PSTAT 126 HW 1

Tamjid Islam

4/12/2020

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.

```
# install.packages('alr4')
library(alr4)
```

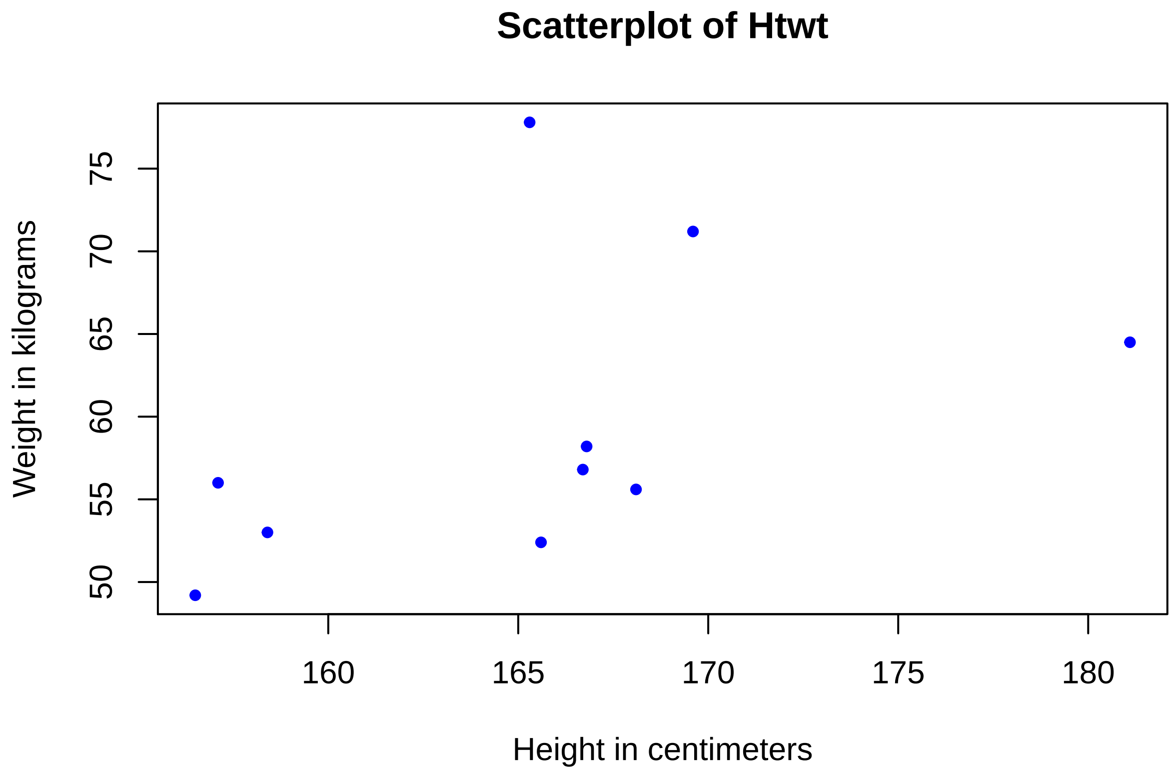
```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

(a) Identify the predictor and response.

Height is predictor and weight is response.

(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?

```
library(alr4)
data(Htwt)
plot(Htwt$ht, Htwt$wt,
     xlab = 'Height in centimeters', ylab = 'Weight in kilograms',
     main = 'Scatterplot of Htwt', pch = 20, col = 'blue')
```



No a simple linear regression model wouldn't make sense for this data because even though there is some scatter there isn't a good linear relationship to height and weight. Therefore, if a regression line was created, the performance of the model would not be the best since the goal of the model is to minimize SSE, meaning we want less scatter.

(c) Show that $\bar{x} = 165.52$, $\bar{Y} = 59.47$, $S_{xx} = 472.08$, $S_{yy} = 731.96$ and $S_{xy} = 274.79$. Compute estimates of the slope and the intercept for the regression of Y on x. Draw the fitted line on your scatterplot.

