PSTAT 126 Homework 1

Shaiyon Hariri

4/9/2020

1. In the Htwt data in the alr4 package, ht = height in centimeters and wt = weight in kilograms for a sample of n = 10 18-year-old girls. Interest is in predicting weight from height.

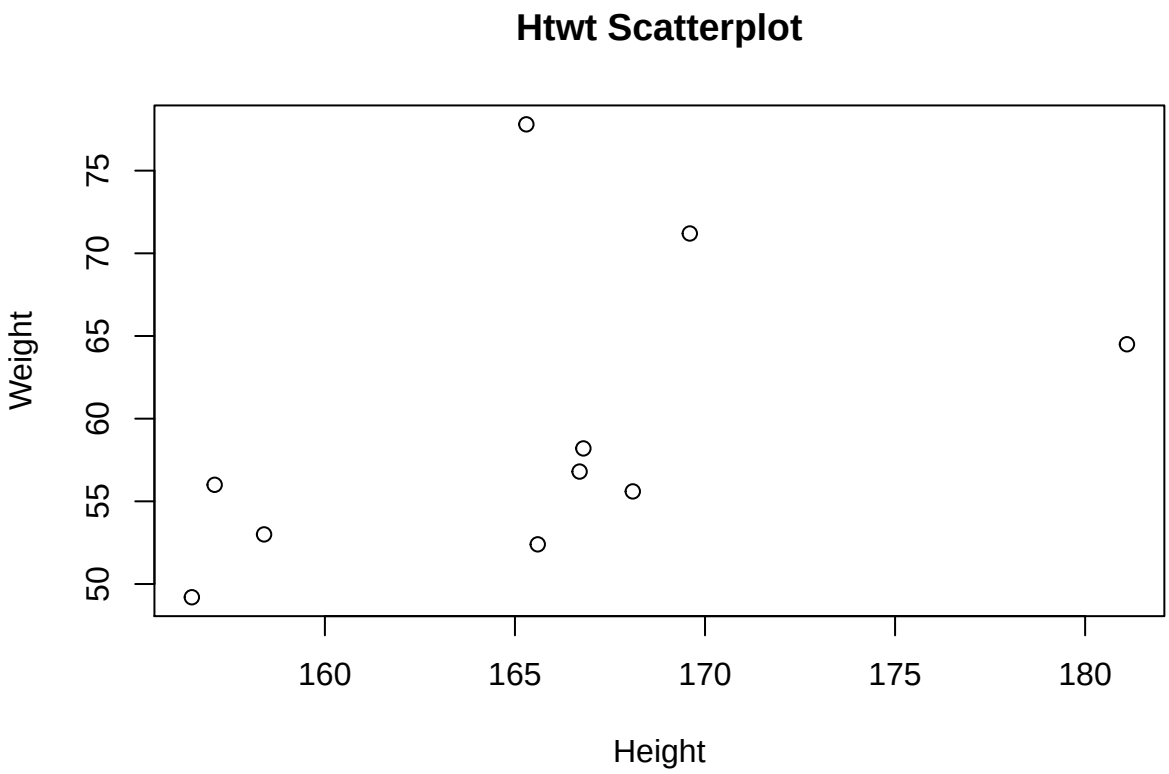
(a) Identify the predictor and response.

Predictor: height Response: weight

(b) Draw a scatterplot of wt on the vertical axis versus ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

```
htwt <- read.csv("Htwt.csv")

plot(x=htwt$ht, y=htwt$wt, xlab="Height", ylab="Weight", main="Htwt Scatterplot")
```



There are some outliers, but the data seems to follow a linear trend of weight increasing as height increases. So yes.

