Rennehan,Richard

NSCC Institute of Technology

DBAS3075 Introduction to Statistical Learning

Multiple Regression Assignment

March 7, 2018

Contents

[Questions and Answers 2](#_Toc508182088)

[References 8](#_Toc508182089)

[R Code: 9](#_Toc508182090)

# Questions and Answers

1. Read in the data from the website provided using the read.csv() function, and be sure to indicate in the arguments that the data contains a header, and is tab delimited. Report the line of code you used to perform this step.
   1. body\_data=read.csv("http://www.statsci.org/data/oz/physical.txt", header=TRUE, sep="\t")
2. List the names of the variables included in the data set using the names() function, and provide a summary of each variable using the summary() function.
   1. Mass, Fore, Bicep, Chest, Neck, Shoulder, Waist, Height, Calf, Thigh, and Head are the variable names.
   2. > summary(body\_data$Mass)

Min. 1st Qu. Median Mean 3rd Qu. Max.

54.50 66.00 73.00 73.93 80.38 94.00

> summary(body\_data$Fore)

Min. 1st Qu. Median Mean 3rd Qu. Max.

24.00 26.50 28.25 27.77 29.38 31.00

> summary(body\_data$Bicep)

Min. 1st Qu. Median Mean 3rd Qu. Max.

28.50 31.00 33.50 33.20 35.88 38.00

> summary(body\_data$Chest)

Min. 1st Qu. Median Mean 3rd Qu. Max.

87.50 94.25 99.50 99.64 105.50 112.00

> summary(body\_data$Neck)

Min. 1st Qu. Median Mean 3rd Qu. Max.

35.00 35.75 38.50 37.59 39.00 40.50

> summary(body\_data$Shoulder)

Min. 1st Qu. Median Mean 3rd Qu. Max.

102.0 105.8 113.5 111.8 117.8 121.0

> summary(body\_data$Waist)

Min. 1st Qu. Median Mean 3rd Qu. Max.

74.00 80.62 84.00 85.57 91.25 101.00

> summary(body\_data$Height)

Min. 1st Qu. Median Mean 3rd Qu. Max.

168.5 173.2 178.8 178.5 182.9 188.0

> summary(body\_data$Calf)

Min. 1st Qu. Median Mean 3rd Qu. Max.

32.00 36.00 38.00 37.34 38.88 42.00

> summary(body\_data$Thigh)

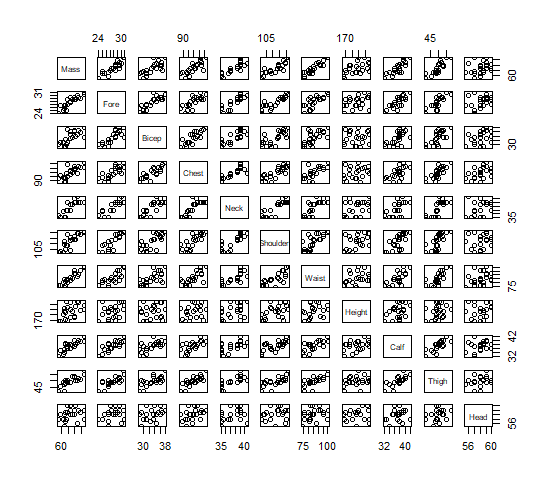
Min. 1st Qu. Median Mean 3rd Qu. Max.

42.00 49.00 49.75 49.73 51.75 57.50

> summary(body\_data$Head)

Min. 1st Qu. Median Mean 3rd Qu. Max.

55.50 57.00 58.00 58.07 59.00 60.00

1. Using the pairs() function, look at the plots of each variable plotted against each of the other variables. Include this graphic in your report.
   1. The following graphic is the result of using the pairs function
2. Perform a linear regression on this data to predict Mass using all 10 of the other variables. Report model with each coefficient rounded to 2 decimal places. Also report the R^2 and adjusted-R^2 in your report. Why might these statistics be unreliable for this particular data?
   1. Mass = -69.52 + 1.78(Fore) + 0.16(Bicep) + 0.19(Chest) - 0.48(Neck) - 0.03(Shoulder) + 0.66(Waist) + 0.32(Height) + 0.44(Calf) + 0.30(Thigh) - 0.92(Head)

Multiple R-Squared: 0.9772

Adjusted R-Squared: 0.9565

The data might be unreliable because we only have 11 degrees of freedom (observations subtract variables). The more variables we have, the more observations we should have to reduce chances of the model from picking up patterns caused by random chance. Twenty-two observations is too little for a model with 11 variables.