```
avg1 <- mean(Htwt$ht)
avg1
```

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

```
avg2 <- mean(Htwt$wt)
avg2
```

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

```
Sxx <- sum((Htwt$ht-avg1)^2)
Sxx
```

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

```
Syy <- sum((Htwt$wt-avg2)^2)
Syy
```

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

```
Sxy <- sum((Htwt$ht-avg1)*(Htwt$wt-avg2))
Sxy
```

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

```r
lm1 <- lm(Htwt$wt ~ Htwt$ht)
lm1
```

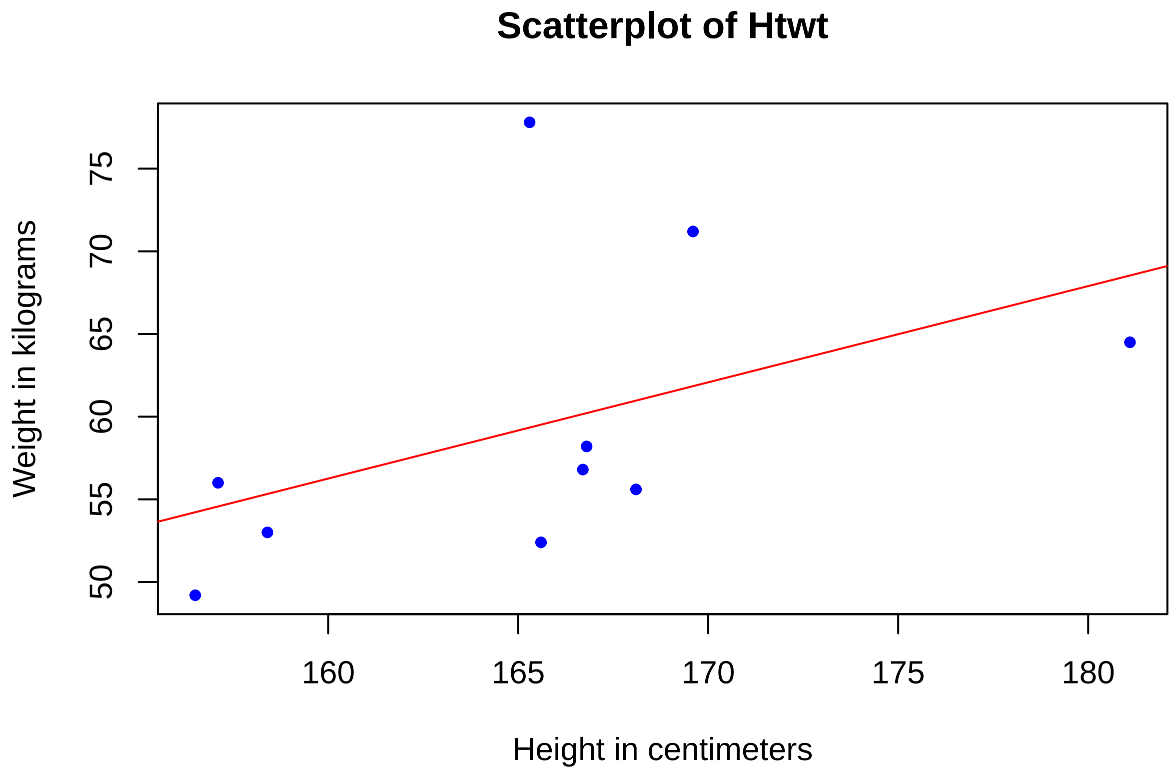
```
##
## Call:
## lm(formula = Htwt$wt ~ Htwt$ht)
##
## Coefficients:
## (Intercept)      Htwt$ht
##    -36.8759       0.5821
```

