**데이터마이닝이론**

**S T A 6 6 0 0**

**김현중 교수님**

**Homework 1**

2018314030

응용통계학과 김단아

**1. Write programs to implement the multiple linear regression analysis. Do not use an R command such as ‘lm’ or ‘glm’.**

a. Make the program to accept file names for data and output from the user. Give the user a prompt to type-in the data file and output file names. (Hint: use ‘readline’ command in R)

|  |
| --- |
| # import data file  data = readline("Enter the data file name (with extension name) : " )  cat("Select the data coding format(1 = 'a b c' or 2 = 'a,b,c'): ")  fm = scan(n=1, quiet=TRUE)  if(fm==1) {form = ""} else {form = ","}  data = read.table(data, sep=form)  # make output file  outputname = readline("Write the output file name you want to save (without extension name): ")  outputname = paste(outputname,".txt",sep="") |

* 위 R 코드는 multiple linear regression analysis 를 수행하는 함수의 일부이다. 불러들일 data file과 저장할 file names을 유저가 직접 치는 프롬프트 창을 이용하는 코드이다. Assignment의 예시를 사용하였다.

b. The program must print out the coefficient for each X variable and save it into

an output file.

|  |
| --- |
| # 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[,1])  x = as.matrix(cbind(one,data[,-1]))  H = x%\*%solve(t(x)%\*%x)%\*%t(x)  H0 = one%\*%solve(t(one)%\*%one)%\*%t(one)  # multiple regression result  b = round(solve(t(x)%\*%x)%\*%t(x)%\*%y, 4)  yhat = x%\*%b  SST = t(y)%\*%(I-H0)%\*%y  SSE = t(y)%\*%(I-H)%\*%y  Rsquare = round(1 - SSE/SST, 4)  MSE = round(SSE/(n-p), 4)    # output - coefficient  name = paste("Beta",c(0:p),":" ,sep="")  name[1] = "Constant:"  row.names(b) = name  # make output file  cat("Coefficients","\n","-------------","\n",file = outputname,sep="")  write.table(b, outputname, sep= " ", row.names=TRUE, col.names=FALSE, append=TRUE, quote=FALSE) |

* H 매트릭스를 사용하여 다중회귀분석을 수행한 후, 그 추정계수를 b를 output 파일에 저장하는 코드의 일부이다.

c. The program must calculate the fitted values and save it into an output file.

|  |
| --- |
| # multiple regression result  yhat = x%\*%b  # output - ID, Actual values, Fitted values  y.values = cbind(c(1:n), y, round(yhat,1))  # make output file  cat("\n","ID, Actual values, Fitted values","\n","--------------------------------","\n",  file = outputname,sep="", append=TRUE)  write.table(y.values, outputname, sep= ", ", row.names=FALSE, col.names=FALSE,  append=TRUE, quote=FALSE) |

* 매트릭스를 이용하여 구한 fiited value(yhat)에 ID와 Actual value를 추가하여 output 파일에 저장하는 코드의 일부이다.

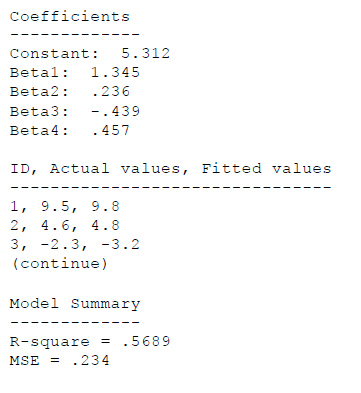
f. Use a data file named “harris.dat” for this assignment. Assume that the first one is the response variable.

|  |
| --- |
| regression.model()  harris.dat  1  harris.regression.output |

|  |
| --- |
| **( R 결과창 )**    **( text file 생성된 결과 )** |

* 첫번째 열을 반응변수로 가정하고 “harris.dat”을 regression.model 함수를 이용하여 회귀분석을 진행한 결과이다. output으로 지정해준 파일 이름 그대로 “harris.regression.output.txt”가 생성됨을 확인할 수 있다.

g. The output file generated by the program must look like the below (the sample output is not the true one).



|  |
| --- |
| **( lm 함수를 이용한 회귀분석 결과창 )** |

|  |
| --- |
| **( harris.regression.output.txt )**  Coefficients  -------------  Constant: 3526.4221  Beta1: 90.0203  Beta2: 1.269  Beta3: 23.4062  Beta4: 722.4607  ID, Actual values, Fitted values  --------------------------------  1, 3900, 4630.1  2, 4020, 4646.3  3, 4290, 5315.2  4, 4380, 4418.3  ………(생략)……..  90, 6840, 5815.7  91, 6900, 5785.3  92, 6900, 6328.4  93, 8100, 6530.8  Model Summary  -------------  R-square = 0.5109  MSE = 254583.5835 |

* 위의 두 포맷이 동일하고 lm 함수로 구한 결과가 동일함을 확인할 수 있다.