Tae Coding
Introduction to Data Science: CS61
Summer 2018
Class Exercise#5

Date Given: June 26, 2018 Due Date:

==========

Problem Number	Answer
1	Answers: Regression Equation: $y = 53.5x - 1354.5$
2	Answers: a. 2.5134 b. 1 additional hour of video, GPA drops by 0.0526, c. No Video GPA = 2.9342 d. Above the average
3	Answers: a. Regression Eq: $Head - Circumference = 0.1827 * Height + 12.4932$ b. If height increases by1, HC increases by 0.1827 c. 17.06 inch d. Residual = -0.16 e. HC can vary f. No
4	Answers: a. MPG = -0.0070*Weight + 44.8793 b. If weight increases by 1 pound, MPG decreases by 0.0070 c. Below average d. No

Python Code

```
Problem #1

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.cross_validation import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

#######################################
# 1. Read Data File

Xlist = [25, 30, 35, 40, 45]

Ylist = [5, 260, 480, 745, 1100]
#######################################
Xarray = np.array(Xlist)
Yarray = np.array(Ylist)
XYarray = Xarray * Yarray
XYarray
Out[16]: array([  125,  7800, 16800, 29800, 49500])

X2array = Xarray**2
X2array
Out[18]: array([ 625,  900, 1225, 1600, 2025], dtype=int32)

################################################
meanX = np.mean(Xarray)
print(meanX)
35.0

meanY = np.mean(Yarray)
print(meanY)
518.0

meanXY = np.mean(XYarray)
print(meanXY)
20805.0

meanX2 = np.mean(X2array)
print(meanX2)
1275.0

stdX = np.std(Xarray)
print(stdX)
7.07106781187
```

```
stdY = np.std(Yarray)
print(stdY)
379.849970383

r = np.corrcoef(Xarray, Yarray)
r[0][1]
Out[33]: 0.99592512157712954

########################################
# Problem 1a : Closed form solution - using only the mean of X, Y,
XY, X^2

slope = (meanXY - (meanX*meanY))/(meanX2 - meanX*meanX)
print(slope)
53.5

intercept = meanY - slope*meanX
print(intercept)
-1354.5
########################################
# Problem 1b: Using Correlation and Standard Deviation
slope = r[0][1]*stdY/stdX
print(slope)
53.5

####################################
# Problem 1c: Regression Using Scikit-Learn
df_x = pd.DataFrame(Xlist)
df_y = pd.DataFrame(Ylist)

reg = linear_model.LinearRegression()
reg.fit(df_x,df_y)
Out[51]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1,
normalize=False)

print(reg.coef_)
[[ 53.5]]

print(reg.intercept_)
[-1354.5]

########################################
```

```
##############################################
# Problem #2
# Video Games and GPA
# y (GPA) = -0.0526 * x (Hours Video Games) + 2.9342

slope = -0.0526
intercept = 2.9342
##############################################
# 2 a
hoursVideoGames = 8
GPA = hoursVideoGames * slope + intercept

print(GPA)
2.5134000000000003

##############################################
# 2 b
# Every additional hour of video game played, decreases the GPA by
0.0526
#
##############################################
# 2 c
#
# When the value of 'x' (number of hours video games played) is zero,

# GPA = 2.9342
##############################################
#  2 d
hoursVideoGames = 7
GPA = hoursVideoGames * slope + intercept

print(GPA)
2.5660000000000003

# Since 2.68 > 2.566

# Above the average
```

```
Problem #3

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import linear_model


C:\ProgramData\Anaconda3\lib\site-
packages\sklearn\cross_validation.py:41: DeprecationWarning: This
module was deprecated in version 0.18 in favor of the model_selection
module into which all the refactored classes and functions are moved.
Also note that the interface of the new CV iterators are different
from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)

from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

#######################################
# 1. Read Data File

Xlist = [27.75, 24.5, 25.5, 26, 25, 27.75, 26.5, 27, 26.75, 26.75,
27.5]

Ylist = [17.5, 17.1, 17.1, 17.3, 16.9, 17.6, 17.3, 17.5, 17.3, 17.5,
17.5]

#######################################

# Problem 3a: Regression Using Scikit-Learn
df_x = pd.DataFrame(Xlist)
df_y = pd.DataFrame(Ylist)
reg = linear_model.LinearRegression()

reg.fit(df_x,df_y)
Out[19]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1,
normalize=False)

print(reg.coef_)
[[ 0.18273245]]

print(reg.intercept_)
[ 12.49316888]
```

```
##########################################
# 3 b
# slope: if height increases by 1 inch, head-circumference increases
by 0.1827 inches

#

# intercept:  If height is zero, head circumference = 12.49.  This is
absurd.

# Therefore, interpretation is outside the scope of the model

##########################################

# 3 d

height = 25

head_cir = height * reg.coef_ + reg.intercept_

residual = head_cir - 16.9

print(residual)
[[ 0.16148008]]
```

```
Problem #4

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import linear_model

C:\ProgramData\Anaconda3\lib\site-
packages\sklearn\cross_validation.py:41: DeprecationWarning: This
module was deprecated in version 0.18 in favor of the model_selection
module into which all the refactored classes and functions are moved.
Also note that the interface of the new CV iterators are different
from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)

# 1. Read Data File

Xlist = [3765, 3984, 3530, 3175, 2580, 3730, 2605, 3772, 3310, 2991,
2752]
Ylist = [19, 18, 21, 22, 27, 18, 26, 17, 20, 25, 26]

#####################################
# Problem 4a: Regression Using Scikit-Learn

df_x = pd.DataFrame(Xlist)
df_y = pd.DataFrame(Ylist)

reg = linear_model.LinearRegression()
reg.fit(df_x,df_y)
Out[19]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1,
normalize=False)
print(reg.coef_)
[[-0.00703632]]
print(reg.intercept_)
[ 44.87932977]

#####################################

# Problem 4b

# slope: For every pound added to the weight of the car will reduce
the gas mileage by 0.007 MPG

# intercept: Interpretation of intercept is outside the scope of the
model
```

```
########################################

# Problem 4c

weight = 2780

MPG = weight * reg.coef_ + reg.intercept_

print(MPG)
[[ 25.31835546]]

# Since 22 < 25.318.  Below average

########################################

# 4 d

# No. This data is for internal combustion engines only.

# Toyota Prius is a hybrid car

#
```