```r
summary(lm1)
```

```
##
## Call:
## lm(formula = Htwt$wt ~ Htwt$ht)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1166 -4.7744 -2.8412  0.5696 18.4581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.8759    64.4728  -0.572    0.583
## Htwt$ht       0.5821     0.3892   1.496    0.173
##
## Residual standard error: 8.456 on 8 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.1208
## F-statistic: 2.237 on 1 and 8 DF,  p-value: 0.1731
```

The estimate on the slope = 0.5821 and estimate on the intercept = -36.8756.
$\hat{y}$ = -36.8756 + 0.5821x
The standard error of the estimate is the residual standard error = 8.456

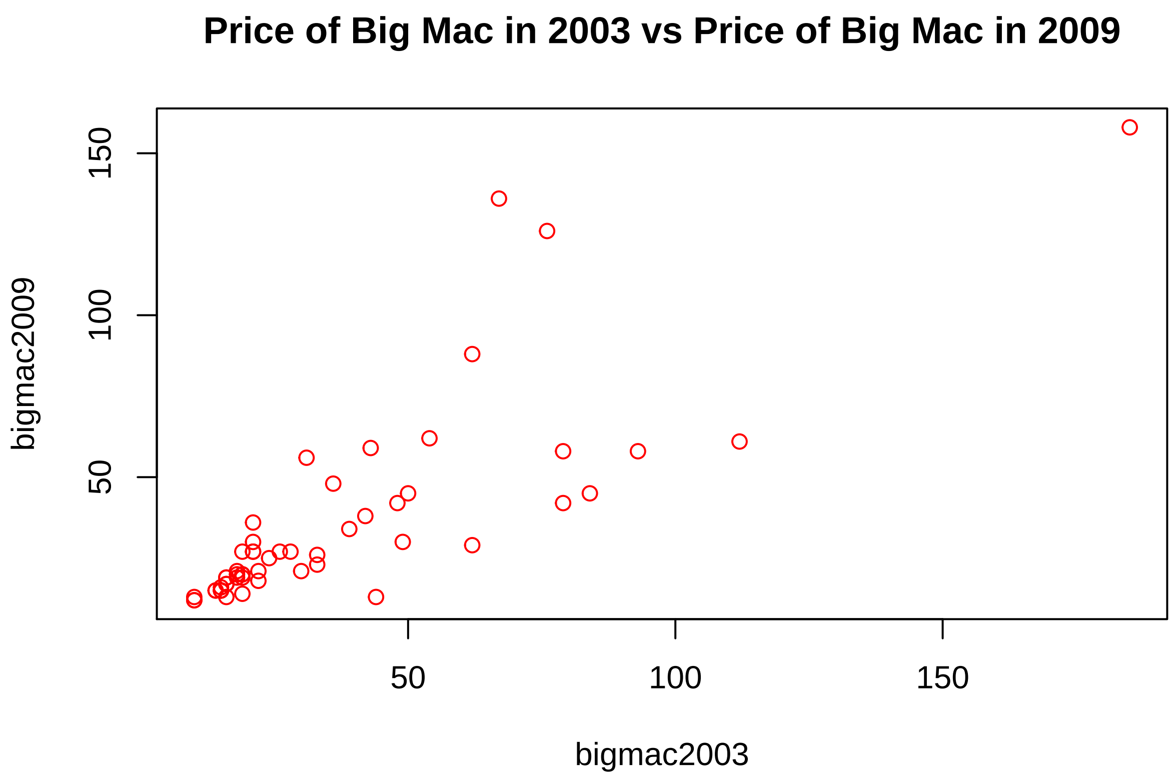
```r
plot(Htwt$ht, Htwt$wt,
     xlab = 'Height in centimeters', ylab = 'Weight in kilograms',
     main = 'Scatterplot of Htwt', pch = 20, col = 'blue')
abline(lm1, col = 'red')
```



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.

```
library(alr4)
data(UBSprices)
plot(UBSprices$bigmac2003, UBSprices$bigmac2009,
     xlab = 'bigmac2003', ylab ='bigmac2009',
     main = 'Price of Big Mac in 2003 vs Price of Big Mac in 2009',
     pch= 1, col = 'red' )
```



The reason why fitting a simple linear regression would not likely be appropriate is beacuse there isn't much scatter even tho there seems to be a upward trend.There is much more of a cluster rather than a scatter. Therefore, there is not a good linear relationship.

(b) Plot log(bigmac2009) versus log(bigmac2003) and explain why this graph is more sensibly summarized with a linear regression. & (c) 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).

