**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 3**

Prepare a short report with relevant output, your comments, and answers to the questions (this does not need to be exhaustive or polished, but should contain enough to show that you completed all tasks and analyses).

Submit the report at the end of the lab session.

Consider again the dataset *record.txt* that you used for Lab 1.

This dataset contains running records obtained from athletes from different countries in various types of athletics events (sprints and middle-distance).

We have data about 55 countries (observations) and 6 records (variables): 100 meters, 200 meters, 400 meters, 800 meters, 1500 meters and 3000 meters.

* Load the dataset *record.txt* in R, using the function *read.table* (remember to set sep=' ')
* Draw a scatterplot and compute the correlation for all pairs of variables in the dataset. Interpret the results you obtained: what can you observe about the relationship among the variables?

> record=read.table("record.txt",header=TRUE,sep=' ')

> head(record)

m100 m200 m400 m800 m1500 m3000

argentin 11.61 22.94 54.50 129.0 265.8 587.4

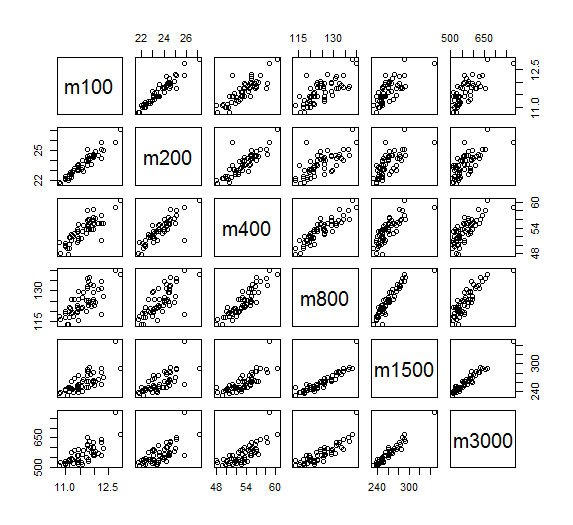
australi 11.20 22.35 51.08 118.8 247.8 544.8

austria 11.43 23.09 50.62 119.4 253.2 560.4

belgium 11.41 23.04 52.00 120.0 248.4 532.8

bermuda 11.46 23.05 53.30 129.6 274.8 588.6

brazil 11.31 23.17 52.80 126.0 269.4 586.2



> cor(record)

m100 m200 m400 m800 m1500 m3000

m100 1.0000000 0.9527911 0.8346918 0.7276888 0.7283709 0.7416988

m200 0.9527911 1.0000000 0.8569621 0.7240597 0.6983643 0.7098710

m400 0.8346918 0.8569621 1.0000000 0.8984052 0.7878417 0.7776369

m800 0.7276888 0.7240597 0.8984052 1.0000000 0.9016138 0.8635652

m1500 0.7283709 0.6983643 0.7878417 0.9016138 1.0000000 0.9691690

m3000 0.7416988 0.7098710 0.7776369 0.8635652 0.9691690 1.0000000

**By observing the graph, all variables are positively correlative with each other.**

**By computing the correlation for all pairs of variables, each pair of variables has strong correlation because all correlations computed is between 1.0 and 0.5.**

Consider the variables m100 and m400.

* Using the equations, compute the least square estimators for the coefficients of a single linear regression model, with response m400 and predictor m100. How do you interpret the slope or the regression line?

