IS4240 Business Intelligence Systems – Assignment 2
Raghavendhra Balaraman – A0123443R

a) Provide the R codes for loading the dataset into a variable bodyfat. The attributes should be given reasonable attribute names based on the description given above. You should remove the density attribute. (1 marks)

```
bodyfat <- read.table("bodyfat.data", sep=",", col.names =
c("bodyfatpercent","age", "weight", "height", "neck_circumference",
"chest_circumference", "abdomen_2_circumference",
"hip_circumference", "thigh_circumference", "knee_circumference",
"ankle_circumference", "biceps_circumference",
"forearm_circumference", "wrist_circumference"),
colClasses="numeric")
```

Note :

- The reasonable attributes names are given using 'col.names'
- The density attribute has been removed

b) Provide the R codes for generating a multiple linear regression model (model) using the percentage of body fat value as response (dependent variable) and the other attributes as the independent variable. (1 marks)

R code for generating Multiple Linear Regression Model:

```
model <- lm(bodyfat$bodyfatpercent~bodyfat$age +
bodyfat$weight+bodyfat$height+bodyfat$neck_circumference+bodyfat$chest_
circumference+bodyfat$abdomen_2_circumference+bodyfat$hip_circumferenc
e+bodyfat$thigh_circumference+bodyfat$knee_circumference+bodyfat$ankle_
circumference+bodyfat$biceps_circumference+bodyfat$forearm_circumferenc
e+bodyfat$wrist_circumference)
```

lm() is the function that performs multiple linear regression on the dataset.

bodyfat$bodyfatpercent – Dependent Attribute
Remaining Columns – Independent Attribute

c) Execute `summary(model)` and show the result. Based on this result, write down the equation of regression model. (1 mark)

R code for generating Multiple Linear Regression Model :

```
summary(model)
```

Result:

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.1687  -2.8639  -0.1014   3.2085  10.0068

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -18.18849   17.34857  -1.048  0.29551
bodyfat$age                     0.06208    0.03235   1.919  0.05618 .
bodyfat$weight                 -0.08844    0.05353  -1.652  0.09978 .
bodyfat$height                 -0.06959    0.09601  -0.725  0.46925
bodyfat$neck_circumference     -0.47060    0.23247  -2.024  0.04405 *
bodyfat$chest_circumference    -0.02386    0.09915  -0.241  0.81000
bodyfat$abdomen_2_circumference  0.95477    0.08645  11.044  < 2e-16 ***
bodyfat$hip_circumference      -0.20754    0.14591  -1.422  0.15622
bodyfat$thigh_circumference     0.23610    0.14436   1.636  0.10326
bodyfat$knee_circumference      0.01528    0.24198   0.063  0.94970
bodyfat$ankle_circumference     0.17400    0.22147   0.786  0.43285
bodyfat$biceps_circumference    0.18160    0.17113   1.061  0.28966
bodyfat$forearm_circumference   0.45202    0.19913   2.270  0.02410 *
bodyfat$wrist_circumference    -1.62064    0.53495  -3.030  0.00272 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.0.1 ' ' 1
Residual standard error: 4.305 on 238 degrees of freedom
Multiple R-squared:  0.749, Adjusted R-squared:  0.7353
      F-statistic: 54.65 on 13 and 238 DF,  p-value: < 2.2e-16
```

Equation of the Regression Model:

bodyfatpercent = -18.18849 + 0.06208age - 0.08844weight - 0.06959height - 0.47060neck_circumference - 0.02386chest_circumference + 0.95477abdomen_2_circumference - 0.20754hip_circumference + 0.23610thigh_circumference + 0.01528knee_circumference + 0.17400ankle_circumference + 0.18160biceps_circumference + 0.45202forearm_circumference - 1.62064wrist_circumference

d) Comment on the model generated in (c). (3 mark)

Level of Significance : From the model generated in (c), we can clearly see that the predictors are 'neck_circumference', 'abdomen_2_circumference', 'forearm_circumference', 'wrist_circumference'. They are identified considering the confidence value as 95%. We can reject the null hypothesis for the above parameters as ther p value is low (<0.05). In short, **attributes/predictors having a low p-value have a significant relationship with the response (bodyfatpercent) i.e., any change in the predictors will affect 'bodyfatpercent'**

Regression coefficient : From the equation of the regression model, the regression coefficient for each predictor can be explained as follows

Parameter	Regression Coefficient	Description
neck_circumference	0.47060	For every additional increase in neck_circumference, the bodyfatpercent **increases** by an average of 0.4760 percent
abdomen_2_circumference	0.95477	For every additional increase in abdomen_2_circumference, the bodyfatpercent **increases** by an average of 0.95477 percent
forearm_circumference	0.45202	For every additional increase in forearm_circumference, the bodyfatpercent **increases** by an average of 0.45202 percent
wrist_circumference	-1.62064	For every additional increase in wrist_circumference, the bodyfatpercent **decreases by** an average of 1.62064 percent

Adjusted R-Squared Value : R-Squared values are the measures to identify how close the data has been fitted to the regression line. **In general, higher the R-squared value, better the model fits with the data**. From our analysis, the adjusted R-squared measure indicates a value of 0.7353. This value indicates that, the model explains 73.5% variability of the dependent data (here, bodyfatpercent).

e) Based on the results from (c), provide the R codes for generating another multiple regression model (model1) with predictors that are significant. Show the results from summary(model1) and write down the equation of model1. (2 mark)