R Code

Problem#1

Linear Regression

The values of 2 variables X and Y are given below. Here X is the predictor variable and Y is the response variable.

X	25	30	35	40	45
Y	5	260	480	745	1100

Build your regression model using the following 3 methods.

a) Closed form solution – using only the mean of 'x', 'y', 'x*y', 'x^2' variables.
b) Closed form solution – using the correlation coefficient between 'x' and 'y' variables and the standard deviation of both variables.
c) R Regression function

Make sure that your answers are the same using all the 3 methods.

Problem#1a

	A	B	C	D	E	F
1						
2		X	Y		X*Y	X^2
3		25	5		125	625
4		30	260		7800	900
5		35	480		16800	1225
6		40	745		29800	1600
7		45	1100		49500	2025
8						
9	SUM	175	2590		104025	6375
10	Average	35	518		20805	1275
11	StdDev	7.905694	424.6852			
12	Correlation	0.995925				

Problem#1a

$$\frac{\frac{\sum y_i x_i}{N} - \frac{\sum y_i \sum x_i}{N.N}}{\frac{\sum x_i^2}{N} - \frac{\sum x_i \sum x_i}{N.N}} = \frac{Mean\ of\ X*Y - (Mean\ of\ X)*(Mean\ of\ Y)}{Mean\ of\ x^2 - (Mean\ of\ X)*(Mean\ of\ X)}$$

$$b = \left(\frac{\sum y_i}{N} - m\frac{\sum x_i}{N}\right)$$

Method#1	
Numerator	2675
Denominator	50
Slope	53.5
Intercept	-1354.5

Regression Equation: $y = 53.5x - 1354.5$

Problem#1b

- $m = r\frac{\sigma_y}{\sigma_x} = Correlation\frac{Std\ Dev\ of\ y}{Std\ Dev\ of\ x}$