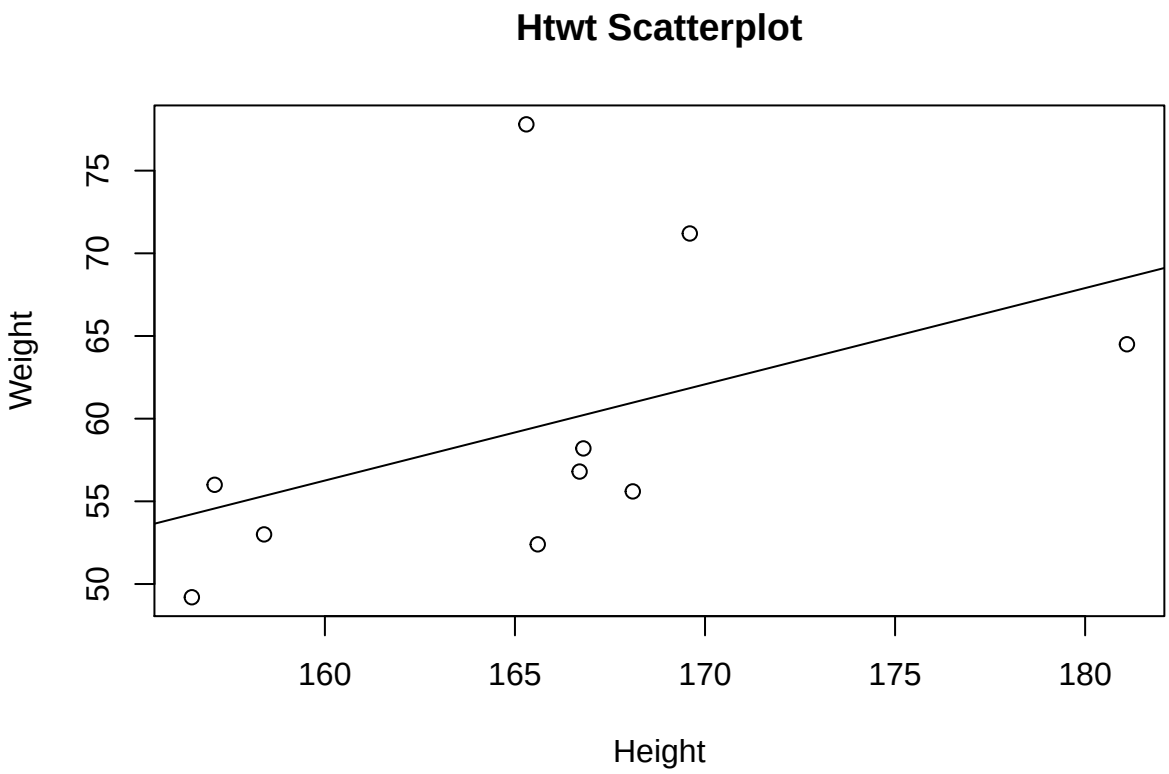
(c) Show that xbar = 165.52, ybar = 59.47, Sxx = 472.08, Syy = 731.96 and Sxy = 274.79. Compute estimates of the slope and the intercept for the regression of Y on x. Draw the fitted line on your scatterplot.

```
x <- htwt$ht
y <- htwt$wt
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum((x - xbar)^2)
Syy <- sum((y - ybar)^2)
Sxy <- sum((x - xbar)*(y - ybar))
# Estimates of the slope and intercept
b1 <- Sxy/Sxx
b0 <- ybar - b1*xbar

cat("xbar:", xbar, " ybar:", ybar, " Sxx:", Sxx, " Syy:", Syy, " Sxy:", Sxy)
```

```
## xbar: 165.52  ybar: 59.47  Sxx: 472.076  Syy: 731.961  Sxy: 274.786
```

```
plot(x=htwt$ht, y=htwt$wt, xlab="Height", ylab="Weight", main="Htwt Scatterplot")
abline(b0, b1)
```

