PSTAT 126 Homework 2

Shaiyon Hariri

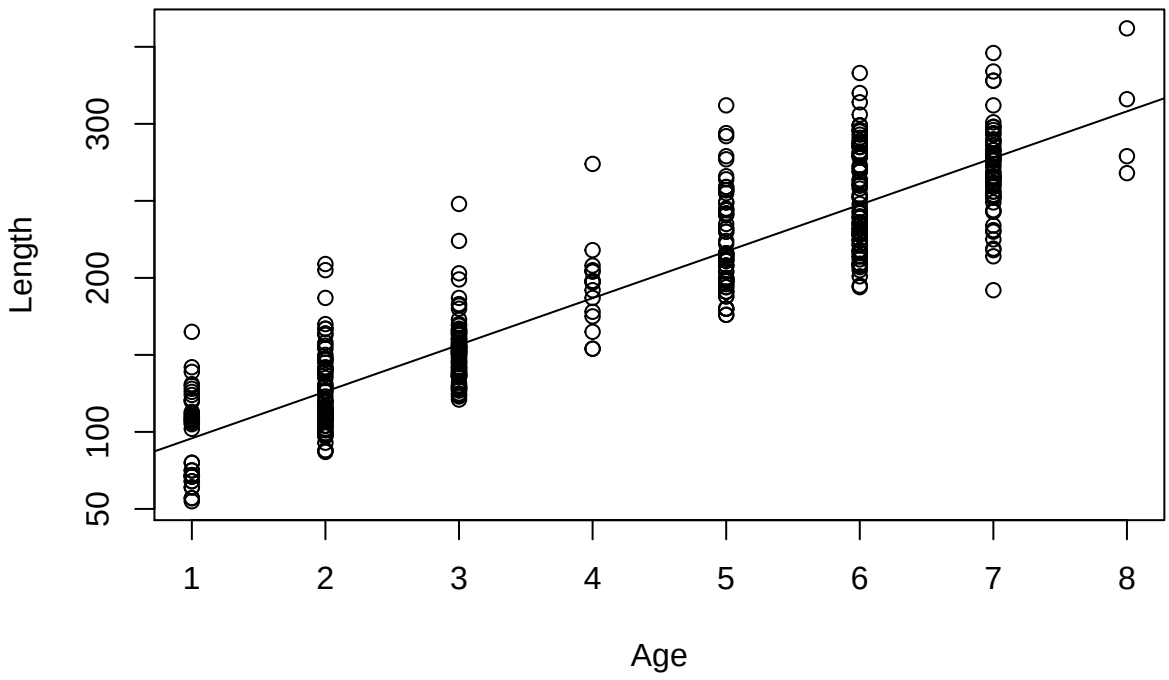
4/23/2020

1. This problem uses the wblake data set in the alr4 package. This data set includes samples of small mouth bass collected in West Bearskin Lake, Minnesota, in 1991. Interest is in predicting length with age. Finish this problem without using lm().

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

(a) Compute the regression of length on age, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the results of these computations.

```
n <- length(wblake$Age)
x <- wblake$Age
y <- wblake$Length
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
yhat <- b0 + b1*x

e <- y - yhat
sigma2hat <- sum(e^2)/(n-2)
sigmahat <- sqrt(sigma2hat)

ssto <- sum((y-ybar)^2)
sse <- sum((y-yhat)^2)
ssr <- sum((yhat-ybar)^2)
r2 <- ssr/ssto

plot(x, y, xlab="Age", ylab="Length")
abline(b0, b1)
```




```
cat("yhat:", b0, "+", b1, "* x", " standard error:", sigmahat, " variance:", sigma2hat,
    " r2:", r2)
```

```
## yhat: 65.52716 + 30.32389 * x  standard error: 28.64585  variance: 820.5847  r2: 0.816477
```

The R squared is very high, suggesting that Age accounts for a lot (81.65%) of the variation in Length. b1 is positive, and the graph shows an increasing slope, so as age increases, length also increases.

