---
title: 'IMT 573: Problem Set 5 - Learning from Data'
author: "Neelam Purswani"
date: 'Due: Tuesday, November 6, 2018'
output: pdf_document

<!-- This syntax can be used to add comments that are ignored during knitting process. -->

Collaborators: Akshay Khanna

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset5.Rmd` file from Canvas. Open `problemset5.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset5.Rmd`.

2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.

3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. If you are using more than just a standard function that you found from another source, please credit the source in the comments. For example:

```{r citing, include=FALSE}
# code adapted from "Example: Multiplication Table"
# https://www.datamentor.io/r-programming/examples/multiplication-table/


# assign num
num = 8
# use for loop to iterate 10 times
for(i in 1:10) {
print(paste(num,'x', i, '=', num*i))
}
```

4. Collaboration on problem sets is acceptable, and even encouraged, but students must turn in an individual write-up in their own words and their own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF` or `Knit Word`, rename the R Markdown file to `YourLastName_YourFirstName_ps5.Rmd`, knit a PDF or DOC and submit both the PDF/DOC and the Rmd file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages. If you have not already installed them, do so in your console before loading the library here.

```{r Setup, message=FALSE, warning=FALSE}
# Load standard libraries
library(tidyverse)
library(Sleuth3) # Contains data for problemset
library(UsingR) # Contains data for problemset
library(MASS) # Modern applied statistics functions
```

Problem 1:
Davis et al. (1998) collected data on the proportion of births that were male in Denmark, the Netherlands, Canada, and the United States for selected years. Davis et al. argue that the proportion of male births is declining in these countries. We will explore this hypothesis. You can obtain this data as follows:

```{r}
# Import male births data
malebirths <- Sleuth3::ex0724
malebirths
```

```{r}
str(malebirths)
```

1a.
Use the \texttt{lm} function in \textbf{R} to fit four (one per country) simple linear regression models of the yearly proportion of males births as a function of the year and obtain the least squares fits. Write down the estimated linear model for each country.

```{r}
USA_MalePopulation <- lm(USA ~ Year, malebirths)
summary(USA_MalePopulation)
```
The equation for linear regression looks like : y^i=β^0+β^1xi+ε^i
β^0 : is the Estimate value in the (Intercept) row (specifically, 6.201e-01)
β^1 : is the Estimate value in the x row (specifically, -5.429e-05)

```{r}
Canada_MalePopulation <- lm(Canada ~ Year, malebirths)
summary(Canada_MalePopulation)
```
The equation for linear regression looks like : y^i=β^0+β^1xi+ε^i
β^0 : is the Estimate value in the (Intercept) row (specifically, 7.338e-01)
β^1 : is the Estimate value in the x row (specifically, -1.112e-04)

```{r}
Netherlands_MalePopulation <- lm(Netherlands ~ Year, malebirths)
summary(Netherlands_MalePopulation)
```
The equation for linear regression looks like : y^i=β^0+β^1xi+ε^i
β^0 : is the Estimate value in the (Intercept) row (specifically, 6.724e-01)
β^1 : is the Estimate value in the x row (specifically, -8.084e-05)

```{r}
Denmark_MalePopulation <- lm(Denmark ~ Year, malebirths)
summary(Denmark_MalePopulation)
```
The equation for linear regression looks like : y^i=β^0+β^1xi+ε^i
β^0 : is the Estimate value in the (Intercept) row (specifically, 5.987e-01)
β^1 : is the Estimate value in the x row (specifically, -4.289e-05)

1b.
Obtain the t-statistic for the test that the slopes of the regression lines are zero, for each of the four countries. Is there evidence that the proportion of births that are male is truly declining over this period?
t-values for -
USA : -5.779
Canada : -4.017
Netherlands : -5.71
Denmark : -2.073
Since p-value is less than 0.05 for all the countries, we can reject the null hypothesis and accept the alternative hypothesis that there is truly a decline in proportion of births.