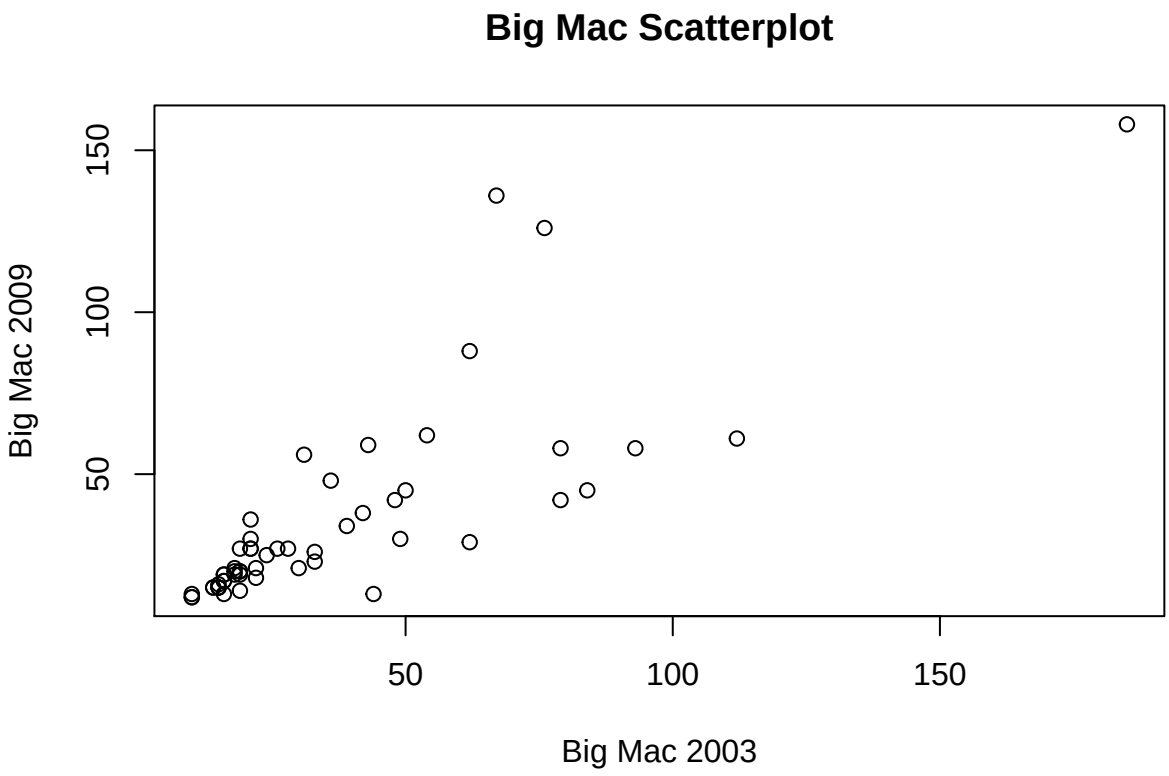


2. This problem uses the UBSprices data set in the alr4 package.

(a) Draw the plot of Y = bigmac2009 versus x = bigmac2003, the price of a Big Mac hamburger in 2009 and 2003. Give a reason why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

```
UBSprices <- read.csv("UBSprices.csv")

plot(x=UBSprices$bigmac2003, y=UBSprices$bigmac2009, xlab="Big Mac 2003",
     ylab="Big Mac 2009", main="Big Mac Scatterplot")
```