(b) Obtain a 99% confidence interval for B1 from the data. Interpret this interval in the context of the data.

```
se_b1 <- sigmahat / sqrt(Sxx)
t_pct <- qt(p = 0.99,df = n-2)
ci_B1_99 <- b1 + c(-1, 1) * t_pct * se_b1
cat("99% Confidence interval for B1:", ci_B1_99)
```

```
## 99% Confidence interval for B1: 28.7181 31.92968
```

This interval is small, so we can say with relative certainty that for each year older a small mouth bass is, it will be approximately 30 mm longer.

(c) Obtain a prediction and a 99% prediction interval for a small mouth bass at age 1. Interpret this interval in the context of the data.

```
ynew <- yhat[x=1]
se_age1 <- sigmahat * sqrt(1 + 1/n + (1-xbar)^2/Sxx)
t_pct <- qt(p = 0.995, df = n-2)
```

```r
ci_age1_99 <- ynew + c(-1, 1) * t_pct * se_age1
cat("Ynew:", ynew, " 99% Prediction interval for Age=1:", ci_age1_99)
```

```
## Ynew: 95.85105  99% Prediction interval for Age=1: 21.43775 170.2644
```

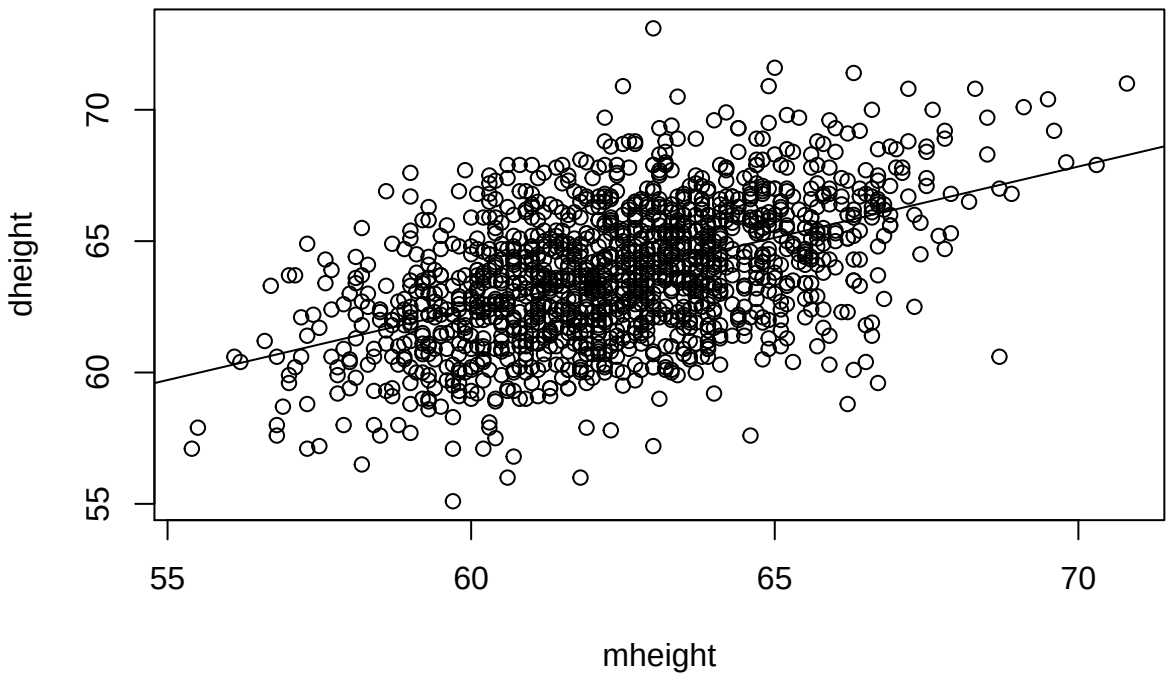
This is a large interval, from tiny (25 mm) to medium sized fish (170mm). This large interval is due to the high variance of length and the small significance level, making pinpointing one estimated value with certainty difficult.

