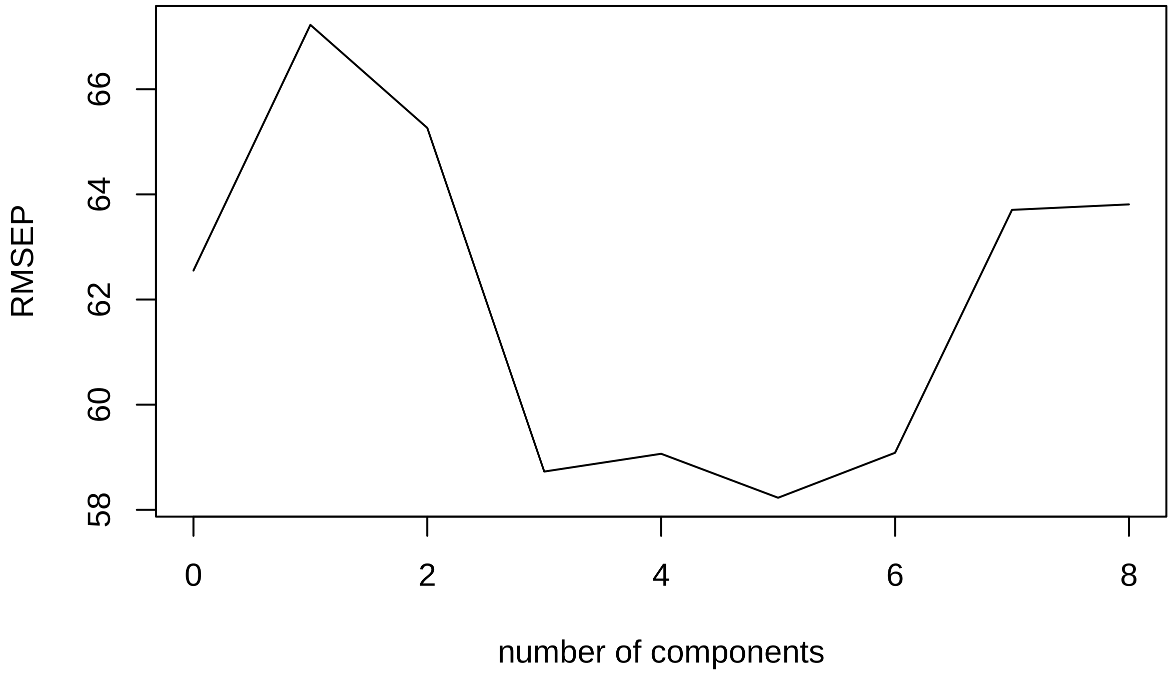
```
trainseat <- seatpos[1:30,]
testseat <- seatpos[31:38,]
pcrmod <- pcr(hipcenter ~ ., data=trainseat, ncomp = 8)

rmse <- function(x,y) sqrt(mean((x-y)^2))

rmse(predict(pcrmod), trainseat$hipcenter)
```

```
## [1] 29.07376
```

```
plot(prseatpos$sdev, type="l", ylab="SD of PC", xlab="PC number")
```



```
pcrmse <- RMSEP(pcrmod, newdata=testseat)
plot(pcrmse,main="")
```



Question Two

The dataset kanga contains data on the skulls of historical kangaroo specimens.

a

Compute a PCA on the 18 skull measurements. You will need to exclude observations with missing values. What percentage of variation is explained by the first principal component?

```
df <- kanga
df <-na.omit(df)
class.labels <- df$sex
class.labels.species <- df$species
df <-subset ( df,select = -c(species,sex))
pca.kanga <- prcomp(df)
summary(pca.kanga)
```

```
## Importance of components:
##                              PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation     288.0382 69.51124 30.74720 27.85580 21.73015 19.42356
## Proportion of Variance   0.9003  0.05243  0.01026  0.00842  0.00512  0.00409
## Cumulative Proportion    0.9003  0.95269  0.96295  0.97136  0.97649  0.98058
##                              PC7     PC8      PC9     PC10     PC11     PC12
## Standard deviation     17.28247 16.6247 14.52310 13.98826 12.35253 12.07402
## Proportion of Variance  0.00324  0.0030  0.00229  0.00212  0.00166  0.00158
## Cumulative Proportion   0.98382  0.9868  0.98911  0.99123  0.99289  0.99447
##                            PC13     PC14     PC15     PC16     PC17     PC18
## Standard deviation     11.94245 10.82939 10.0735 8.46081 7.16825 5.01246
## Proportion of Variance  0.00155  0.00127  0.0011 0.00078 0.00056 0.00027
## Cumulative Proportion   0.99602  0.99729  0.9984 0.99917 0.99973 1.00000
```

b

Provide the loadings for the first principal component. What variables are prominent?

```r
library(pander)
```

```
## Warning: package 'pander' was built under R version 4.1.3
```

```r
pander( data.frame(first.pc.loadings =round(pca.kanga$rotation[,1], 3)), caption ="First Principal Compo
```