1. The correlations between variables can be viewed using the cor() function. These correlations can range between -1 and 1. Values close to 1 are highly correlated. Values close to -1 are highly anti-correlated. Values close to 0 are uncorrelated. Use the cor() function on the data set and examine the correlations between Mass and each other variable (so you should have 10 correlations in total). Report these correlation. Note: A variable will have a correlation of 1 with itself.
   1. mass/waist: 0.92

mass/fore: 0.90

mass/shoulder: 0.86

mass/thigh: 0.84

mass/calf: 0.83

mass/neck: 0.81

mass/chest: 0.78

mass/bicep: 0.73

mass/height: 0.49

mass/head: 0.25

1. For this assignment, we will consider any correlations between -0.5 and 0.5 between a predictor variable and the Mass variable to be not good predictors. Identify these variables and then exclude these variables from the data set. You should end up with Mass and 8 predictors.
   1. The correlations with mass vs height and mass vs head lie between -0.5 and 0.5, so we will get rid of the height and head variables.
2. When building models, we typically prefer to have the best predictive model using the fewest number of predictor variables. When two predictor variables are highly correlated, it is possible that we do not need them both. We can see these correlations visually in the plot produced in question #3. Alternatively, we can explore the correlations between the predictors. For any pairs of predictor variables with correlations greater than 0.85, remove the predictor with the lower correlation with Mass. You should end up with Mass and 6 predictors.
   1. The first pair with correlations greater than 0.85 is mass/shoulder (0.86) with mass/thigh (0.84). The second pair is mass/waist (0.92) with mass/fore (0.90). Therefore, we will get rid of the thigh and fore variables.
3. Perform linear regression on Mass using the 6 remaining predictor variables. Report model with each coefficient rounded to 2 decimal places. Also report the R^2 and adjusted-R^2 in your report.
   1. Mass = -97.15 - 0.13(Bicep) - 0.0025(Chest) + 0.53(Neck) + 0.41(Shoulder) + 0.68(Waist) + 1.37(Calf)

Multiple R-squared is 0.9412

Adjusted R-squared is 0.9177

1. Using the hypotheses tests included in the summary output of your 6 predictor variable model, which variables have p-values less than 0.1, and are therefore considered significant in the model? Construct a new model using only the significant predictors. Report model with each coefficient rounded to 2 decimal places. Also report the R^2 and adjusted-R^2 in your report.
   1. The significant predictors are:

Shoulder: 0.07361

Waist: 0.0030

Calf: 0.00494

New model:

Mass = -90.10 + 0.48(Shoulder) + 0.71(Waist) + 1.33(Calf)

Multiple R-Squared is 0.945

Adjusted R-Squared is 0.9358

# References

Data Set: <http://www.statsci.org/data/oz/physical.txt>

## R Code:

body\_data=read.csv("http://www.statsci.org/data/oz/physical.txt", header=TRUE, sep="\t")

names(body\_data)

summary(body\_data$Mass)

summary(body\_data$Fore)

summary(body\_data$Bicep)

summary(body\_data$Chest)

summary(body\_data$Neck)

summary(body\_data$Shoulder)

summary(body\_data$Waist)

summary(body\_data$Height)

summary(body\_data$Calf)

summary(body\_data$Thigh)

summary(body\_data$Head)

pairs(body\_data)

model1=lm(Mass~., body\_data)

model1sum=summary(model1)

model1sum

cor(body\_data)

names(body\_data)

body\_data\_no\_weak\_cor = body\_data[,-c(8,11)] #New dataset that removes height and head attributes due to low correlation

names(body\_data\_no\_weak\_cor)

body\_data\_no\_pairs = body\_data\_no\_weak\_cor[,-c(9,2)] #New dataset that removes thigh and fore

model2=lm(Mass~.,body\_data\_no\_pairs)

summary(model2)

names(body\_data\_no\_pairs)

body\_data\_only\_significant = body\_data\_no\_pairs[,-c(2,3,4)] #Leaves us with only mass, shoulder, waist, and calf for our final model

names(body\_data\_only\_significant)

model3=lm(Mass~.,body\_data\_only\_significant)

summary(model3)