```r
x <- log(UBSprices$bigmac2003)
x
```

```
##  [1] 2.772589 3.044522 2.944439 3.912023 3.091042 2.772589 4.532599
##  [8] 3.988984 2.890372 4.369448 3.761200 4.330733 2.302585 2.772589
## [15] 3.091042 2.708050 2.708050 2.772589 2.944439 3.583519 4.204693
## [22] 3.496508 4.430817 4.369448 3.496508 3.663562 2.772589 2.708050
## [29] 3.044522 3.178054 4.127134 2.302585 3.044522 2.639057 3.401197
## [36] 4.718499 5.220356 2.890372 2.944439 3.737670 3.871201 3.258097
## [43] 3.332205 3.044522 3.433987 2.944439 2.944439 2.890372 3.891820
## [50] 2.302585 2.639057 2.708050 4.127134 3.784190
```

```r
y <- log(UBSprices$bigmac2009)
y
```

```
##  [1] 2.944439 3.401197 2.944439 3.806662 3.044522 2.944439 4.060443
##  [8] 4.127134 2.944439 3.737670 4.077537 4.836282 2.484907 2.833213
## [15] 2.890372 2.708050 2.708050 2.833213 3.295837 3.871201 4.912655
## [22] 3.258097 3.806662 4.060443 3.135494 3.526361 2.564949 2.708050
## [29] 3.295837 3.218876 4.477337 2.564949 3.295837 2.708050 3.044522
## [36] 4.110874 5.062595 3.044522 2.995732 3.637586 3.737670 3.295837
## [43] 3.295837 3.583519 4.025352 2.995732 2.639057 2.995732 3.401197
## [50] 2.484907 2.708050 2.772589 3.367296 2.564949
```

```
plot(x,y, xlab = 'bigmac2003', ylab = 'bigmac2009',
     main = 'Price of Big Mac in 2003 vs Price of Big Mac in 2009',
     pch = 1, col = 'red' )
mean1 <- mean(x)
mean1
```

```
## [1] 3.349187
```

```
mean2 <- mean(y)
mean2
```

```
## [1] 3.329467
```

```
Sxy <- sum((x-mean1)*(y-mean2))
Sxy
```

```
## [1] 19.70271
```

```
Sxx <-sum((x-mean1)^2)
Sxx
```

```
## [1] 24.53861
```

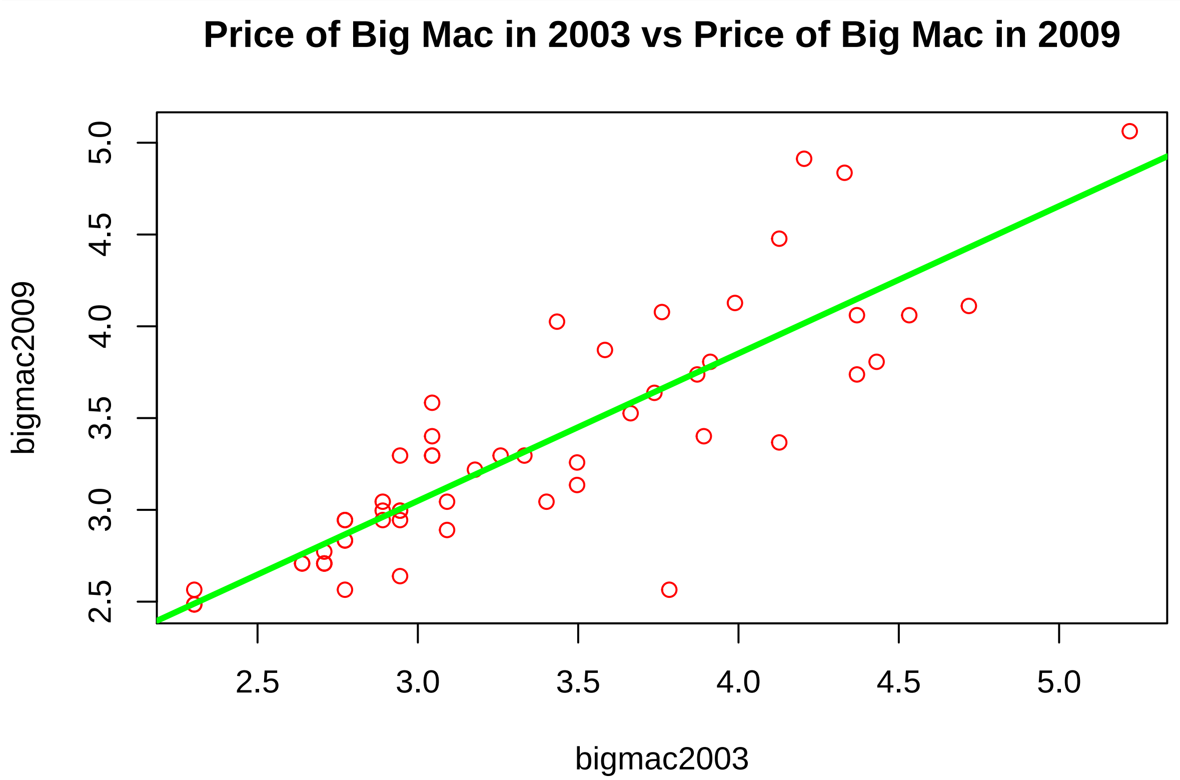
```
slope <- Sxy / Sxx
slope
```

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

