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: |
| 2 | Answers:   1. 2.5134 2. 1 additional hour of video, GPA drops by 0.0526, 3. No Video GPA = 2.9342 4. Above the average |
| 3 | Answers:   1. Regression Eq: 2. If height increases by1, HC increases by 0.1827 3. 17.06 inch 4. Residual = -0.16 5. HC can vary 6. No |
| 4 | Answers:   1. MPG = -0.0070\*Weight + 44.8793 2. If weight increases by 1 pound, MPG decreases by 0.0070 3. Below average 4. 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.

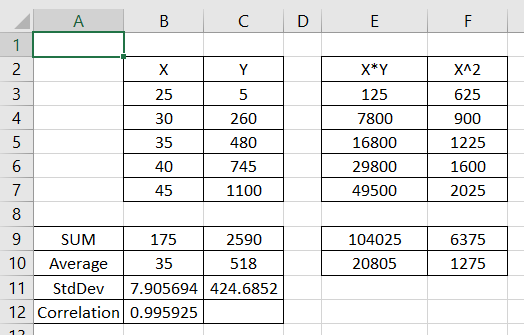
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | 25 | 30 | 35 | 40 | 45 |
| Y | 5 | 260 | 480 | 745 | 1100 |

Build your regression model using the following 3 methods.

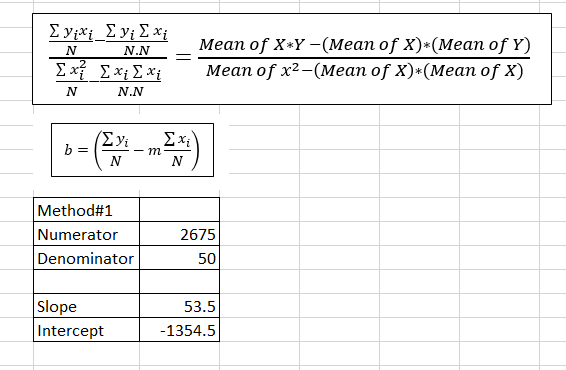
1. Closed form solution – using only the mean of ‘x’, ‘y’, ‘x\*y’, ‘x2’ variables.
2. Closed form solution – using the correlation coefficient between ‘x’ and ‘y’ variables and the standard deviation of both variables.
3. R Regression function

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

**Problem#1a**

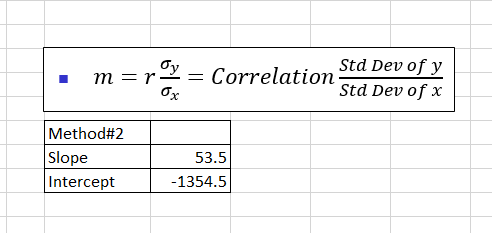


**Problem#1a**



Regression Equation:

