PSTAT 126 - Homework 2

Due: 11:55 p.m. Friday, April 24

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()**.

   (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.

   (b) Obtain a 99% confidence interval for β_1 from the data. Interpret this interval in the context of the data.

   (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.

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

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

   (b) For this problem, give an interpretation for β_0 and β_1.

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

3. The simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$ can also be written as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Using matrix notations, the model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In this problem, we will show that that the least squares estimate is given by:

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X'X)^{-1}X'Y$$

   (a) Using straightforward matrix multiplication, show that

$$X'X = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & S_{xx}/n + \bar{x}^2 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ S_{xy} + n\bar{x}\bar{Y} \end{pmatrix}$$

(b) Using the identity

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

for a 2×2 matrix, show that

$$(X'X)^{-1} = \frac{1}{S_{xx}}\begin{pmatrix} S_{xx}/n + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

(c) Combine your answers from (a) and (b) to show that

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X'X)^{-1}X'Y$$

where $b_1 = \frac{S_{xy}}{S_{xx}}$ and $b_0 = \bar{Y} - b_1\bar{x}$ are the least squares estimates from simple linear regression.

(d) Simulate a data set with $n = 100$ observation units such that $Y_i = 1 + 2x_i + \varepsilon_i$, $i = 1, \ldots, n$. ε_i follows the standard normal distribution, i.e., a normal distribution with zero mean and unit variance. Use the result in (c) to compute b_0 and b_1. Show that they are the same as the estimates by **lm()**. Start with generating x as

```
n = 100
x = seq(0, 1, length = n)
```

(Hint: check the help page of **rnorm()** about how to simulate normally distributed random variables. Use **solve()** to get an inverse matrix and use **t()** to get a transpose matrix.).

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.

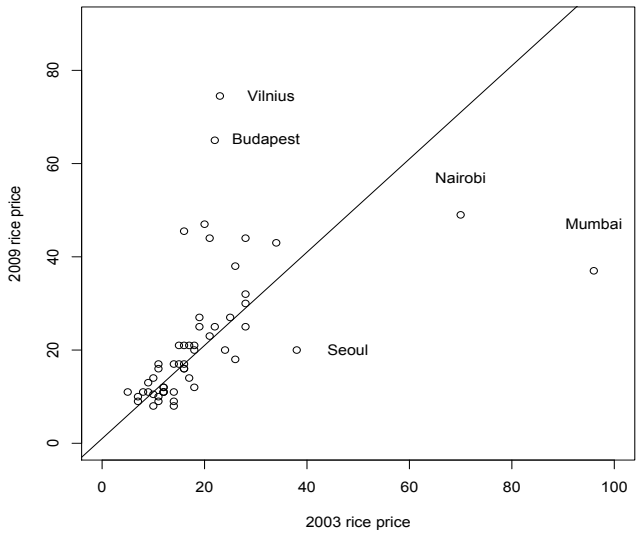
An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a "typical" worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession.
The first graph below is the plot of $Y =$ rice2009 versus $x =$ rice2003, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points marked.

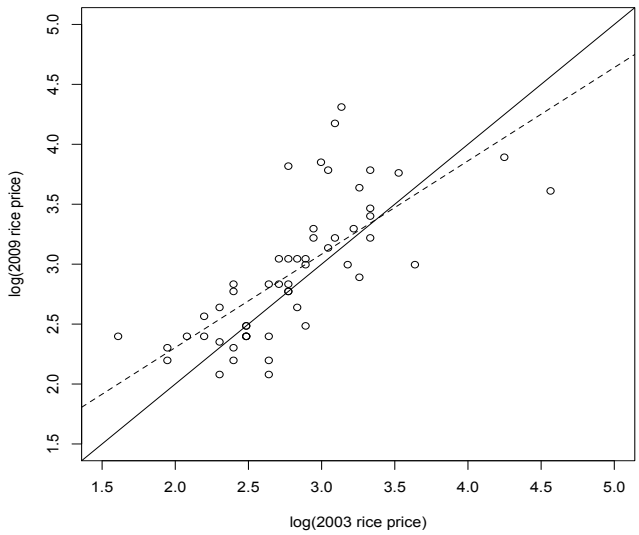
(a) The line with equation $Y = x$ is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?

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

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

(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.


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