Method#2	
Slope	53.5
Intercept	-1354.5

Problem#1c

```
> ######################################################
> # Introduction to Statistics Using R
> #
> ######################################################
> # Homework#6
> # Problem#1
> ######################################################
> x = c(25,30,35,40,45)
> y = c(5,260,480,745,1100)
>
> result = lm(y~x)
> summary(result)

Call:
lm(formula = y ~ x)

Residuals:
    1     2     3     4     5
 22.0   9.5 -38.0 -40.5  47.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1354.500     99.874  -13.56 0.000867 ***
x              53.500      2.797   19.13 0.000312 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.22 on 3 degrees of freedom
Multiple R-squared:  0.9919,   Adjusted R-squared:  0.9892
F-statistic: 365.9 on 1 and 3 DF,  p-value: 0.0003121



>
```

Problem#2

17. You Explain It! Video Games and GPAs NW A student at Joliet Junior College conducted a survey of 20 randomly selected full-time students to determine the relation between the number of hours of video game playing each week, x, and grade-point average, y. She found that a linear relation exists between the two variables. The least-squares regression line that describes this relation is $\hat{y} = -0.0526x + 2.9342$.

(a) Predict the grade-point average of a student who plays video games 8 hours per week.

(b) Interpret the slope.

(c) If appropriate, interpret the y-intercept.

(d) A student who plays video games 7 hours per week has a grade-point average of 2.68. Is this student's grade-point average above or below average among all students who play video games 7 hours per week?

```
> ######################################################
> # Introduction to Statistics Using R
> #
> ######################################################
> # Homework#6
> # Problem#2
> ######################################################
> # a
> x = 8
> (y = -0.0526*x + 2.9342)
[1] 2.5134
>
> #b
> # For additional hour of video game played
> # GPA drops by 0.0526
>
> #c
> #Y-intercept: If number of hours played = 0
> # GPA = 2.9342
>
>
> #d
> x = 7
> (y = -0.0526*x + 2.9342)
[1] 2.566
>
> # Computed GPA = 2.566
> # If GPA = 2.68, above the average


>
```

Problem#3

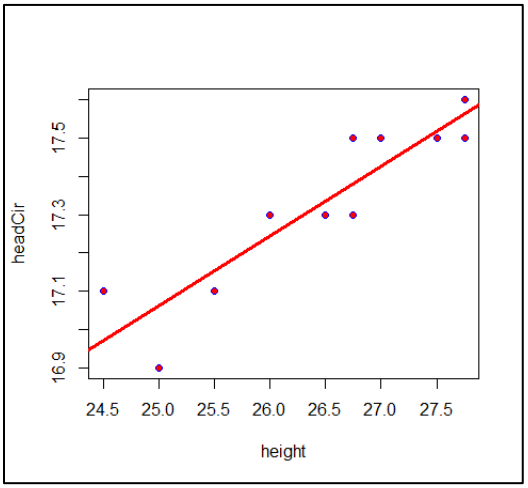
21. Height versus Head Circumference (Refer to Problem 23, Section 4.1) A pediatrician wants to determine the relation that exists between a child's height, x, and head circumference, y. She randomly selects 11 children from her practice, measures their heights and head circumferences and obtains the following data.