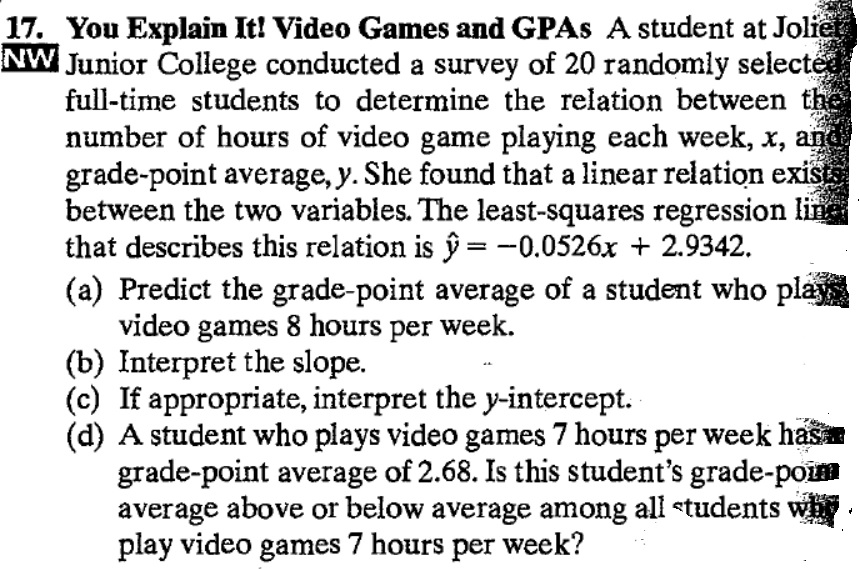
**Problem#1b**



**Problem#1c**

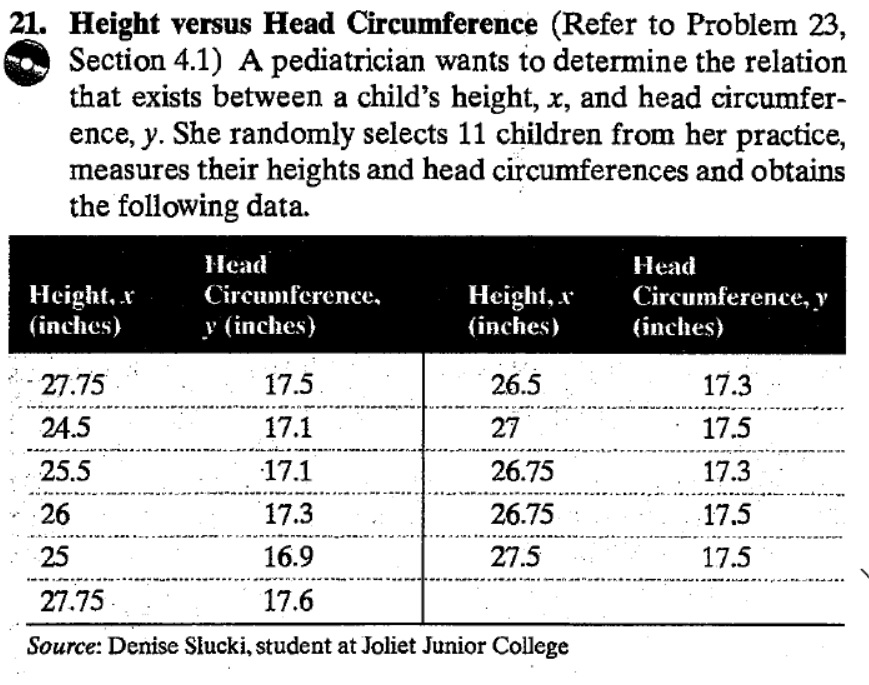
|  |
| --- |
| > #####################################################  > # Introduction to Statistics Using R  > #  > ####################################################  > # Homework#6  > # Problem#1  > ####################################################  > x = c(25,30,35,40,45)  > y = c(5,260,480,745,1100)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

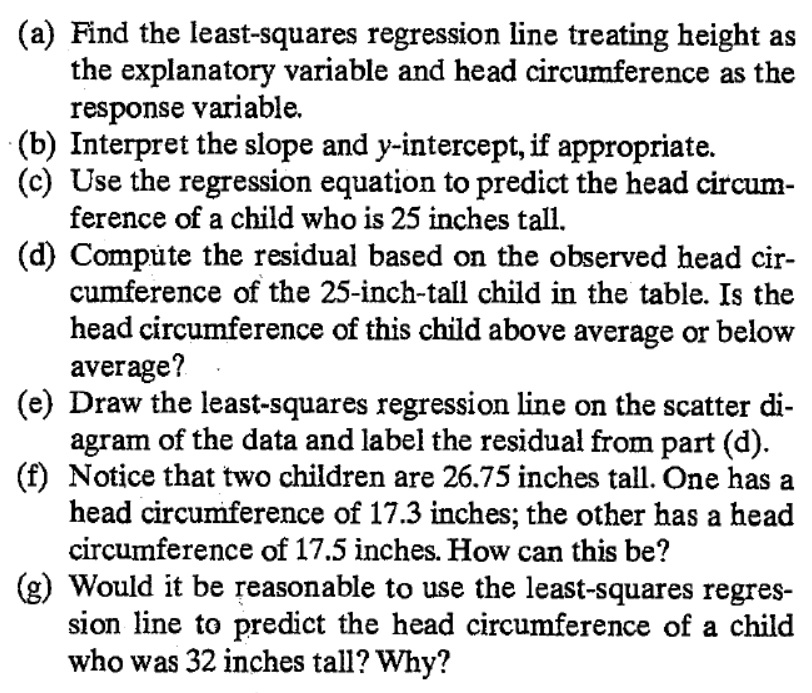
**Problem#2**



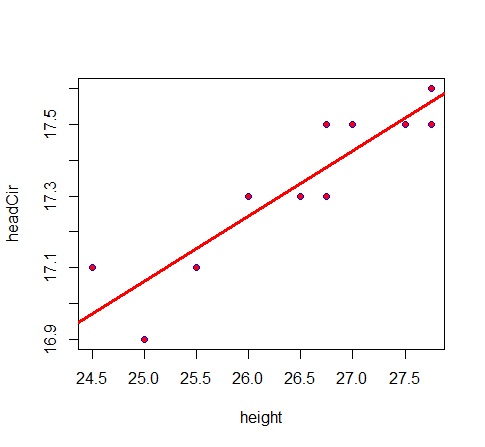
|  |
| --- |
| > #####################################################  > # Introduction to Statistics Using R  > #  > ####################################################  > # Homework#6  > # Problem#2  > ####################################################  > # a  > x = 8  > (y = -0.0526\*x + 2.9342)  [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**





|  |
| --- |
| > #####################################################  > # 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")  >  >  > # |
|  |
| |  | | --- | | > | |