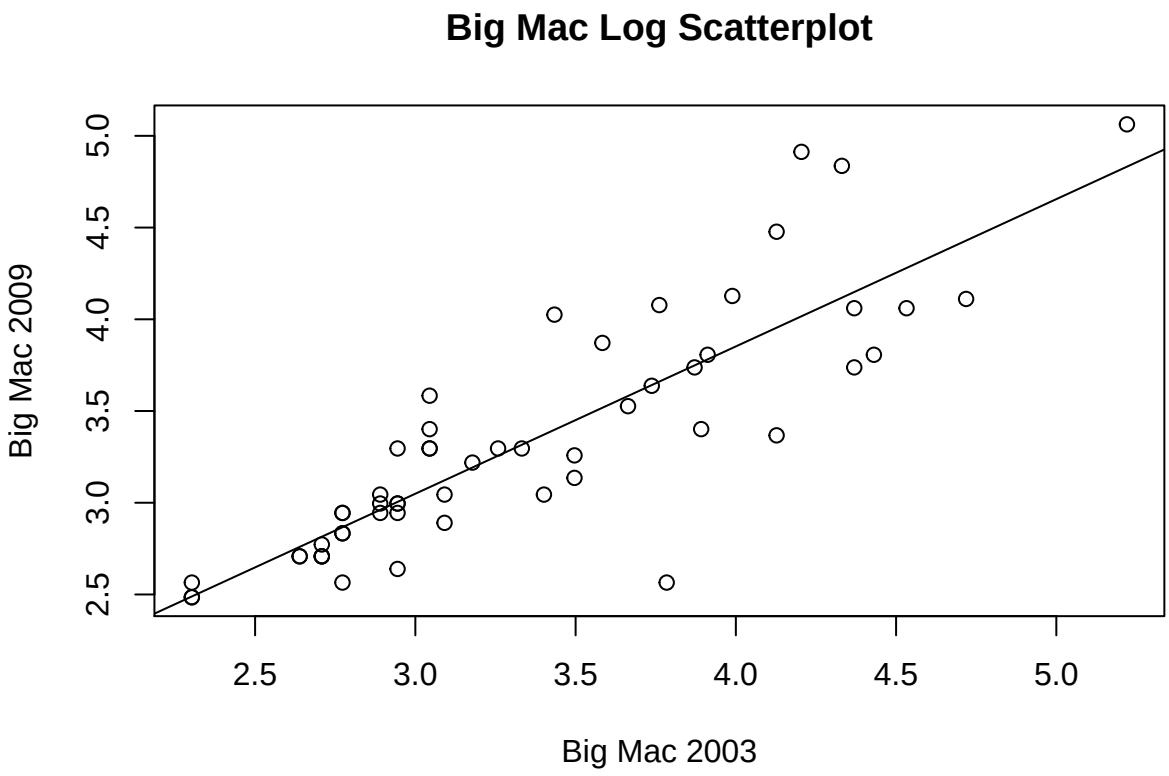


There is a large cluster of points in the bottom left corner, and it fans out and becomes more scarce as you traverse the axes. This clustering suggests that the data may have an exponential trend, and a normalizing function is necessary for modeling.

(b) & (c) Plot log(bigmac2009) versus log(bigmac2003) and explain why this graph is more sensibly summarized with a linear regression. Without using the R function lm(), find the least-squares fit regressing log(bigmac2009) on log(bigmac2003) and add the line in the plot in (b).

```
plot(x=log(UBSprices$bigmac2003), y=log(UBSprices$bigmac2009), xlab="Big Mac 2003",
     ylab="Big Mac 2009", main="Big Mac Log Scatterplot")

x = log(UBSprices$bigmac2003)
y = log(UBSprices$bigmac2009)
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum((x - xbar)^2)
Syy <- sum((y - ybar)^2)
Sxy <- sum((x - xbar)*(y - ybar))
b1 <- Sxy/Sxx
b0 <- ybar - b1*xbar
abline(b0, b1)
```



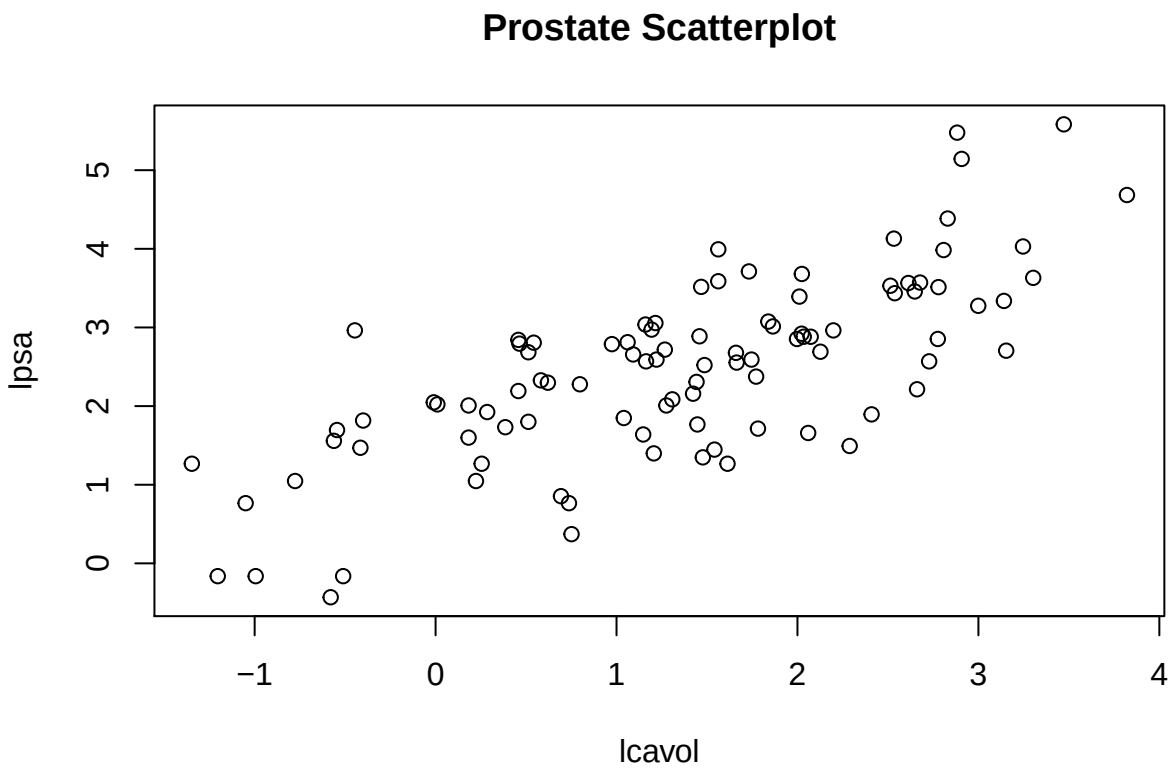
Compared to the plot in (a), the data is much more evenly distributed across the plot, and the regression line gives us more insight in this situation.