```
model1 <- lm(formula = bodyfat$bodyfatpercent ~
bodyfat$neck_circumference +
bodyfat$abdomen_2_circumference +
bodyfat$forearm_circumference +
bodyfat$wrist_circumference)
```

summary(mode1l)

Result:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.6179  -3.1114  -0.1505   2.9468  11.7076

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -6.4679     5.9591  -1.085 0.278814
bodyfat$neck_circumference        -0.6817     0.2202  -3.095 0.002192 **
bodyfat$abdomen_2_circumference    0.8144     0.0406  20.059  < 2e-16 ***
bodyfat$forearm_circumference      0.3044     0.1855   1.641 0.102109
bodyfat$wrist_circumference       -1.7874     0.4736  -3.774 0.000201 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.512 on 247 degrees of freedom
Multiple R-squared:  0.7139,     Adjusted R-squared:  0.7093
F-statistic: 154.1 on 4 and 247 DF,  p-value: < 2.2e-16
```

Equation of the Regression Model 1:

```
bodyfatpercent = -6.4679 -
0.6817neck_circumference +
0.8144abdomen_2_circumference +
0.3044forearm_circumference -
1.7874wrist_circumference
```

f) Other than R^2, we often make use of Mean Squared Error (MSE) to measure the quality of fit of regression model. Provide the R codes for calculating the MSE for model and model1 and write down the MSE of the 2 models. (1 mark)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

where *f(x)* is the prediction for observation *i* based on the model

Function to calculate Mean Squared Error (MSE):

```
mse <- function(sum) {
    mse <-
mean(sum$residuals^2)
    return(mse)
}
```

MSE of Model :

```
Model
> mse(sum)
[1] 17.50575
```

MSE of Model 1:

```
Model 1
> mse(sum1)
[1] 19.95858
```

g) Is there any difference between the MSE values? Discuss whether this is expected. (1 mark)

Yes, the MSE Values differs.

The Mean Squared values are helpful in identifying how accurate the attributes predicts the dependent variable (bodyfatpercent). Again, **lower the MSE value, more accurate the predictors are.** An MSE of zero indicates that the independent variables predicts the observation of the dependent variable (bodyfatpercent) with perfect accuracy. **As a result, 'Model' predicts the observation of 'bodyfatpercent' more accurately than 'Model1'.** The following are some of the analysis as to why MSE of 'Model' is lower than the 'Model1'.

- The adjusted R-squared value of 'Model' (0.7353) is greater than the 'Model 1' (0.7093). This explains that the 'Model' best fits the data than 'Model1'
- The removal of some of the attributes from 'Model' could have also influenced a higher MSE for 'Model 1'.
- The estimators that are chosen for 'Model1' may not account for the information that could help in identifying more accurate results.