```
intercept <- mean2 - (mean1*slope)
intercept
```

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

```
abline(intercept, slope, col = 'green', lwd = 3)
```



This graph is more sensibly summarized with linear regression because it is more normally distributed and shows an upward trend linearly in relation to bigmac2003 and bigmac2009. The line of best fit is $\hat{y}$ = 0.6403 + 0.8029x

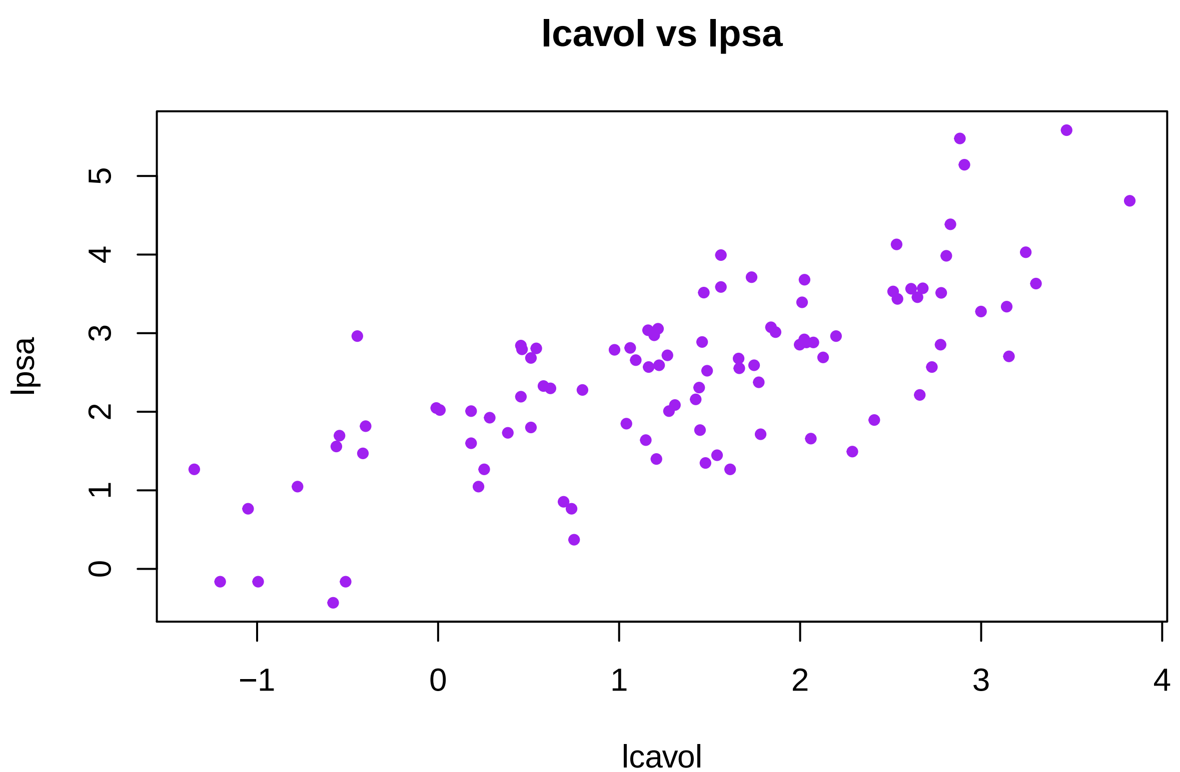
3. This problem uses the prostate data set in the faraway package.