Height, x (inches)	Head Circumference, y (inches)	Height, x (inches)	Head Circumference, y (inches)
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College

(a) Find the least-squares regression line treating height as the explanatory variable and head circumference as the response variable.
(b) Interpret the slope and y-intercept, if appropriate.
(c) Use the regression equation to predict the head circumference of a child who is 25 inches tall.
(d) Compute the residual based on the observed head circumference of the 25-inch-tall child in the table. Is the head circumference of this child above average or below average?
(e) Draw the least-squares regression line on the scatter diagram of the data and label the residual from part (d).
(f) Notice that two children are 26.75 inches tall. One has a head circumference of 17.3 inches; the other has a head circumference of 17.5 inches. How can this be?
(g) Would it be reasonable to use the least-squares regression line to predict the head circumference of a child who was 32 inches tall? Why?

```
> ####################################################
> # Introduction to Statistics Using R
> #
> ####################################################
> # Homework#6
> # Problem#3
> ####################################################
> # a
> height = c(27.75, 24.5, 25.5, 26, 25, 27.75, 26.5, 27, 26.75, 26.75, 27.5)
> headCir = c(17.5, 17.1, 17.1, 17.3, 16.9, 17.6, 17.3, 17.5, 17.3, 17.5, 17.5)
>
> plot(height,headCir,pch=21,col="blue",bg="red")
> result = lm(headCir~height)
> summary(result)

Call:
lm(formula = headCir ~ height)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16148 -0.05842 -0.01831  0.06442  0.12989

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.49317    0.72968   17.12 3.56e-08 ***
height       0.18273    0.02756    6.63 9.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09538 on 9 degrees of freedom
Multiple R-squared:  0.8301,   Adjusted R-squared:  0.8112
F-statistic: 43.96 on 1 and 9 DF,  p-value: 9.59e-05

> abline(result,lwd=3,col="red")
>
>
> #


>
```



a. Regression Equation: $Head - Circumference = 0.1827 * Height + 12.4932$

b. Slope: If height increases by 1 inch, Head-Circumference increases by 0.1827 inches

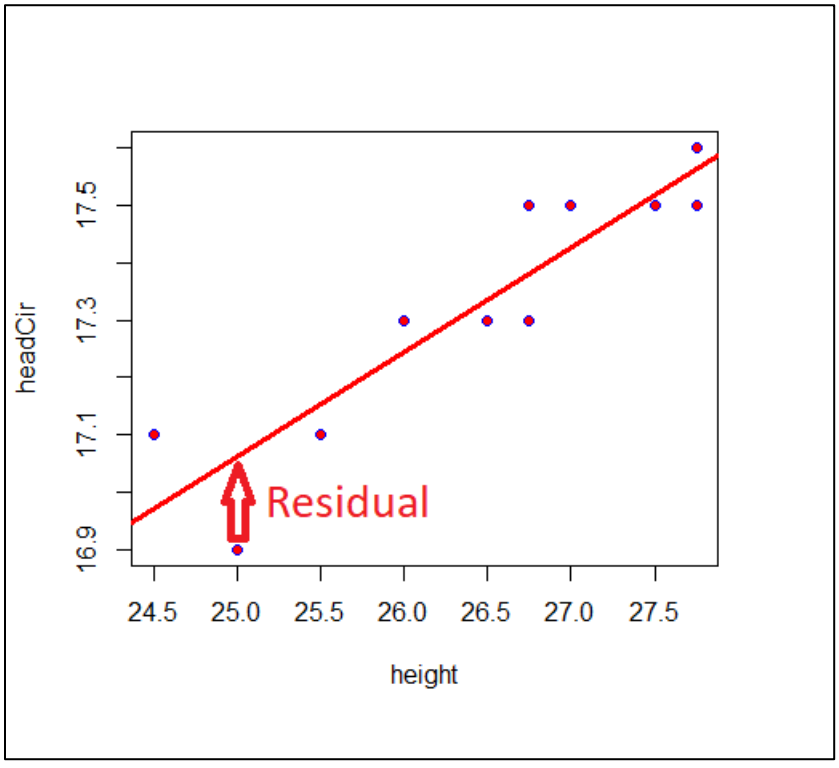
Intercept: If Height is 0, Head-circumference = 12.49. This is absurd. Therefore, interpretation of intercept is outside the scope of the model.

c.

```
> ##############################################
> #c
> h.25 = 25
>
> (headCir.25 = result[[1]][2]*h.25 + result[[1]][1])
  height
17.06148
>
> #


>
```

d. Residual = 16.9 – 17.06 = -0.16: Below average

e.



f. The head-circumference of 2 children for the same height can vary.

g. No. A height of 32 inch would be outside the scope of the model.