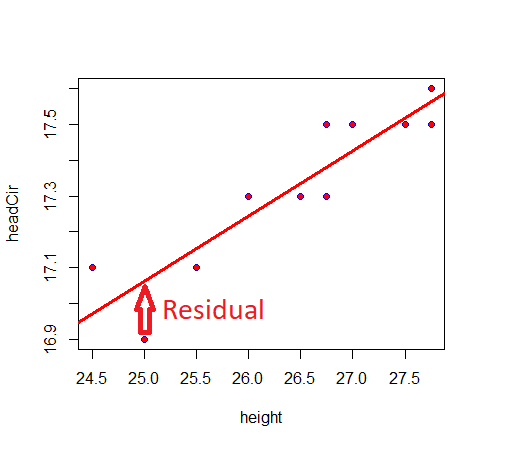


1. Regression Equation:
2. Slope: If height increases by 1 inch, Head-Circumference increases by 0.1827 inches

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

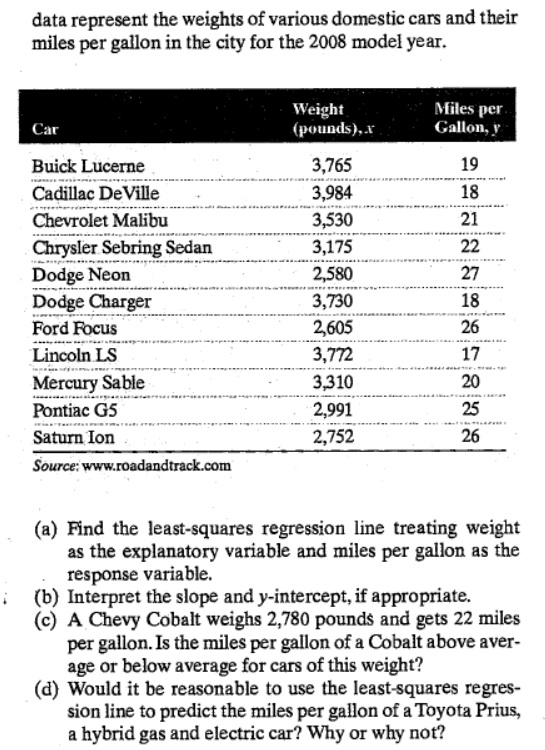
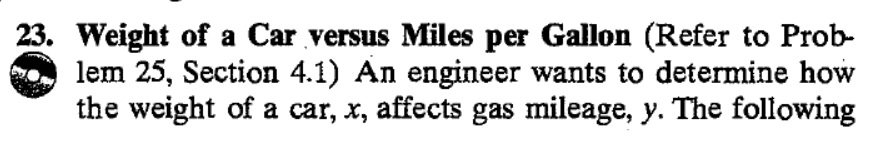
|  |
| --- |
| > ##############################################  > #c  > h.25 = 25  >  > (headCir.25 = result[[1]][2]\*h.25 + result[[1]][1])  height  17.06148  >  > # |
|  |
| |  | | --- | | > | |

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



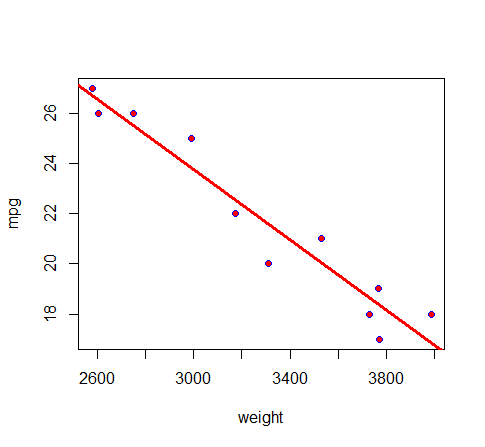
1. The head-circumference of 2 children for the same height can vary.
2. No. A height of 32 inch would be outside the scope of the model.

**Problem#4**



1. 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") |
|  |
| |  | | --- | | > | |



1. 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

1. 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

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