PSTAT126 HW1

zhongyun zhang

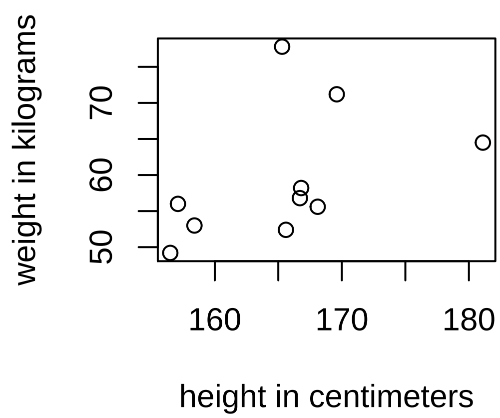
2020/4/11

Problem 1

#a) Predictor is height, and response is weight.

#b) Simple linear regression model does not make sense in this data. Since they are too scattered away from each other that we cannot see the pattern for the data and they seems not showing a linear pattern. Moreover, this is a such small sample size(only 10 samples) that we are not able to see a clear pattern.

```
setwd("~/Desktop/2020 Spring/PSTAT126/hw1/hw1 datasets")
Htwt <- read.table("Htwt.csv", header=TRUE,sep = ",")
#data(Htwt)
x=Htwt$ht
y=Htwt$wt
#View(Htwt)
n = length(x)
plot(x,y,xlab="height in centimeters",ylab="weight in kilograms")
```



#c)

```
xbar = mean(x)
xbar
```

```
## [1] 165.52
```

```
ybar = mean(y)
ybar
```

```
## [1] 59.47
```

```
Sxx = sum((x - xbar)^2)
Sxx
```

```
## [1] 472.076
```

```
Syy = sum((y - ybar)^2)
Syy
```

```
## [1] 731.961
```

```r
Sxy = sum((x - xbar)*(y - ybar))
Sxy
```

```
## [1] 274.786
```

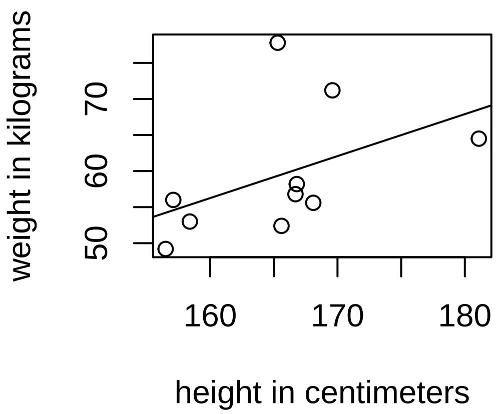
```r
#correlation coefficient between x and y
r = Sxy/sqrt(Sxx*Syy)
#slope
b1 = r*sqrt(Syy/Sxx)
b1
```

```
## [1] 0.58208
```

```r
#y-intercept
b0 = ybar - b1*xbar
b0
```

```
## [1] -36.87588
```

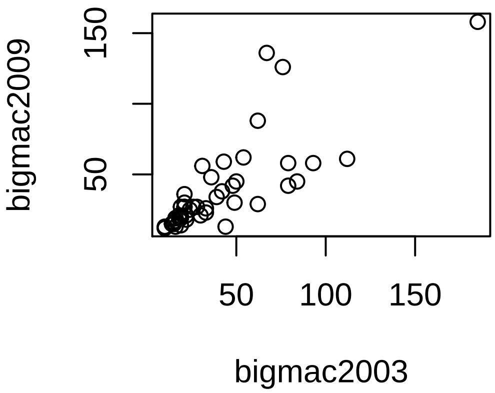
```r
#add the least squares fit to the scatterplot
plot(x, y, xlab = "height in centimeters", ylab = "weight in kilograms")
#intercept and slope
abline(b0, b1)
```