Problem#4

23. Weight of a Car versus Miles per Gallon (Refer to Problem 25, Section 4.1) An engineer wants to determine how the weight of a car, x, affects gas mileage, y. The following data represent the weights of various domestic cars and their miles per gallon in the city for the 2008 model year.

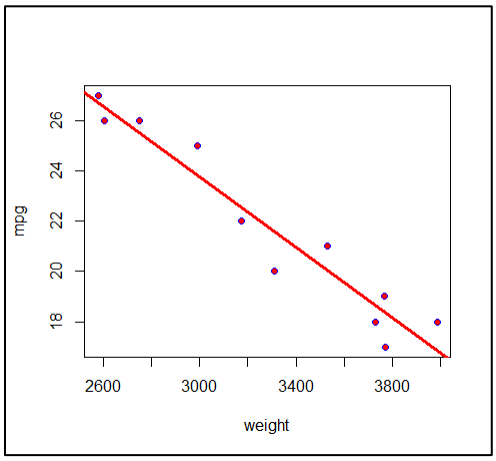
Car	Weight (pounds), x	Miles per Gallon, y
Buick Lucerne	3,765	19
Cadillac DeVille	3,984	18
Chevrolet Malibu	3,530	21
Chrysler Sebring Sedan	3,175	22
Dodge Neon	2,580	27
Dodge Charger	3,730	18
Ford Focus	2,605	26
Lincoln LS	3,772	17
Mercury Sable	3,310	20
Pontiac G5	2,991	25
Saturn Ion	2,752	26

Source: www.roadandtrack.com

(a) Find the least-squares regression line treating weight as the explanatory variable and miles per gallon as the response variable.
(b) Interpret the slope and y-intercept, if appropriate.
(c) A Chevy Cobalt weighs 2,780 pounds and gets 22 miles per gallon. Is the miles per gallon of a Cobalt above average or below average for cars of this weight?
(d) Would it be reasonable to use the least-squares regression line to predict the miles per gallon of a Toyota Prius, a hybrid gas and electric car? Why or why not?

a. Regression

```
> ####################################################
> # Introduction to Statistics Using R
> #
> ####################################################
> # Homework#6
> # Problem#4
> ####################################################
> # a
> weight = c(3765, 3984, 3530, 3175, 2580, 3730, 2605, 3772, 3310, 2991, 2752)
> mpg = c(19, 18, 21, 22, 27, 18, 26, 17, 20, 25, 26)
>
> plot(weight,mpg,pch=21,col="blue",bg="red")
> result = lm(mpg~weight)
> summary(result)

Call:
lm(formula = mpg ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5891 -0.5918  0.2744  0.7856  1.1663

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.8793298  2.1487116   20.89 6.19e-09 ***
weight      -0.0070363  0.0006461  -10.89 1.75e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.033 on 9 degrees of freedom
Multiple R-squared:  0.9295,   Adjusted R-squared:  0.9216
F-statistic: 118.6 on 1 and 9 DF,  p-value: 1.752e-06

> abline(result,lwd=3,col="red")


>
```



$MPG = -0.0070 * Weight + 44.8793$

b. Slope: For every pound added to the weight of the car will reduce the gas mileage by 0.0070 MPG

Interpretation of slope is outside the scope of the model

c. If weight = 2780, MPG = -0.0070*2780+44.8793 = 25.32

```
> ##############################################
> #c
> w.2780 = 2780
>
> (mpg.2780 = result[[1]][2]*w.2780 + result[[1]][1])
  weight
25.31836


>
```

Chevy Cobalt getting 22 MPG is below average

d. No. This data is only for internal combustion engines only. Toyota Prius is a hybrid car.