Problem 2:

Regression was originally used by Francis Galton to study the relationship between parents and children. One relationship he considered was height. Can we predict a man's height based on the height of his father? This is the question we will explore in this problem. You can obtain data similar to that used by Galton as follows:

```{r}
# Import and look at the height data
heightData <- tbl_df(get("father.son"))
heightData
```

```

### 2a.
Perform an exploratory analysis of the dataset. Describe what you find. At a minimum you should produce statistical summaries of the variables, a visualization of the relationship of interest in this problem, and a statistical summary of that relationship.
```{r}
summary(heightData)
```
```{r}
p <-ggplot(data = heightData ) + geom_point(mapping = aes(x = fheight , y = sheight, color=fheight))
p+labs(x="father's height", y="Son's height")
```
The scatter plot shows that son's height is directly proportional to father's height.

2b.
Use the \texttt{lm} function in R to fit a simple linear regression model to predict son's height as a function of father's height. Write down the model, $$\hat{y}_{\mbox{\texttt{sheight}}} = \hat{\beta}_0 + \hat{\beta_i} \times \mbox{\texttt{fheight}}$$ filling in estimated coefficient values and interpret the coefficient estimates.
```{r}
linearmodelfs <- lm(sheight ~ fheight, data=heightData)
summary(linearmodelfs)
```

y=mx+c
son's height = m*father's height + y-intercept

sheight= 0.51409*fheight + 33.88660

2c.
Find the 95\% confidence intervals for the estimates. You may find the \texttt{confint()} command useful.
```{r}
confint(linearmodelfs)
```


2d.
Produce a visualization of the data and the least squares regression line.
```{r}
plot(heightData)
abline(linearmodelfs$coefficients)
```
```{r}
p <-ggplot(data = heightData ) + geom_smooth(mapping = aes(x = fheight , y = sheight, color=fheight))
p+labs(x="father's height", y="Son's height")
```

2e.
Produce a visualization of the residuals versus the fitted values. (You can inspect the elements of the linear model object in R using \texttt{names()}). Discuss what you see. Do you have any concerns about the linear model?
```{r}
plot(linearmodelfs, which=c(1,1))
```


2f.
Using the model you fit in part (b) predict the height was 5 males whose father are 50, 55, 70, 75, and 90 inches respectively. You may find the \texttt{predict()} function helpful.
```{r}
df.father <- data.frame(fheight= c(50,55,70,75,90))
predict(linearmodelfs,df.father)
```


Extra Credit:

EC(a).
What assumptions are made about the distribution of the explanatory variable in the normal simple linear regression model?
1.For each value of X, the distribution of possible Y values is normal.
2.The normal distribution of Y values corresponding to a particular value of X has standard deviation
σ{Y |X}. That standard deviation is usually assumed to be the same for all values of X so that we may
write σ{Y |X} = σ.
*For points 1 and 2,Referred http://www.public.iastate.edu/~dnett/S401/wreginf.pdf

EC(b).
Why can an R^2 close to one not be used as evidence that the simple linear regression model is appropriate?
Although higher value of R shows that a model is perfect fit, the reason it cannot be used is that it can be increased artificially by adding more independent variables in the dataset.

EC(c).
Consider a regression of weight on height for a sample of adult males. Suppose the intercept is 5 kg. Does this imply that males of height 0 weigh 5 kg, on average? Would this imply that the simple linear regression model is meaningless?
Yes, this sounds absurd because the height of the person cannot be 0

EC(d).
Suppose you had data on pairs (X,Y) which gave the scatterplot been below. How would you approach the analysis?


```{r pressure, echo=FALSE, fig.cap="Scatterplot", out.width = '100%'}
knitr::include_graphics("scatterplot.png")
```
Looking at the scatterplot, we can find correlation between the explanatory and response variables.
For example, over here there is a positive correlation at least. Although correlation does not necessarily imply causation.
Then maybe I will see if I can fit a line through the data so that I have very few outliers.
Then, i will take into consideration the variables into consideration and try to use some background knowledge to make inferences/assumptions.
```