Problem 2

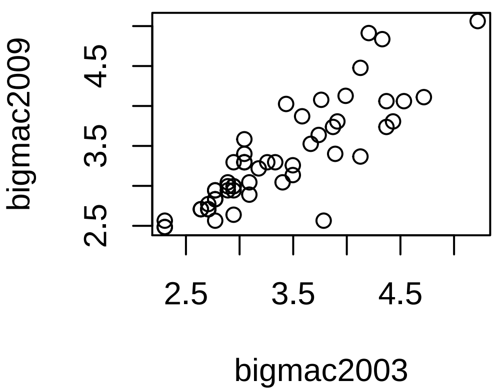
#a) Fitting simple linear regression to the figure in this problem is not likely to be appropriate since there are not enough correlation between these x and Y. Moreover, data are lied in a small area with some outlierswhich does not follow the pattern of simple linear regression model.

```r
setwd("~/Desktop/2020 Spring/PSTAT126/hw1/hw1 datasets")
UBSprices <- read.table("UBSprices.csv", header=TRUE,sep = ",")
#data(UBSprices)
#View(UBSprices)
x=UBSprices$bigmac2003
y=UBSprices$bigmac2009
plot(x,y,xlab="bigmac2003",ylab="bigmac2009")
```



#b) This graph is more sensibly summarized with a linear regression, since the appearance of plots look like linear, which data are gethering around the regression line. Moreover, the errors seems to be normally distributed.

```
plot(log(x),log(y),xlab="bigmac2003",ylab="bigmac2009")
```



#c)

```
xbar = mean(log(x))
ybar = mean(log(y))
Sxx = sum((log(x) - xbar)^2)
Syy = sum((log(y) - ybar)^2)
Sxy = sum((log(x) - xbar)*(log(y) - ybar))
r = Sxy/sqrt(Sxx*Syy)
b1 = r*sqrt(Syy/Sxx)
b1
```

```
## [1] 0.8029268
```

```
b0 = ybar - b1*xbar
b0
```

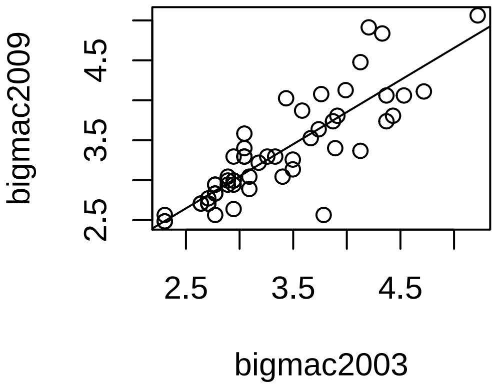
```
## [1] 0.6403147
```

```
yhat2 = b0 + b1*x
#total sum of squars
ssto = sum((log(y)- ybar)^2)
#error sum of squares
sse = sum((log(y) - yhat2)^2)
#regression sum of squares
ssr = sum((yhat2 - ybar)^2)
r2 = ssr/ssto
```

```r
r2
```

```
## [1] 3383.337
```

```r
plot(log(x),log(y),xlab="bigmac2003",ylab="bigmac2009")
abline(b0, b1)
```



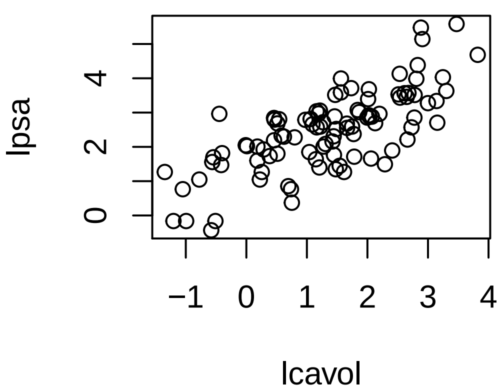
PROBLEM 3

#a)

```r
setwd("~/Desktop/2020 Spring/PSTAT126/hw1/hw1 datasets")
prostate <- read.table("prostate.csv", header=TRUE,sep = ",")
#data(prostate)
x=prostate$lcavol
y=prostate$lpsa
fit=lm(y~x)
coef(fit)
```

```
## (Intercept)           x
##   1.5072979   0.7193201
```