3. This problem uses the prostate data set in the faraway package.

(a) Plot lpsa against lcavol. Use the R function lm() to fit the regressions of lpsa on lcavol and lcavol on lpsa.

```
library(faraway)

plot(x=prostate$lcavol, y=prostate$lpsa, xlab="lcavol", ylab="lpsa",
     main="Prostate Scatterplot")
```

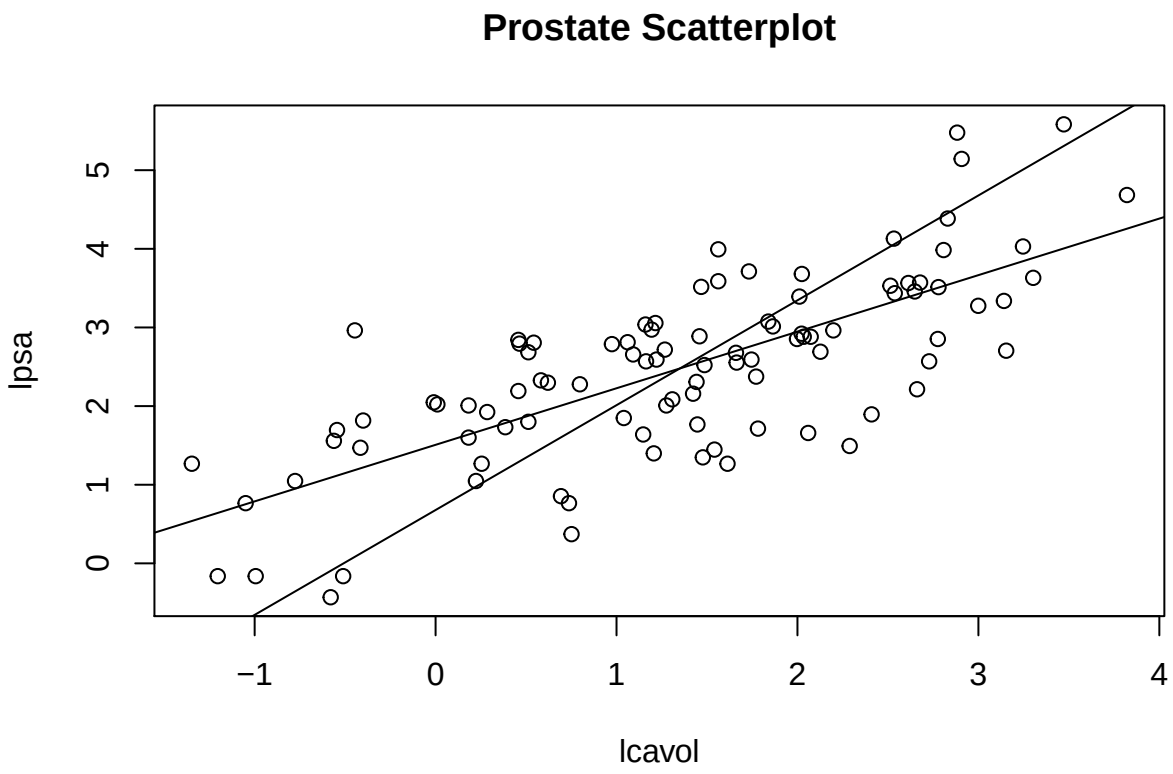



```
xbar <- mean(prostate$lcavol)

lpsa_on_lcavol <- lm(formula=lpsa~lcavol, data=prostate)
lcavol_on_lpsa <- lm(formula=lcavol~lpsa, data=prostate)
```