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

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

(a) Compute the regression of dheight on mheight, and report the estimates, their standard errors, the value of the coeffcient of determination, and the estimate of the variance.

```r
n <- length(heights$mheight)
x = heights$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))
b1 <- Sxy/Sxx
b0 <- ybar - b1*xbar
yhat <- b0 + b1*x

e <- y - yhat
sigma2hat <- sum(e^2)/(n-2)
sigmahat <- sqrt(sigma2hat)

ssto <- sum((y-ybar)^2)
sse <- sum((y-yhat)^2)
ssr <- sum((yhat-ybar)^2)
r2 <- ssr/ssto

plot(x, y, xlab="mheight", ylab="dheight")
abline(b0, b1)
```




```
cat("yhat:", b0, "+", b1, "* x", " standard error:", sigmahat, " variance:", sigma2hat,
    " r2:", r2)
```

```
## yhat: 29.91744 + 0.541747 * x  standard error: 2.266311  variance: 5.136167  r2: 0.2407957
```

(b) For this problem, give an interpretation for B0 and B1.

B0 is the population intercept and B1 is the population coefficient for mheight. This means that when mheight is 0, dheight will be the value of Bo, but as height cannot be zero or very close to it, this result cannot have significant insights extracted from it. However, the value of B1 implies a positive slope of the regression line, and that for every 1 value of mheight, dheight will increase by B1.

(c) Obtain a prediction and a 99% prediction interval for a daughter whose mother is 64 inches tall.

```
ynew <- yhat[x=64]
se_mh64 <- sigmahat * sqrt(1 + 1/n + (1-xbar)^2/Sxx)
t_pct <- qt(p = 0.995, df = n-2)
ci_mh64_99 <- ynew + c(-1, 1) * t_pct * se_age1
cat("Ynew:", ynew, " 99% Prediction interval for mheight=64:", ci_mh64_99)
```

```
## Ynew: 62.85566  99% Prediction interval for mheight=64: -11.33584 137.0472
```

4. This problem uses the UBSprices data set in the alr4 package. The international bank UBS regularly produces a report (UBS, 2009) on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1 kg of rice, a 1 kg loaf of bread, and the price of a Big Mac hamburger at McDonalds.

```
library(alr4)
```

(a) The line with equation Y = x is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?

Points above the line have had an increase in the price of rice relative to a typical worker's wages from 2003 to 2009, and points below the line have had a decrease in price.

(b) Which city had the largest increase in rice price? Which had the largest decrease in rice price?

```
diff <- UBSprices$rice2009 - UBSprices$rice2003
maxprice <- UBSprices[which(diff == max(diff)),]
minprice <- UBSprices[which(diff == min(diff)),]

cat("Largest increase in rice price:", row.names(maxprice),
    " Largest decrease in rice price:", row.names(minprice))
```

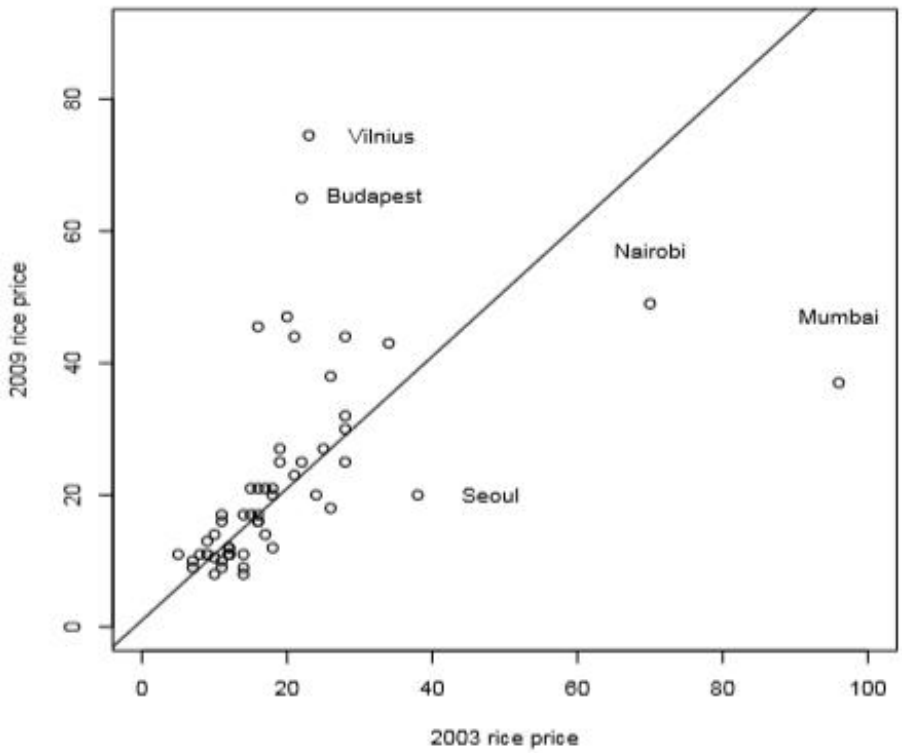
```
## Largest increase in rice price: Vilnius  Largest decrease in rice price: Mumbai
```