> x=record$m100

> y=record$m400

> beta1=sum((y-mean(y))\*(x-mean(x)))/sum((x-mean(x))^2)

> beta0=mean(y)-beta1\*mean(x)

> c(beta0,beta1)

[1] -4.032628 4.943686

**beta0= -4.032628**

**beta1= 4.943686**

**Since the beta1= 4.943686, which is the slope of the regression line, is positive, variables m100 and m400 are positively correlated.**

**That means:**

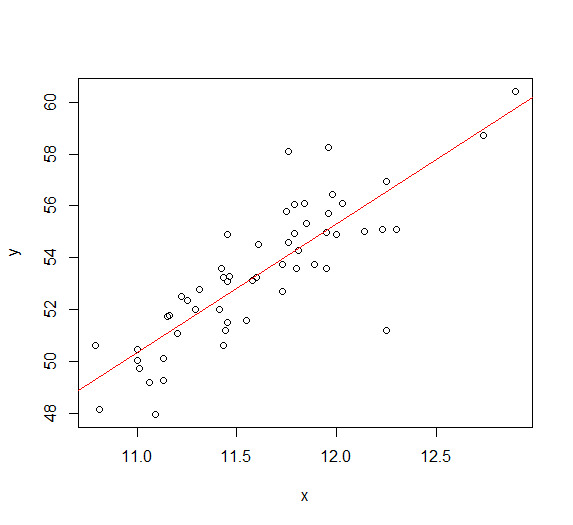
**An increase of 1 second 100 meters running is associated with an average increase of 4.943686 seconds of 400 meters running.**

* Produce a scatter plot for m400 vs m100 with the fitted regression line superimposed.

|  |  |
| --- | --- |
| |  | | --- | |  | |
| |  | | --- | |  | |

> plot(x,y)

> abline(beta0,beta1,col='red')



R code:

setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/math462")

getwd()

record=read.table("record.txt",header=TRUE,sep=' ')

head(record)

plot(record)

cor(record)

x=record$m100

y=record$m400

beta1=sum((y-mean(y))\*(x-mean(x)))/sum((x-mean(x))^2)

beta0=mean(y)-beta1\*mean(x)

c(beta0,beta1)

plot(x,y)

abline(beta0,beta1,col='red')

After running:

> setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/stat462")

> getwd()

[1] "\\\\udrive.win.psu.edu/Users/j/q/jql5883/Desktop/stat462"

>

> record=read.table("record.txt",header=TRUE,sep=' ')

> head(record)

m100 m200 m400 m800 m1500 m3000

argentin 11.61 22.94 54.50 129.0 265.8 587.4

australi 11.20 22.35 51.08 118.8 247.8 544.8

austria 11.43 23.09 50.62 119.4 253.2 560.4

belgium 11.41 23.04 52.00 120.0 248.4 532.8

bermuda 11.46 23.05 53.30 129.6 274.8 588.6

brazil 11.31 23.17 52.80 126.0 269.4 586.2

>

> plot(record)

>

> cor(record)

m100 m200 m400 m800 m1500 m3000

m100 1.0000000 0.9527911 0.8346918 0.7276888 0.7283709 0.7416988

m200 0.9527911 1.0000000 0.8569621 0.7240597 0.6983643 0.7098710

m400 0.8346918 0.8569621 1.0000000 0.8984052 0.7878417 0.7776369

m800 0.7276888 0.7240597 0.8984052 1.0000000 0.9016138 0.8635652

m1500 0.7283709 0.6983643 0.7878417 0.9016138 1.0000000 0.9691690

m3000 0.7416988 0.7098710 0.7776369 0.8635652 0.9691690 1.0000000

>

> x=record$m100

> y=record$m400

>

> beta1=sum((y-mean(y))\*(x-mean(x)))/sum((x-mean(x))^2)

> beta0=mean(y)-beta1\*mean(x)

> c(beta0,beta1)

[1] -4.032628 4.943686

>

> plot(x,y)

> abline(beta0,beta1,col='red')

The following part will NOT be considered in grading the lab report.

* Re-compute the least square estimators of beta0 and beta1 using the matrix equation.
* Re-compute the least square estimators of beta0 and beta1 using the R function *lm*, and visualize the summary of the regression.
* Which are n and p for the considered regression model?
* Compute the fitted values and the residuals, using the estimated regression line.
* Consider the 25th and 75th percentiles of the variable m100 on the dataset. Use the estimated regression line to estimate the mean of the variable m400 at each of these two percentiles.