**데이터마이닝이론**

**S T A 6 6 0 0**

**김현중 교수님**

**Homework 2**

**2018314030**

**응용통계학과 김단아**

**1. Assume the location of the response variable can be different for each dataset. Modify your R program in Assignment #1 to do followings.**

a. Prompt the user to ask in which column the response variable is recorded.

|  |
| --- |
| **(R code)**  # enter the Column number  cat("Enter which column the response variable is recorded: ")  num = scan(n=1, quiet=TRUE)  # design matrix  n = dim(data)[1]  p = dim(data)[2]-1  one = matrix(1, nrow=n, ncol=1)  I = diag(n)    y = as.matrix(data[,**num**])  x = as.matrix(cbind(one,data[,-**num**]))  H = x%\*%solve(t(x)%\*%x)%\*%t(x)  H0 = one%\*%solve(t(one)%\*%one)%\*%t(one) |

* Prompt 창에 유저가 직접 반응변수의 열을 지정해주도록 하는 함수를 Assingment1에서 만든 회귀분석 함수에 추가해주었다.

b. Make your R program (in Assignment #1) to run with the column number entered in (a) as the response variable.

|  |
| --- |
| **(R code)**  regression.analysis()  harris.dat  1  1  HW2KimDA\_R\_output |

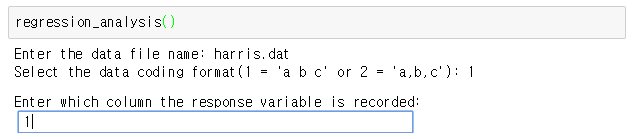
|  |
| --- |
| **(HW2KimDA\_R\_output.txt)**  Coefficients  -------------  Constant: 3526.422  Beta1: 90.02  Beta2: 1.269  Beta3: 23.406  Beta4: 722.461  ID, Actual values, Fitted values  --------------------------------  1, 3900, 4630.1  2, 4020, 4646.3  (….. 중략 …..)  92, 6900, 6328.4  93, 8100, 6530.8  Model Summary  -------------  R-square = 0.5109  MSE = 254583.5835 |

* Harris.dat을 이용하여 Regression.analysis()를 실행한 결과 성공적으로 회귀분석을 진행한 것을 확인하였다.

**2. Write a program to implement the multiple linear regression analysis. Use the ‘Python’ without direct command for regression such as ‘sklearn.linear\_model.LinearRegression’.**

a. Prompt the user to ask in which column the response (class) variable is recorded.

|  |
| --- |
| **(Python 코드)**  # prompt user to enter the data information  data = input("Enter the data file name: ")  fm = input("Select the data coding format(1 = 'a b c' or 2 = 'a,b,c'): ")  num = int(input("Enter which column the response variable is recorded: "))-1  if fm=='1':  form = ' '  else:  form= ',' |



* 1번 열을 response variable로 유저가 직접 지정하는 창의 모습이다. 파이썬의 열은 0 부터 시작하기 때문에 유저가 입력한 숫자에서 1을 감하여 num 이라는 함수명에 저장해주었다.

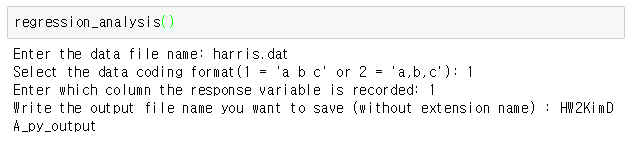
b. Make your Python program to run with the column number entered in (a).

|  |
| --- |
| **(Python 코드)**  # import data  data = pd.read\_csv(data, header=None, sep=form)  response = data[num]  explanatory = data.drop(num, axis=1) |

* 입력한 열의 위치를 response variable로, 나머지 열을 explanatory variable로 지정하였다.

c. Do questions in Assignment #1 using Python.

|  |
| --- |
| **(Python 코드)**  # set the working directory  import os as os  os.chdir("C:/Users/User/Desktop/18-2/DM/data")  print(os.getcwd())  def regression\_analysis():    # import packages  import numpy as np  import numpy.linalg as lin  import pandas as pd    # prompt user to enter the data information  data = input("Enter the data file name: ")  fm = input("Select the data coding format(1 = 'a b c' or 2 = 'a,b,c'): ")  num = int(input("Enter which column the response variable is recorded: "))-1  if fm=='1':  form = ' '  else:  form= ','  data = pd.read\_csv(data, header=None, sep=form)    # import data  response = data[num]  explanatory = data.drop(num, axis=1)    # design matrix  n = data.shape[0]  p = data.shape[1]-1  one = pd.DataFrame(np.ones((n,1)))  I = np.eye(n)  X = pd.concat([pd.DataFrame(one),explanatory], axis=1)  Y = response  H = X.dot(lin.inv(X.T.dot(X))).dot(X.T)  H0 = one.dot(lin.inv(one.T.dot(one))).dot(one.T)  inv = lin.inv(X.T.dot(X))  b = inv.dot(X.T).dot(Y).round(3)  # multiple regression result  yhat = H.dot(Y).round(1)  SST = Y.T.dot(I-H0).dot(Y)  SSE = Y.T.dot(I-H).dot(Y)  Rsquare = round(1 - SSE/SST, 4)  MSE = round(SSE/(n-p), 4)  # output file name  outputname = input("Write the output file name you want to save (without extension name) : ")  outputname = outputname+'.txt'  # outport the result  with open(outputname,"w") as text\_file:    print("Coefficients", file=text\_file)  print("-------------", file=text\_file)  for i in range(p+1):  if i==0:  print("Constant:", b[i],sep=" ", file=text\_file)  else:  print("Beta",i,": ",b[i],sep="", file=text\_file)  print("",file=text\_file)  print("ID, Actual values, Fitted values", file=text\_file)  print("--------------------------------", file=text\_file)  for i in range(n):  print(i+1, Y[i], yhat[i], sep=", ", file=text\_file)  print("",file=text\_file)    print("Model Summary", file=text\_file)  print("-------------", file=text\_file)  print("R-square = ", Rsquare, sep="", file=text\_file)  print("MSE = ", MSE, sep="", file=text\_file) |



|  |
| --- |
| **(HW2KimDA\_py\_output.txt)**  Coefficients  -------------  Constant: 3526.422  Beta1: 90.02  Beta2: 1.269  Beta3: 23.406  Beta4: 722.461  ID, Actual values, Fitted values  --------------------------------  1, 3900, 4630.1  2, 4020, 4646.3  (….. 중략 …..)  92, 6900, 6328.4  93, 8100, 6530.8  Model Summary  -------------  R-square = 0.5109  MSE = 254583.5835 |

* Python으로 만든 Regression.Analysis 함수를 harris.dat로 실행해준 결과, Assignment #1과 동일한 결과를 얻었다.