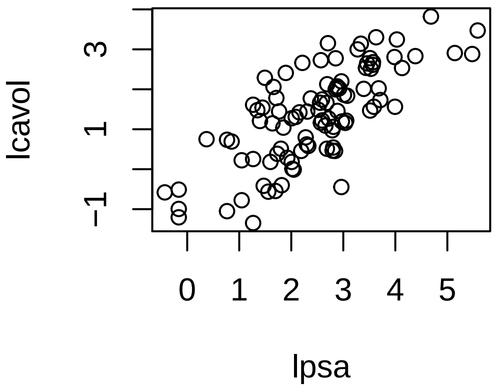
```r
plot(x,y,xlab="lcavol",ylab="lpsa")
```




```r
fit=lm(x~y)
coef(fit)
```

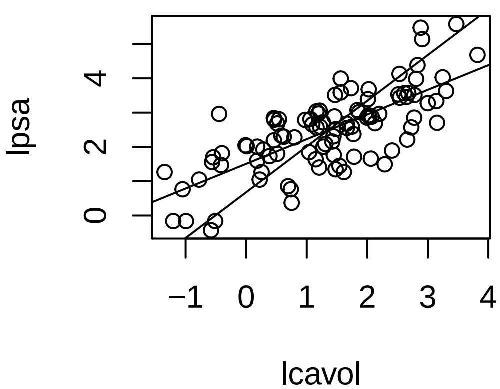
```
## (Intercept)           y
##  -0.5085802   0.7499191
```

```r
plot(y,x,ylab="lcavol",xlab="lpsa")
```



#b) Two lines intersect at the point of mean of the mean of x and y(lcavol and lpsa). Since The least-squares regression line always goes through point xbar,ybar.

```
plot(x,y,xlab="lcavol",ylab="lpsa")
b0=1.5072979
b1=0.7193201
abline(b0, b1)
b2=0.5085802/0.7499191
b3=1/0.7499191
abline(b2,b3)
```



PROBLEM 4

#a)

```
setwd("~/Desktop/2020 Spring/PSTAT126/hw1/hw1 datasets")
Heights <- read.table("Heights.csv", header=TRUE,sep = ",")
#data(Heights)
x=Heights$mheight
y=Heights$dheight
plot(x,y,xlab='mheight',ylab='dheight')
coef(fit)
```

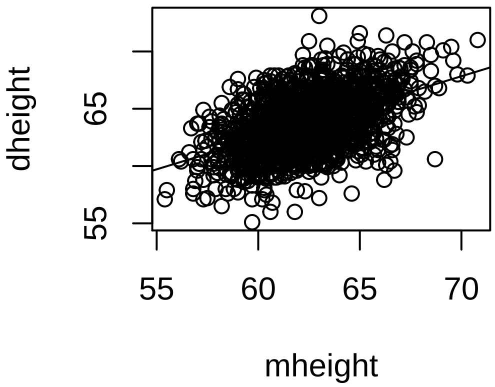
```
## (Intercept)           y
##  -0.5085802   0.7499191
```

```
xbar=mean(x)
ybar=mean(y)
Sxx=sum((x-xbar)^2)
Syy = sum((y - ybar)^2)
Sxy = sum((x - xbar)*(y - ybar))
```

```r
#correlation coefficient between x and y
r = Sxy/sqrt(Sxx*Syy)
#slope
b1 = r*sqrt(Syy/Sxx)
#y-intercept
b0 = ybar - b1*xbar
fit=lm(y~x)
coef(fit)
```

```
## (Intercept)           x
##   29.917437    0.541747
```

```r
#intercept and slope
abline(b0, b1)
```



#b) the rxy is 0.4907, which means that there is moderatly positive linear relationship between daughters' height and mothers' height.

```r
rxy = Sxy/sqrt(Sxx*Syy)
rxy
```

```
## [1] 0.4907094
```