Table 1: First Principal Component

	first.pc.loadings
basilar.length	0.484
occipitonasal.length	0.456
palate.length	0.366
palate.width	0.084
nasal.length	0.248
nasal.width	0.075
squamosal.depth	0.064
lacrymal.width	0.119
zygomatic.width	0.207
orbital.width	0.014
.rostral.width	0.106
occipital.depth	0.178
crest.width	-0.082
foramina.length	0.01
mandible.length	0.436
mandible.width	0.03
mandible.depth	0.058
ramus.height	0.209

c

Repeat the PCA but with the variables all scaled to the same standard deviation. How do the percentage of variation explained and the first principal component differ from those found in the previous PCA?

```r
pca.kanga.scaled <- prcomp(df,scale. = TRUE)
summary(pca.kanga.scaled)
```

```
## Importance of components:
##                             PC1     PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     3.5321 1.30672 1.1006 0.8443 0.6463 0.56426 0.51064
## Proportion of Variance 0.6931 0.09486 0.0673 0.0396 0.0232 0.01769 0.01449
## Cumulative Proportion  0.6931 0.78796 0.8553 0.8949 0.9181 0.93575 0.95024
##                            PC8     PC9   PC10    PC11   PC12    PC13    PC14
## Standard deviation     0.45185 0.43863 0.3723 0.30491 0.2815 0.24345 0.22317
## Proportion of Variance 0.01134 0.01069 0.0077 0.00517 0.0044 0.00329 0.00277
## Cumulative Proportion  0.96158 0.97227 0.9800 0.98514 0.9895 0.99283 0.99560
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.18583 0.15031 0.11849 0.08949
## Proportion of Variance 0.00192 0.00126 0.00078 0.00044
## Cumulative Proportion  0.99752 0.99878 0.99956 1.00000
```

d

Give an interpretation of the second principal component.

```
pander( data.frame(first.pc.loadings =round(pca.kanga$rotation[,2], 3)), caption ="Second Principal Comp
```

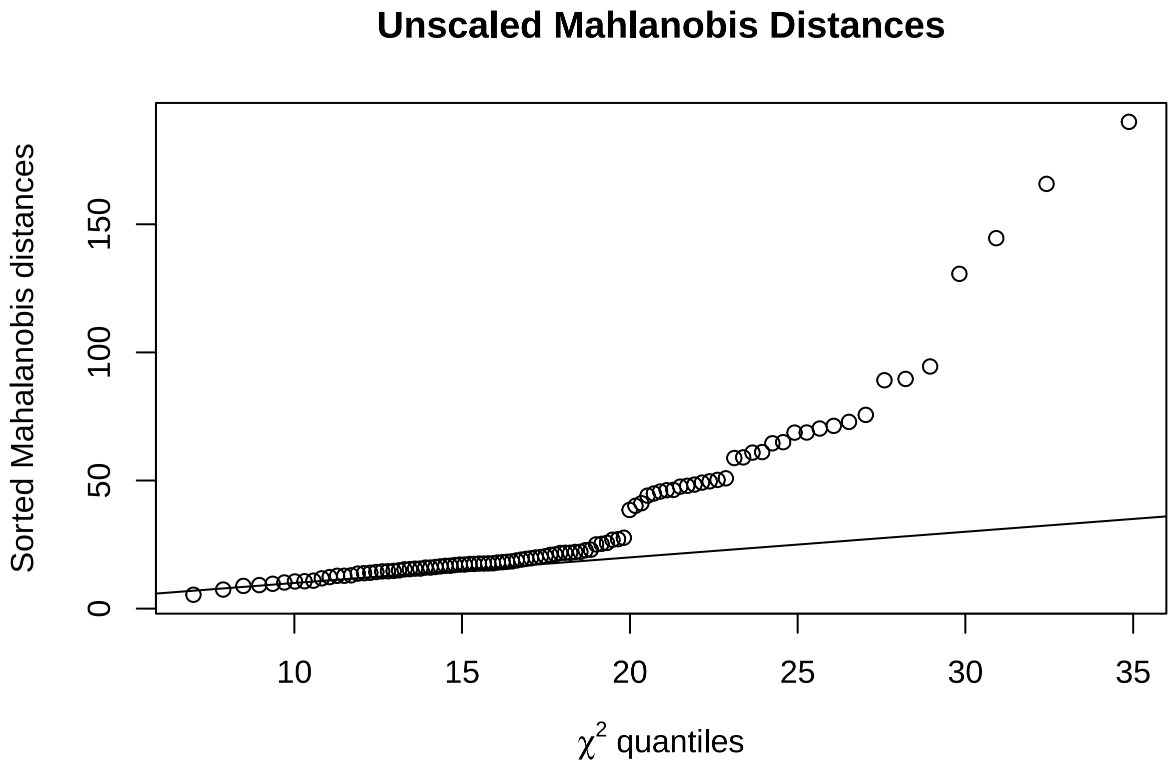
Table 2: Second Principal Component

	first.pc.loadings
basilar.length	-0.138
occipitonasal.length	0.414
palate.length	-0.002
palate.width	-0.023
nasal.length	0.584
nasal.width	0.127
squamosal.depth	-0.105
lacrymal.width	-0.04
zygomatic.width	-0.41
orbital.width	0.001
.rostral.width	-0.063
occipital.depth	-0.079
crest.width	-0.245
foramina.length	0.061
mandible.length	-0.212
mandible.width	-0.092
mandible.depth	-0.106
ramus.height	-0.365

e

Compute the Mahalanobis distances and plot appropriately to check for outliers.

