# HW03

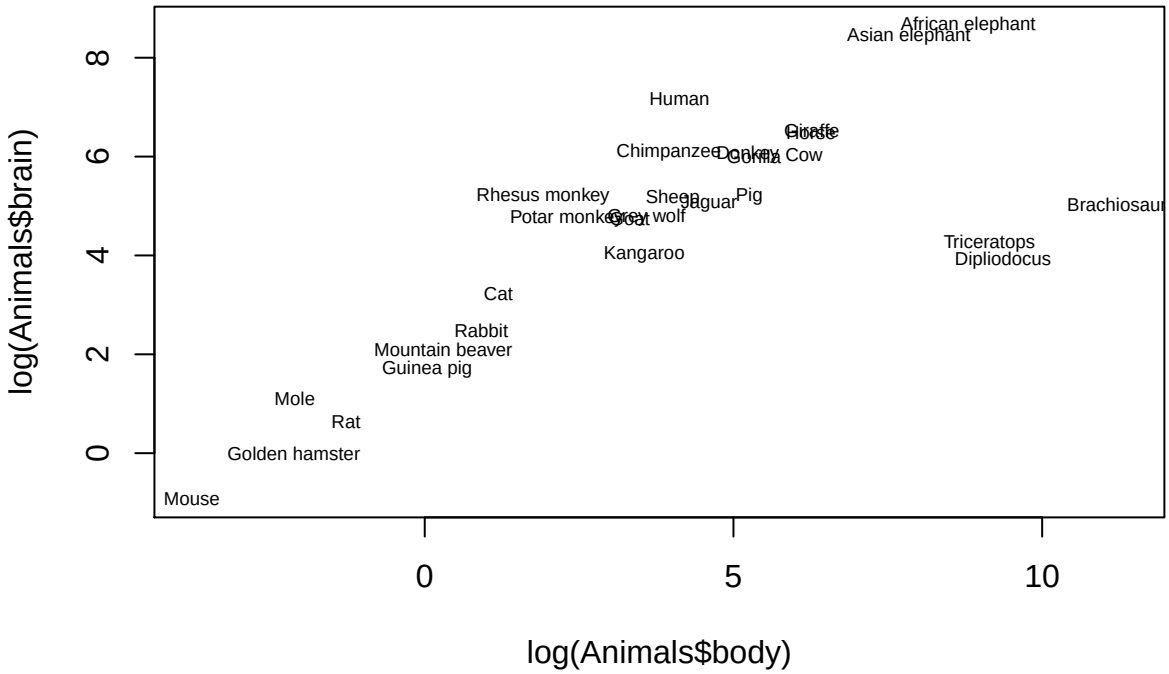
*Kaiyuan Hu*

*2017/4/18*

## Question 2

### a)

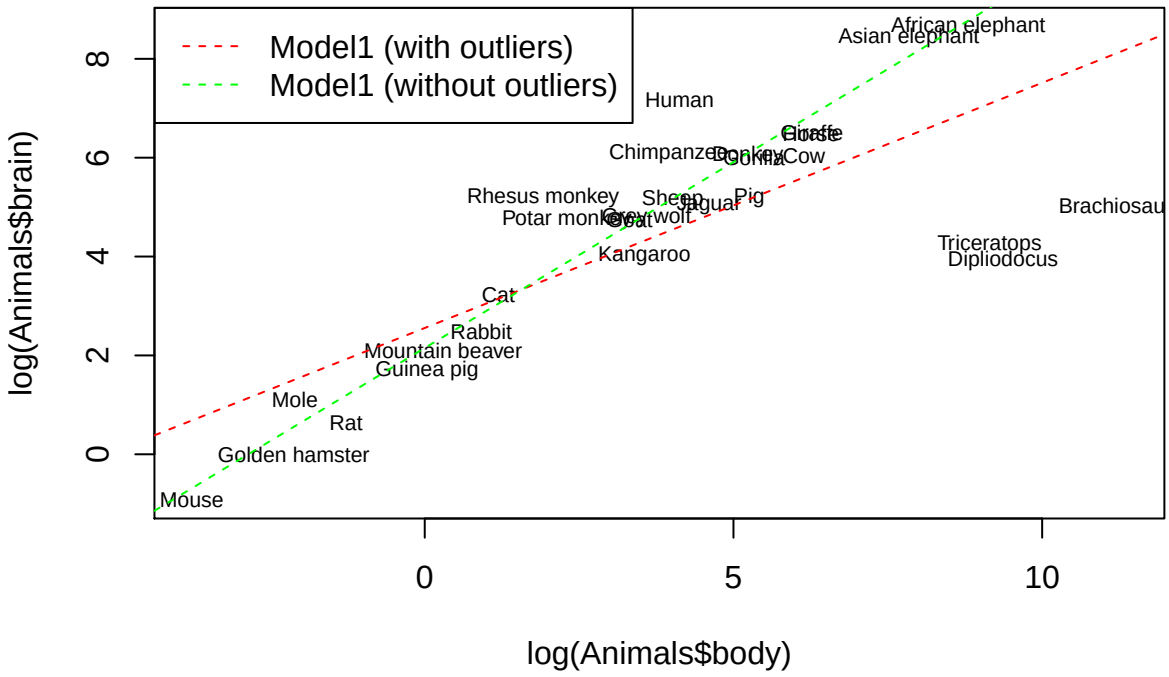
```
library(MASS)
data(Animals)
plot(log(Animals$body),log(Animals$brain),type = 'n')
text(log(Animals$body),log(Animals$brain),row.names(Animals),cex=.6)
```




There seems to be a linear trend on the double-log scale. The three most obvious outliers are *Brachiosaurus*, *Triceratops*, and *Dipliodocus*, since they have smaller brain sizes corresponding to their large body sizes.

### b)

```
model = lm(log(brain)~log(body), data = Animals)
model_wo=lm(log(brain)~log(body), data=Animals[-c(6,16,26),])
plot(log(Animals$body),log(Animals$brain),type = 'n')
text(log(Animals$body),log(Animals$brain),row.names(Animals),cex=.7)
abline(model,col='red', lty=2)
abline(model_wo,col='green', lty=2)
legend('topleft',c('Model1 (with outliers)','Model1 (without outliers)'),col=c('red','green'),lty=2)
```




```
# The coefficients of Model1 with outliers
modelcoef <- c(exp(model$coefficients[1]),model$coefficients[2])
names(modelcoef) <- c('K','r')
modelcoef
```

```
##          K          r
## 12.8699886  0.4959947
```

```
#The coefficients of Model1 without outliers
modelwcoef <- c(exp(model_wo$coefficients[1]),model_wo$coefficients[2])
names(modelwcoef) <- c('K','r')
modelwcoef
