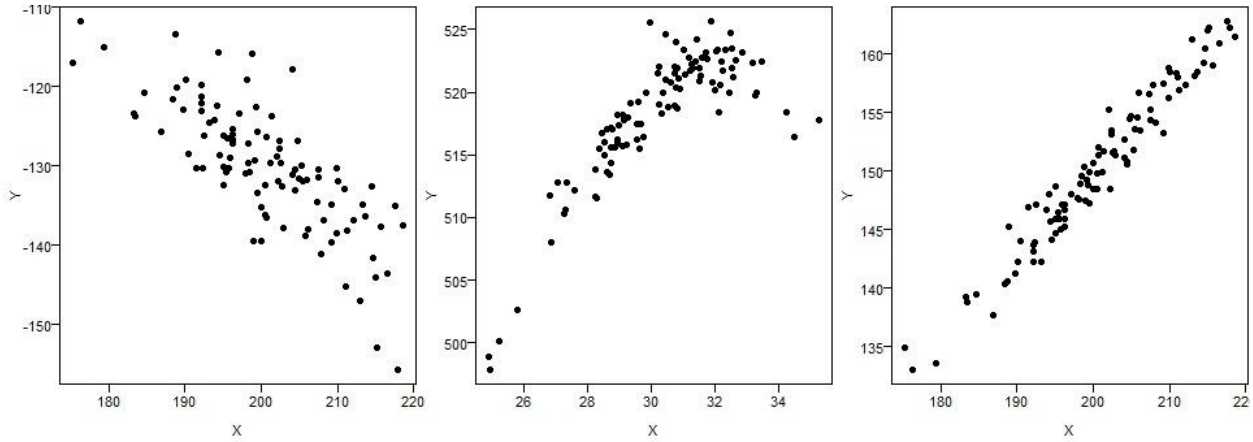
Name: Steve G. Mwangi
Date: December 8, 2019
Assignment: TMATH 390 R Lab 10 Document.

Lab 10: Correlation and linear regression

C1. (4) Submit your R script for this lab.
Submitted

C2. (2) For each scatterplot decide whether linear regression and correlation analysis is appropriate. Explain your reasoning.



Scatter plot 1: Linear regression and correlation analysis is appropriate. Most of the observations follow a linear correlation pattern, even though it is negative in nature.

Scatter plot 2: Linear regression and correlation analysis is not appropriate. At the upper right quadrant of the scatter plot, it seems the correlation in the data is not necessarily linear.

Scatter plot 3: Linear regression and correlation analysis is appropriate. Seems many of the observations follow a positive linear correlation.

C3. (1) In your lab document identify the response and predictor variable for this linear model.
Response: height (cm)
Predictor: shoe print length (cm)

C4 (1) In your lab document identify the estimated intercept and the estimate slope values.
β_0 = 80.930 cm
β_1 = 3.219

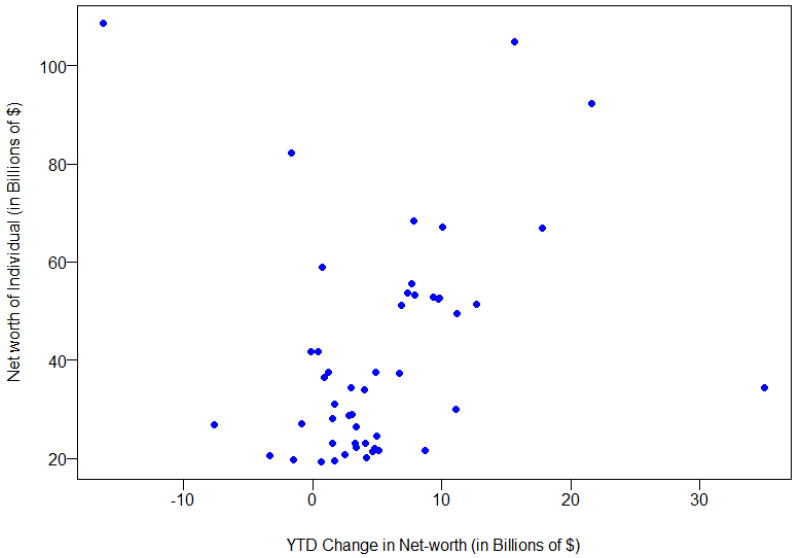
C5. (2) For each shoe print length, use the linear model to predict the height of a suspect. Indicate whether you believe prediction is reasonable given the value of X.
Shoe.Print = {32 cm,28 cm,15 cm}
Corresponding Heights = {183.938cm, 171.062cm, 129.215cm}
r = 0.6609, which shows moderate correlation. Prediction is reasonable for the first 2 heights, but the height at 129.215cm is too small, since it might mean that the criminal is either a midget or underage.

Part 2: Linear regression on your own data
C6. (2) Describe the two quantitative variables in the data set you obtained for HW 1. Choose one to be the response variable (Y), and one to be the predictor variable (X). Explain your choice. Do you think these variables might be related? Why or why not?

Read your data into the current R session (see Lab 3). If you kept your script from lab 3, you can just reuse those lines of code.

X = YTD Change in Net-worth (independent). This change is a better ***predictor*** of Net worth of individuals.
Y = Net worth of Individual (dependent). This ***responds*** to changes in YTD Change.
All values are in Billions of Dollars.
These variables are related since YTD has a direct effect on the overall net worth of the individuals.

C7. (1) Produce a publication-quality scatterplot with your response-variable on the y-axis and your predictor variable on the x-axis. Below is example code for creating publication-quality scatterplot, assuming your data object is data.df, and your column for the response variable is called "Y" and the column for the predictor variable is called "X". Copy and paste your plot into your lab document.



Scatterplot 1 Networth vs YTD Change (in Billions of $)

C8. (1) Describe what you see in your scatterplot. In particular do you believe a linear correlation is meaningful or appropriate? Explain.

The linear correlation in this case might not be meaningful since it is all lumped together. There is negative correlation between YTD Change and Networth, for those whose return on investment was less than 0 (negative YTD Change); and positive correlation between YTD Change and Networth, for those whose return on investment was greater than 0 (positive YTD Change).

C9. (1) Compute the correlation coefficient between X and Y (cor function in R). See below for an example.
r = 0.1776805
Weak correlation

C10. (1) Interpret the value for your correlation coefficient. Does this indicate there might be a strong linear relationship between Y and X?
r = 0.1776805 suggests Weak correlation. Hence there is a weak linear relationship between Net worth and YTD Change.

C11. (1) Now we will estimate the least-squares regression line for your data, using the R function lm (for linear model). Report the estimated slope and intercept for the line fit to your two variables.
β_0 = $37.9380977 billion
β_1 = 0.5246192

```
(Intercept)  YTD.Change
 37.9380977   0.5246192
> summary(my.lm)
Call:
lm(formula = Net.worth ~ YTD.Change, data = data.df)
Residuals:
    Min      1Q  Median      3Q     Max
-22.000 -16.474  -6.636   9.533  79.161
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.9381     3.8233   9.923 3.26e-13 ***
YTD.Change    0.5246     0.4194   1.251    0.217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 21.98 on 48 degrees of freedom
Multiple R-squared:  0.03157, Adjusted R-squared:  0.01139
F-statistic: 1.565 on 1 and 48 DF,  p-value: 0.217
```

C12 (3). Give an interpretation of your regression line. Include an interpretation of the slope value, and what you think it means for the relationship between your two quantitative variables.
Slope suggests that for every YTD Change ($Billions), Networth changes by 0.5246192($Billions). Regression line is not necessarily accurate since correlation value is weak (0.03<0.5). Values Suggest positive correlation between the two values.