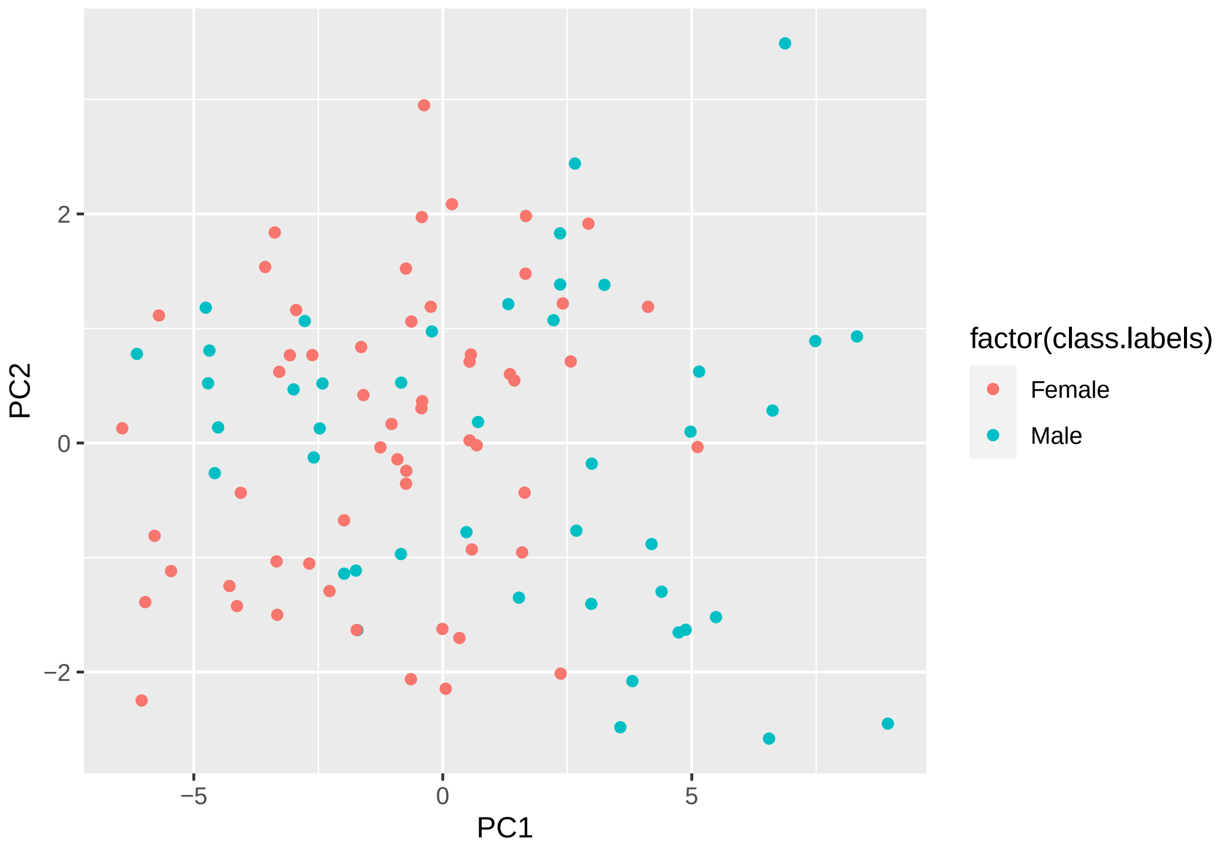
```
require(MASS)
rob.kanga <- cov.rob(df)
mahalanobis.distances <- mahalanobis(df, center=rob.kanga$center, cov=rob.kanga$cov)
n <- nrow(df)
p <- ncol(df)
plot(qchisq(1:n/(n+1),p), sort(mahalanobis.distances), xlab=expression(paste(chi^2," quantiles")), ylab
abline(0,1)
```



f

Make a scatterplot of the first and second principal components using a different plotting symbol depending on the sex of the specimen. Do you think these two components would be effective in determining the sex of a skull?

```
scores <- data.frame(class.labels, pca.kanga.scaled$x[,1:2])
qplot(x=PC1, y=PC2, data=scores, colour=factor(class.labels))
```


2
0
−2
PC2
−5
0
5
PC1
factor(class.labels)
● Female
● Male