```

```
##         K         r
## 8.5883969 0.7522607
```

**c)**

$$EQ = \frac{Brain}{\hat{Brain}} = \frac{\exp(\log Brain)}{\exp(\log \hat{Brain})} = \exp(\log Brain - \log \hat{Brain}) = \exp(e_i)$$

**d)**

```
result = sort(Animals$brain/exp(predict(model_wo,Animals)),T)
result
```

```
##            Human    Rhesus monkey       Chimpanzee     Potar monkey
##      6.891535162      4.928073515      2.616103835      2.368781609
##             Mole   Asian elephant              Cat             Goat
##      1.700218567      1.468612121      1.214143989      1.101886824
##            Sheep            Donkey       Grey wolf African elephant
##      0.993031125      0.953037679      0.932673076      0.884954831
```

```
##          Gorilla            Mouse  Mountain beaver          Giraffe
##      0.855855460      0.795345213      0.752537143      0.707698219
##           Rabbit            Horse       Guinea pig              Rat
##      0.707161536      0.689539124      0.621780464      0.576398430
##   Golden hamster           Jaguar              Cow        Kangaroo
##      0.573830551      0.572092326      0.485074230      0.449504656
##              Pig      Triceratops      Dipliodocus    Brachiosaurus
##      0.401534245      0.008362920      0.005066657      0.003461068
```

EQ of the outlier rank lower.