(b) Display both regression lines on the plot. At what point do the two lines intersetct? Give a brief explanation.

```
plot(x=prostate$lcavol, y=prostate$lpsa, xlab="lcavol", ylab="lpsa",
     main="Prostate Scatterplot")
abline(coef(lpsa_on_lcavol)["(Intercept)"], coef(lpsa_on_lcavol)["lcavol"])
# Solve equation for x
abline(-coef(lcavol_on_lpsa)["(Intercept)"]/coef(lcavol_on_lpsa)["lpsa"],
       1/coef(lcavol_on_lpsa)["lpsa"])
```



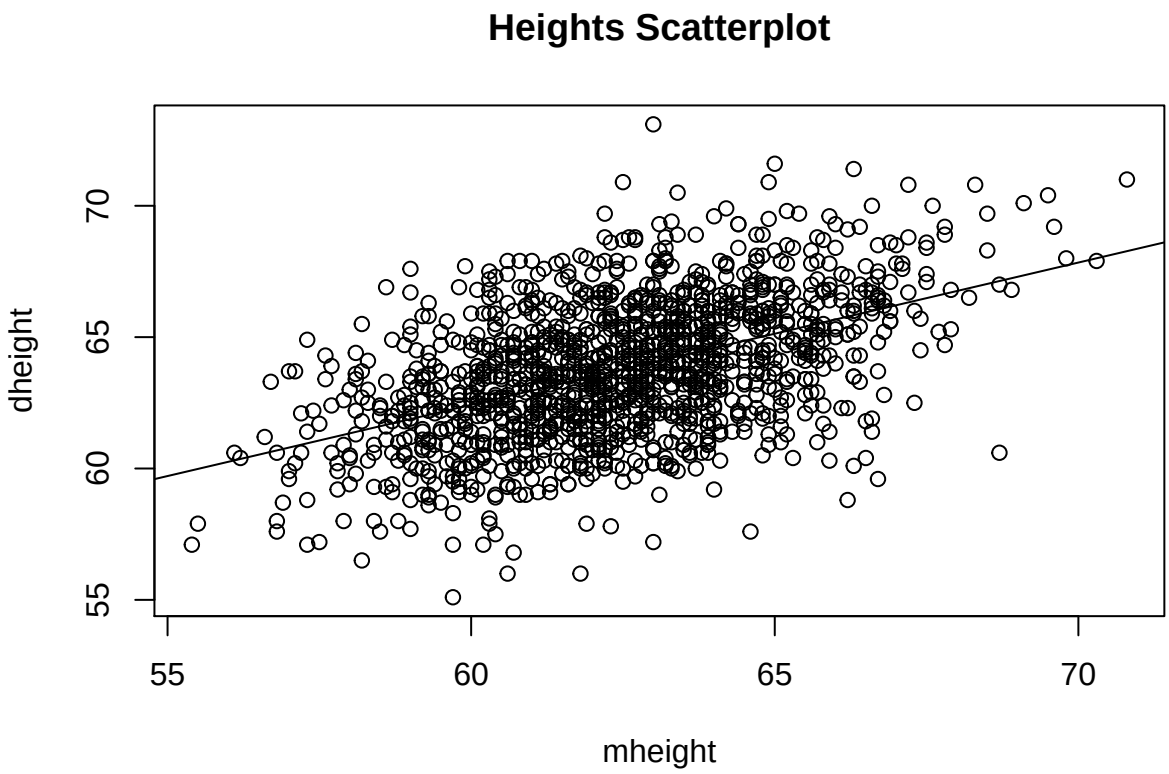
The lines intersect at (xbar, ybar), as every regression line must pass through this point.

4. This problem uses the data set Heights in the alr4 package. Interest is in predicting dheight by mheight.

(a) Use the R function lm() to fit the regression of the response on the predictor. Draw a scatterplot of the data and add your fitted regression line.

```
heights <- read.csv("Heights.csv")

model <- lm(formula=dheight~mheight, data=heights)
plot(x=heights$mheight, y=heights$dheight, xlab="mheight", ylab="dheight", main="Heights Scatterplot")
abline(model)
```



(b) Compute the (Pearson) correlation coefficient rxy. What does the value of rxy imply about the relationship between dheight and mheight?

```
x = heights$mheight
y = heights$dheight
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum((x - xbar)^2)
Syy <- sum((y - ybar)^2)
Sxy <- sum((x - xbar)*(y - ybar))
rxy <- Sxy/sqrt(Sxx*Syy)
cat("rxy: ", rxy)
```

```
## rxy:  0.4907094
```

The correlation coefficient implies a moderate positive linear relationship between mheight and dheight