```
# install.packages('faraway')
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins

## The following objects are masked from 'package:car':
##
##     logit, vif
```

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

```
library(faraway)
data(prostate)
plot(prostate$lcavol, prostate$lpsa,
     xlab = 'lcavol', ylab = 'lpsa', main = 'lcavol vs lpsa',
     pch = 20, col = 'purple')
```




```
fit1 <- lm(prostate$lpsa ~ prostate$lcavol)
fit1
```

```
##
```

```
## Call:
## lm(formula = prostate$lpsa ~ prostate$lcavol)
##
## Coefficients:
##     (Intercept)  prostate$lcavol
##          1.5073           0.7193
```

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = prostate$lpsa ~ prostate$lcavol)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.50730    0.12194   12.36   <2e-16 ***
## prostate$lcavol    0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

```r
fit2 <-lm(prostate$lcavol ~ prostate$lpsa)
fit2
```

```
##
## Call:
## lm(formula = prostate$lcavol ~ prostate$lpsa)
##
## Coefficients:
##   (Intercept)  prostate$lpsa
##       -0.5086         0.7499
```

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = prostate$lcavol ~ prostate$lpsa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15948 -0.59383  0.05034  0.50826  1.67751
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.50858    0.19419  -2.619   0.0103 *
## prostate$lpsa  0.74992    0.07109  10.548   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.8041 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

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

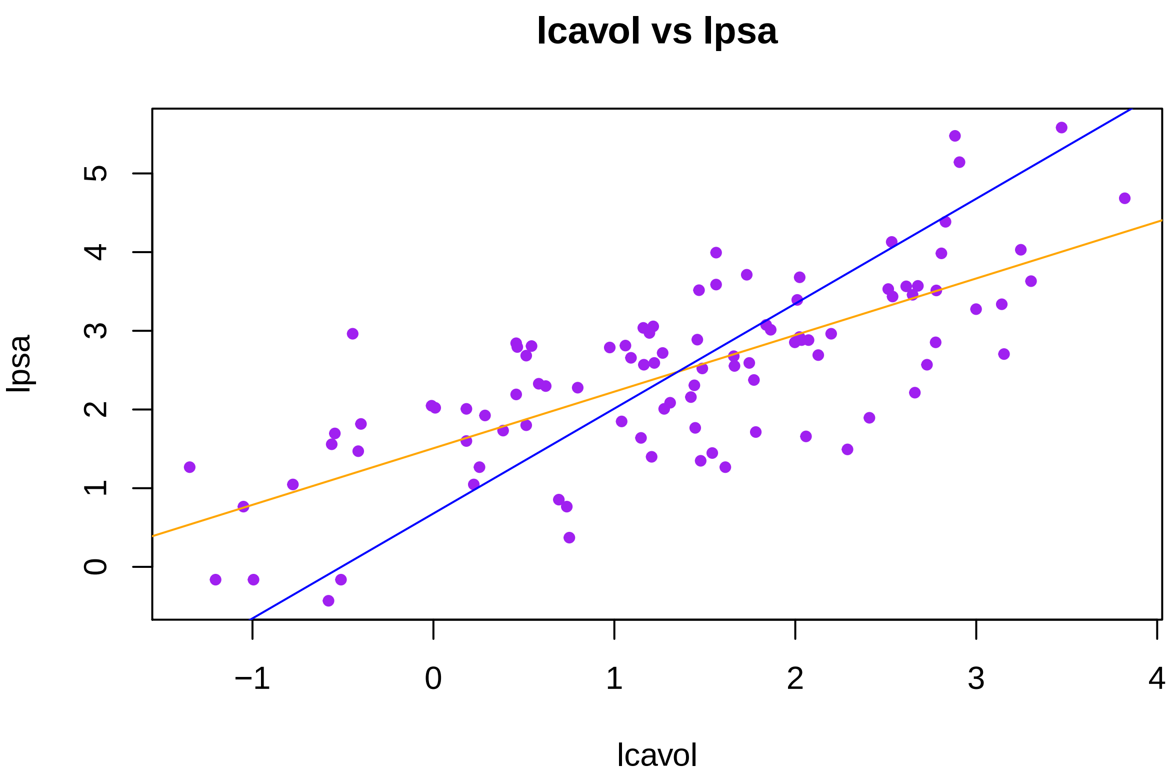
```
plot(prostate$lcavol, prostate$lpsa, xlab = 'lcavol', ylab = 'lpsa',
     main = 'lcavol vs lpsa', pch = 20, col = 'purple')
abline(fit1, col = 'orange')
fit2_slope <- 1/ (fit2$coeff[2])
fit2_slope
```