---

# Question 3

## a)

```
data(biopsy)
biopsy <- na.omit(biopsy)
X <-  as.matrix(biopsy[,2:10])
colnames(X) <- c('F1','F2','F3','F4','F5','F6','F7','F8','F9')
head(X)
```

```
##   F1 F2 F3 F4 F5 F6 F7 F8 F9
## 1  5  1  1  1  2  1  3  1  1
## 2  5  4  4  5  7 10  3  2  1
## 3  3  1  1  1  2  2  3  1  1
## 4  6  8  8  1  3  4  3  7  1
## 5  4  1  1  3  2  1  3  1  1
## 6  8 10 10  8  7 10  9  7  1
```

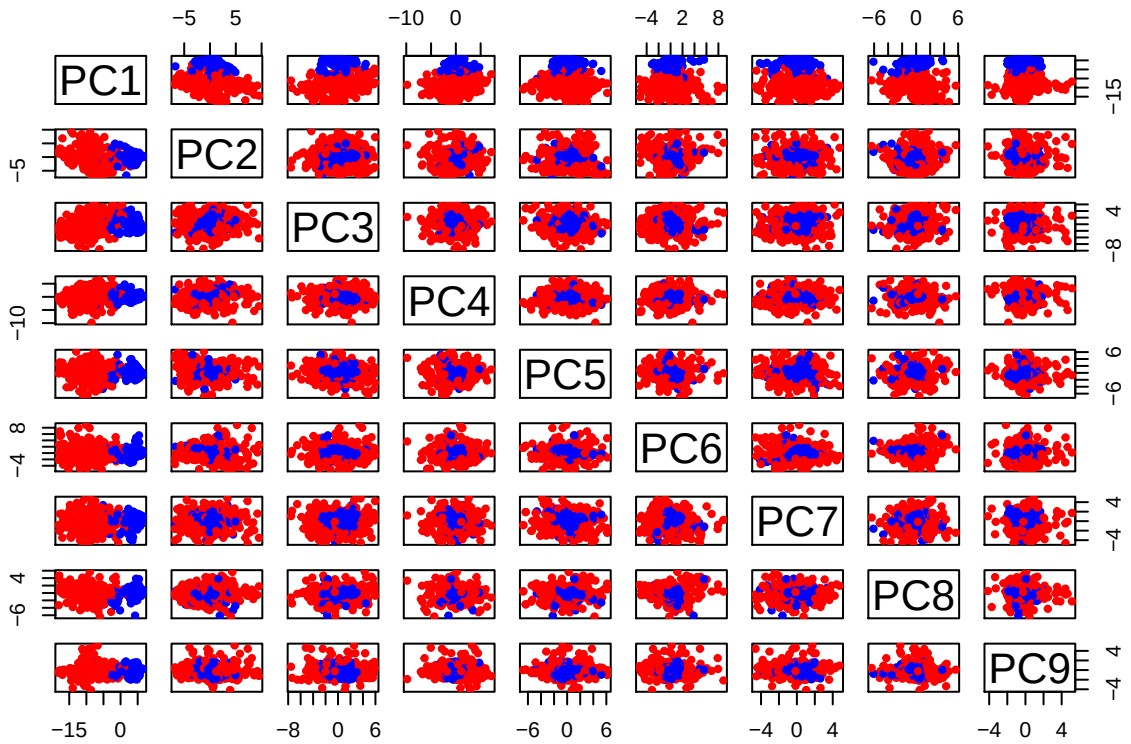
## b)

```
P <- prcomp(X,center=T,scale. = F)
Xbar <- X-P$center
colnames(Xbar) <- c('F1_bar','F2_bar','F3_bar','F4_bar','F5_bar','F6_bar','F7_bar','F8_bar','F9_bar')
head(Xbar)
```