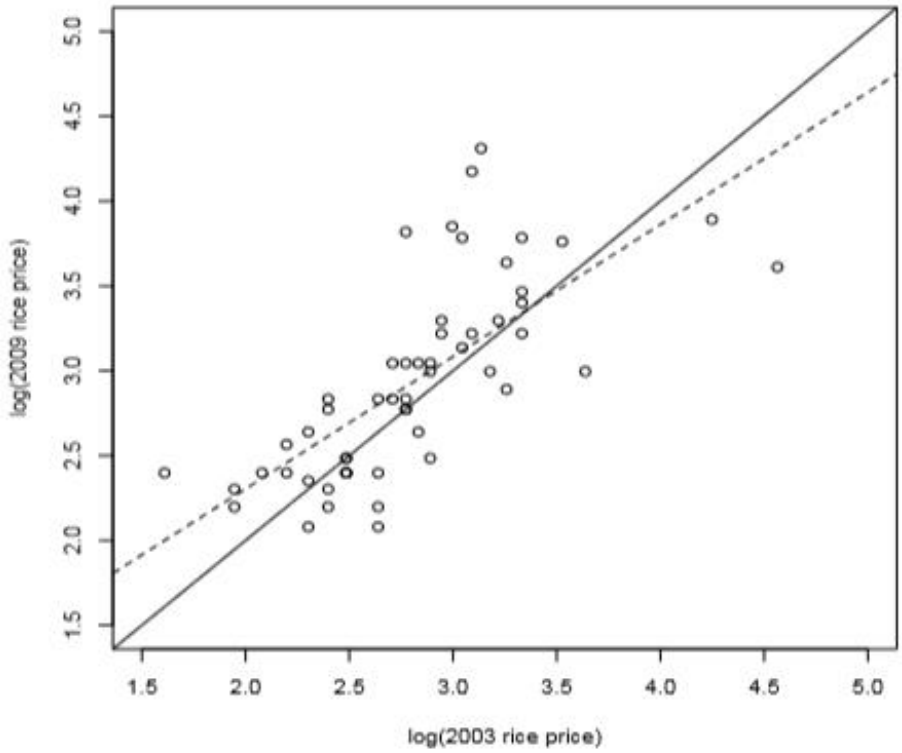
(c) Give at least one reason why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

The data does not seem to be distributed in a way that can be modeled lienarly (e.g.: normal). It clusters in the bottom left of the graph and fans out as x axis increases,

(d) The second graph represents Y and x using log scales. Explain why this graph and the previous graph suggests that using log scales is preferable if fitting simple linear regression is desired. The linear model is shown by the dashed line.

The graph using the log scales shows that the data isn't clustered in the same way as the first graph, and is more normally distributed around the regression line. The residuals will be lower as a result, leading to a better model with better predictive power.


Vilnius
Budapest
Nairobi
Mumbai
Seoul
2009 rice price
2003 rice price
0
20
40
60
80
100


log(2009 rice price)
log(2003 rice price)
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0