```
## prostate$lpsa
##      1.333477
```

```
fit2_intercept <- (-fit2$coeff[1]/fit2$coeff[2])
fit2_intercept
```

```
## (Intercept)
##     0.67818
```

```
abline(fit2_intercept,fit2_slope, col = 'blue')
```




```
mean2 <-mean(prostate$lcavol)
mean2
```

```
## [1] 1.35001
```

```
mean3 <-mean(prostate$lpsa)
mean3
```

```
## [1] 2.478387
```

The point where both regression lines intersect is where the mean of lpsa and the mean of lcavol meet. The point of intersection is (1.35001,2.478387).

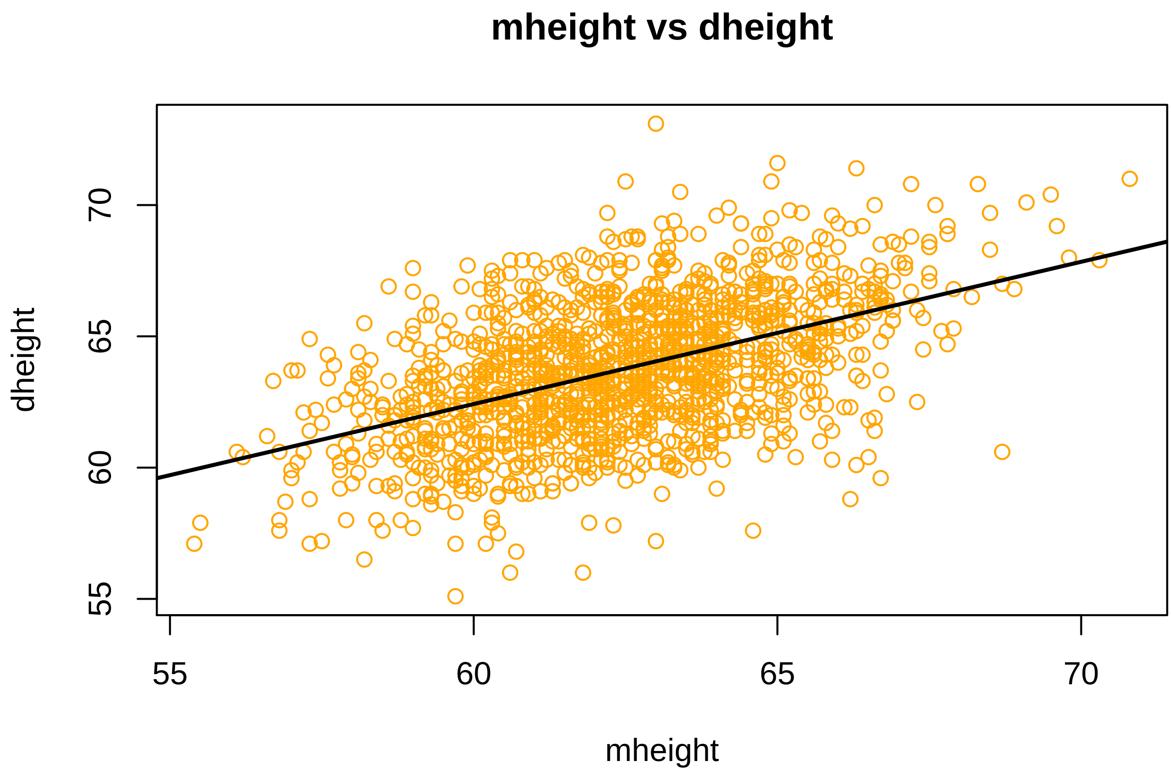
4. This problem uses the data set Heights in the alr4 package. Interest is in predicting dheight by mheight.

(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.