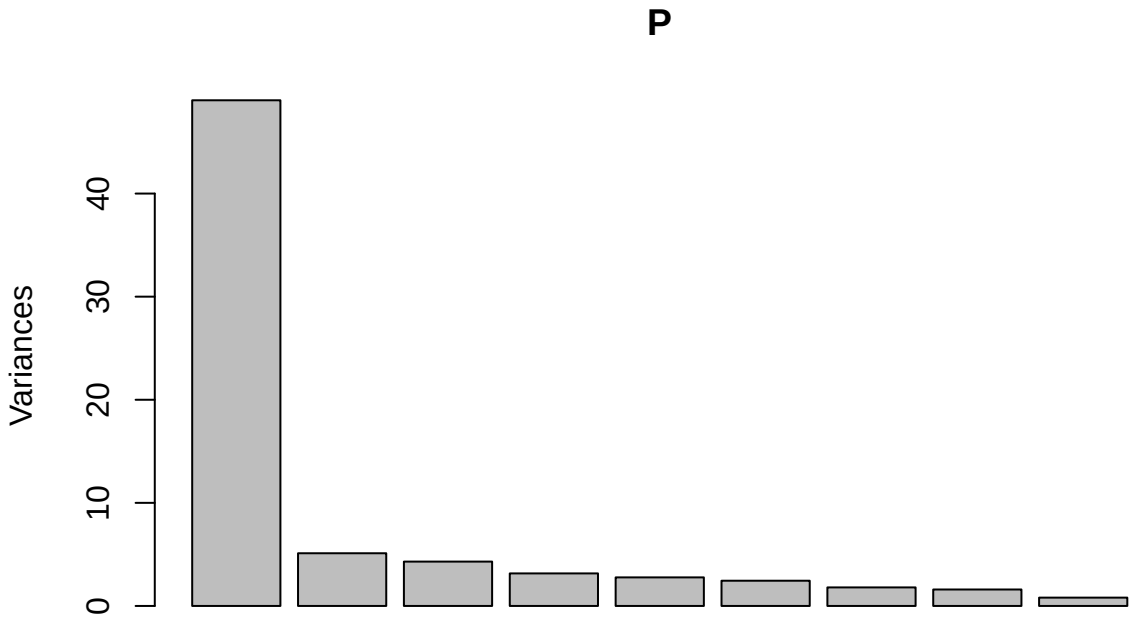
```
##       F1_bar     F2_bar    F3_bar     F4_bar     F5_bar     F6_bar
## 1  0.5578331 -0.6032211 -1.869693 -2.4450952 -1.5446559 -2.2342606
## 2  1.8491947 -0.4421669  2.396779  2.1303075  3.5549048  6.4553441
## 3 -0.2152269 -2.1508053 -3.442167 -0.6032211 -0.8696925 -1.4450952
## 4  3.1698389  4.7847731  4.849195 -3.4421669  1.3967789  1.1303075
## 5  0.7657394 -1.8301611 -2.215227 -0.1508053 -2.4421669 -0.6032211
## 6  4.4553441  6.7657394  7.169839  4.7847731  3.8491947  5.5578331
##       F7_bar     F8_bar    F9_bar
## 1  0.1698389 -2.2152269 -2.150805
## 2 -0.2342606 -0.8301611 -2.215227
## 3 -0.5446559 -2.2342606 -1.830161
## 4 -0.4450952  3.4553441 -2.234261
## 5  0.1303075 -2.4450952 -2.544656
## 6  7.3967789  4.1303075 -2.445095
```

c)

```
pairs(P$x,col=c('blue','red')[biopsy$class],pch=20)
```




```
screeplot(P)
```




The first component explains the majority of variance.

