TMATH 390 R Lab 10

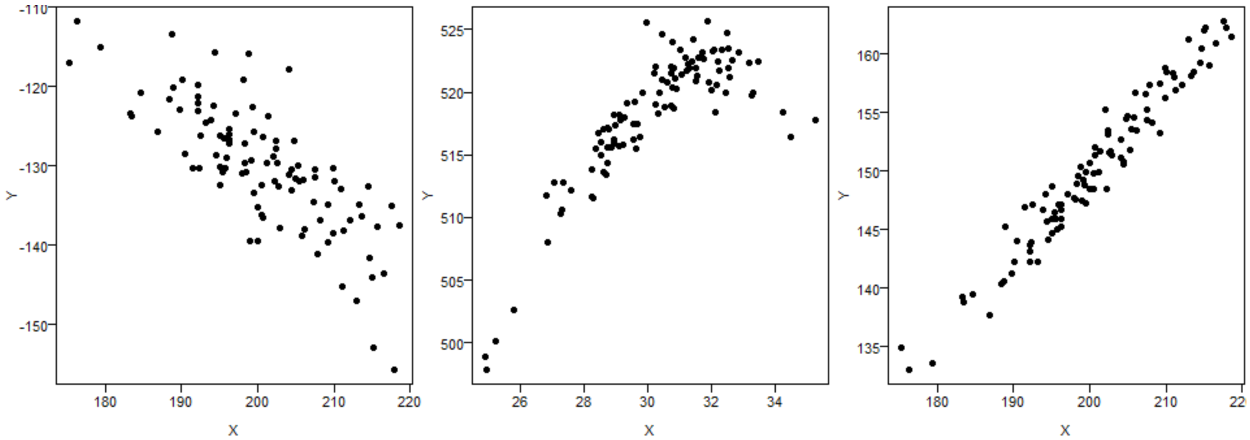
Correlation and linear regression

This lab is divided into 2 parts. For the first part you will answer questions regarding correlation and regression analysis. For the second part you will perform correlation and regression analysis for the dataset you obtained in HW 1, and that you evaluated in R Lab 3. Here we will concentrate on the two quantitative variables in your dataset.

C1. (4) Submit the R script you used to complete this computer lab assignment.

Part 1

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



In forensics it is often desirable to predict the stature of an unseen suspect from evidence (such as a foot print) left at a crime scene. (Rohren, B. 2006. Estimation of Stature from Foot and Shoe Length: Applications in Forensic Science, obtained from Triola Elementary Statistics). The code on the next page estimates the linear model between shoe print length (cm) and height (cm) for 40 individual adults.

```
# first we read in the data. Note you don't have to run this code, it's
# for demonstration purposes.
foot.df=read.csv("g:/My Drive/TeachingRScripts/TMATH410/HWRProj/FOOT.csv")
# In this dataframe we're interested in predicting the column named "Height"
# using the column named "Shoe.Print". First we can summarize each variable
summary(foot.df$Shoe.Print)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.80   26.70   28.50   29.02   31.40   34.50
```

```
summary(foot.df$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   152.4   166.2   175.3   174.3   180.8   195.0
```

```
# Here we use the R function lm (for linear model), which uses the same
#formula syntax as boxplot
foot.lm=lm(Height~Shoe.Print,data=foot.df)
# and now we can access the estimated coefficients
summary(foot.lm)
```

```
##
## Call:
## lm(formula = Height ~ Shoe.Print, data = foot.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8911  -4.9409   0.3969   3.1982  15.1181
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.930     10.893   7.429 6.50e-09 ***
## Shoe.Print     3.219      0.374   8.606 1.86e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.944 on 38 degrees of freedom
## Multiple R-squared:  0.6609, Adjusted R-squared:  0.652
## F-statistic: 74.06 on 1 and 38 DF,  p-value: 1.863e-10
```

C3. (1) In your lab document identify the response and predictor variable for this linear model.

C4 (1) In your lab document identify the estimated intercept and the estimate slope values.

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}

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.

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.

```
# first set up the plotting window. mfrow creates a 1x1 window
# mar sets the margin sizes (bottom, left, top, right)
# mgp sets the axis (scale, ticks, line)
# las orients the axis scale (here we set it to horizontal)
par(mfrow=c(1,1),mar=c(3.75,3.75,0.5,0.5),mgp=c(2.75,0.5,0),las=1)
# now we create the scatterplot. xlab gives the label for the x-axis,
# ylab the label for the y-axis. Make sure to update these with your own
# variables. pch gives the point type (16=solid points)
# see ?points for options. col gives the color for the points (?colors)
plot(data.df$X,data.df$Y,xlab="Predictor variable name (units)",
     ylab="Response variable name (units)",pch=16,col="blue")
```

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

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

```
# using the cor function in R to calculate the correlation between X and Y
# substitute the name of your dataframe and the appropriate column names
cor(data.df$X,data.df$Y)
```

C10. (1) Interpret the value for your correlation coefficient. Does this indicate there might be a strong linear relationship between Y and X?

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.

```
# lm uses the same formula syntax that boxplot does.
# this code estimates the linear relationship between Y and X
# for the dataframe data.df. Replace with the column names for your
# Y and X variables, and your own dataset.
my.lm=lm(Y~X,data=data.df)
# the my.lm object contains a description of the linear model,
# including the estimated coefficients. To retrieve the coefficients:
my.lm$coefficients
summary(my.lm)
```

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.