```
library(alr4)
data(Heights)
plot(Heights$mheight, Heights$dheight,
     xlab = 'mheight', ylab = 'dheight',
     main = 'mheight vs dheight', pch = 1, col = 'orange')
fit3 <- lm(Heights$dheight ~ Heights$mheight)
fit3
```

```
##
## Call:
## lm(formula = Heights$dheight ~ Heights$mheight)
##
## Coefficients:
##     (Intercept)  Heights$mheight
##         29.9174           0.5417
```

```
abline(fit3, col = 'black', lwd = 2)
```



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

```
mean_m <- mean(Heights$mheight)
mean_m
```

```
## [1] 62.4528
```

```r
mean_d <- mean(Heights$dheight)
mean_d
```

```
## [1] 63.75105
```

```r
Sxx1 <- sum((Heights$mheight-mean_m)^2)
Sxx1
```

```
## [1] 7620.907
```

```r
Syy1 <- sum((Heights$dheight-mean_d)^2)
Syy1
```

```
## [1] 9288.616
```

```r
Sxy1 <- sum((Heights$mheight-mean_m)*(Heights$dheight-mean_d))
Sxy1
```

```
## [1] 4128.603
```

```r
correlation <- (Sxy1/(sqrt(Sxx1*Syy1)))
correlation
```

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

The Pearson correlation is 0.4907. This shows that the relationship between dheight and mheight is an upward positive linear relation where the strength of the correlation is measured at 0.4907.

5. We are now given data on n observations $(x_i,\ Y_i)$, i = 1, . . . , n. Assume we have a linear model, so that $E(Y_i) = \beta_0 + \beta_1 x_i$, and let $b_1 = \frac{S_{xy}}{S_{xx}}$ and $b_0 = \bar{Y} - b_1\bar{x}$ be the least-square estimates given in lecture.

(a) Show that $E(S_{xy}) = \beta_1 S_{xx}$ and $E(\bar{Y}) = \beta_0 + \beta_1\bar{x}$, and use this to conclude that $E(b_1) = \beta_1$ and $E(b_0) = \beta_0$. In other words, these are unbiased estimators.

$$E(S_{xy}) = \sum_{i=1}^{n}[E(x_i - \bar{x})(y_i - \bar{y})] = \sum_{i=1}^{n}[E(x_i - \bar{x})(y_i)] = \sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_1\sum_{i=1}^{n}(x_i - \bar{x})(x_i) = \beta_1 S_{xx}$$

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}[E(Y_i)] = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)] = \beta_0 + \beta_1\bar{x}$$

$$E(b_1) = \frac{1}{S_{xx}}\sum_{i=1}^{n}[E(S_{xy})] = \frac{1}{S_{xx}}\sum_{i=1}^{n}\beta_1(S_{xx})] = \beta_1$$

$$E(b_0) = E[\bar{Y} - b_1\bar{x}] = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

(b) The fitted values $\hat{Y}_i = b_0 + b_1 x_i$ are used as estimates of $E(Y_i)$, and the residuals $e_i = Y_i - \hat{Y}_i$ are used as surrogates for the unobservable errors $\epsilon_i = Y_i - E(Y_i)$. By assumption, $E(\epsilon_i) = 0$. Show that the residuals satisfy a similar property, namely, $\sum_{i=1}^{n} e_i = 0$.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i) = n\bar{Y} - (n\bar{Y} - nb_1\bar{x}) - nb_1\bar{x} = n\bar{Y} - n\bar{Y} + nb_1\bar{x} - nb_1\bar{x} = 0$$