## d)

```
# I The 1st principal component is doing a weighted sum over all features.
P$rotation[,1]
```

```
##         F1         F2         F3         F4         F5         F6
## -0.2967358 -0.4039707 -0.3927586 -0.3312021 -0.2497398 -0.4426135
##         F7         F8         F9
## -0.2920783 -0.3545360 -0.1245763
```

```
# II F6 has the largest loading for the 2nd principal component.
names(which.max(abs(P$rotation[,2])))
```

```
## [1] "F6"
```

```
# III F1 has the largest loading for the 3rd principal component.
names(which.max(abs(P$rotation[,3])))
```

```
## [1] "F1"
```

## e)

```
B <-  matrix(rep(0,81),9,9)
for (i in 1:9) B[i,i]=1
B[,9] <- rep(1/9,9)
W <-  X%*%B
head(W)
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]     [,9]
## 1    5    1    1    1    2    1    3    1 1.777778
## 2    5    4    4    5    7   10    3    2 4.555556
## 3    3    1    1    1    2    2    3    1 1.666667
## 4    6    8    8    1    3    4    3    7 4.555556
## 5    4    1    1    3    2    1    3    1 1.888889
## 6    8   10   10    8    7   10    9    7 7.777778
```

Columns 1,6,9.

## f)

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-5
```

```
set.seed(3)
lasso <- cv.glmnet(W,biopsy$class,family='binomial',alpha=1)
coef(lasso,s=lasso$lambda.1se)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -6.0048115
## V1           0.1071590
```

```
## V2           .
## V3           .
## V4           .
## V5           .
## V6           0.1210325
## V7           .
## V8           .
## V9           1.3164962
```

*lambda.1se* is the value of $\lambda$ in the list has error within 1 standard error of the best model. $\alpha = 1$ is only LASSO constraint. Column 1,6,9 are important features. This result is same as principal componets.

## g)

```
library(AUC)
```

```
## AUC 0.3.0
```

```
## Type AUCNews() to see the change log and ?AUC to get an overview.
```

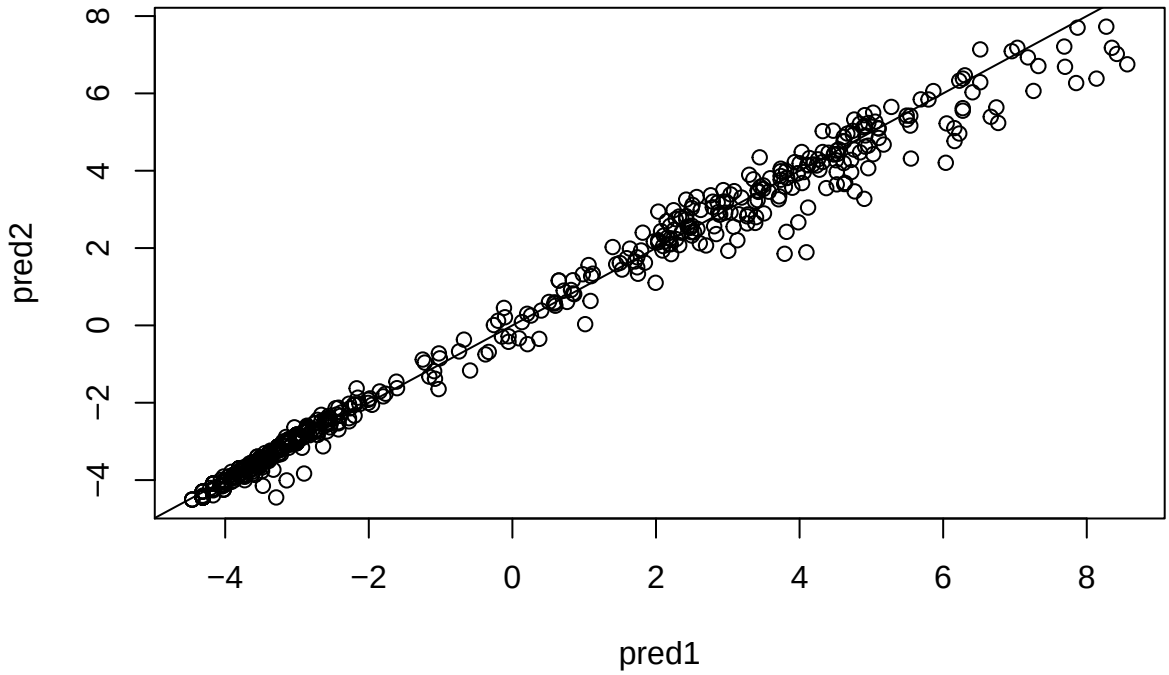
```
##
## Attaching package: 'AUC'
```

```
## The following object is masked from 'package:glmnet':
##
##     auc
```

```
set.seed(3)
lasso2 <- cv.glmnet(X,biopsy$class,family='binomial',alpha=1)
coef(lasso2,s=lasso2$lambda.1se)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept) -5.89369465
## F1           0.30068991
## F2           0.12537234
## F3           0.19395365
## F4           0.10025259
## F5           0.05512987
## F6           0.28648144
## F7           0.20808581
## F8           0.11950373
## F9           .
```

```
pred1 <- predict(lasso,W,s=lasso$lambda.1se)
pred2 <- predict(lasso2,X,s=lasso2$lambda.1se)
plot(pred1,pred2)
abline(0,1)
```




```
auc(roc(pred1,biopsy$class))
```

```
## [1] 0.9959808
```

```
auc(roc(pred2,biopsy$class))
```

```
## [1] 0.9959667
```

Those two model are really similar.

## h)

### I)

$\bar{S}$ is the most important feature used in this model.

### II)

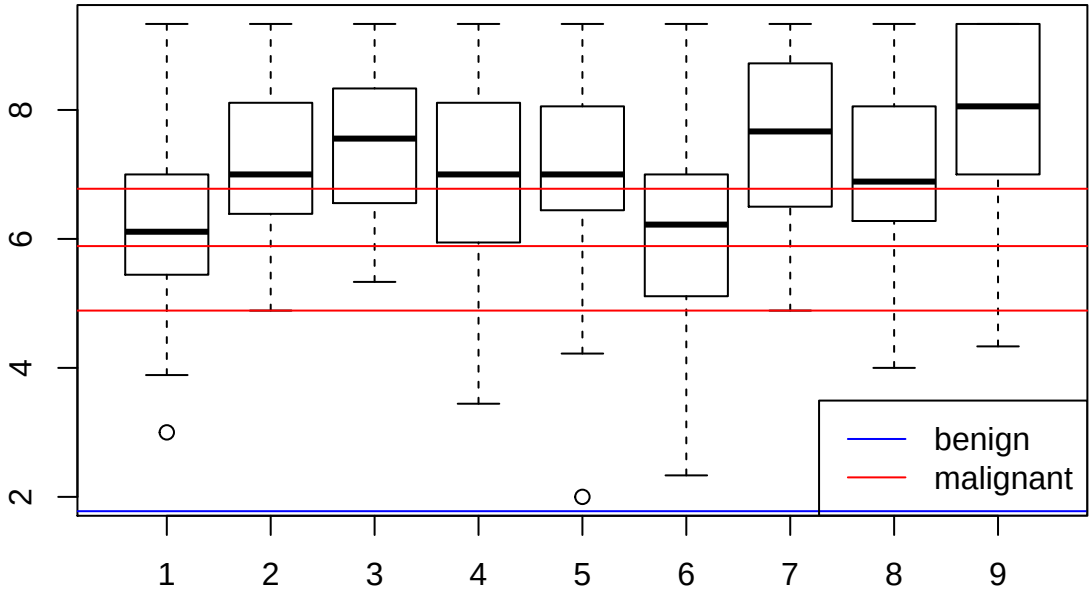
```
obs <- biopsy
obs <- obs[rowSums(obs[,c(2:10)]==10)>0,]
summary(obs$class)
```

```
##    benign malignant
##         5       204
```

### III)

```
S_bar <- W[,9]
l <- list()
for (i in 1:9) l[[i]] <- S_bar[biopsy[,i+1]==10]
boxplot(l)
```

```
abline(h=quantile(S_bar[biopsy$class=='benign'],c(1:3)/4),col='blue')
abline(h=quantile(S_bar[biopsy$class=='malignant'],c(1:3)/4),col='red')
legend('bottomright',c('benign','malignant'),col=c('blue','red'),lty = 1)
```




As we can see, *benign* and *malignant* are important and highly correlated to our prediction. So this is the reason why the model in part (f) select those two features.

---