A1 Methods of Machine Learning

Sebastian Doka

2023-01-30

Question 1: We then load the `penguins` dataset and Select only the columns containing the species, sex, body mass, flipper length and bill length. At the end we remove the rows with missing values in any of these columns.

```
## Warning: package 'palmerpenguins' was built under R version 4.1.3
```

```
##       species          island    bill_length_mm  bill_depth_mm  
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10  
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60  
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30  
##                                  Mean   :43.92   Mean   :17.15  
##                                  3rd Qu.:48.50   3rd Qu.:18.70  
##                                  Max.   :59.60   Max.   :21.50  
##                                  NA's   :2       NA's   :2      
##  flipper_length_mm  body_mass_g       sex           year     
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007  
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007  
##  Median :197.0     Median :4050   NA's  : 11   Median :2008  
##  Mean   :200.9     Mean   :4202                Mean   :2008  
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009  
##  Max.   :231.0     Max.   :6300                Max.   :2009  
##  NA's   :2         NA's   :2
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1 
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0 
## v readr   2.1.3      v forcats 1.0.0
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

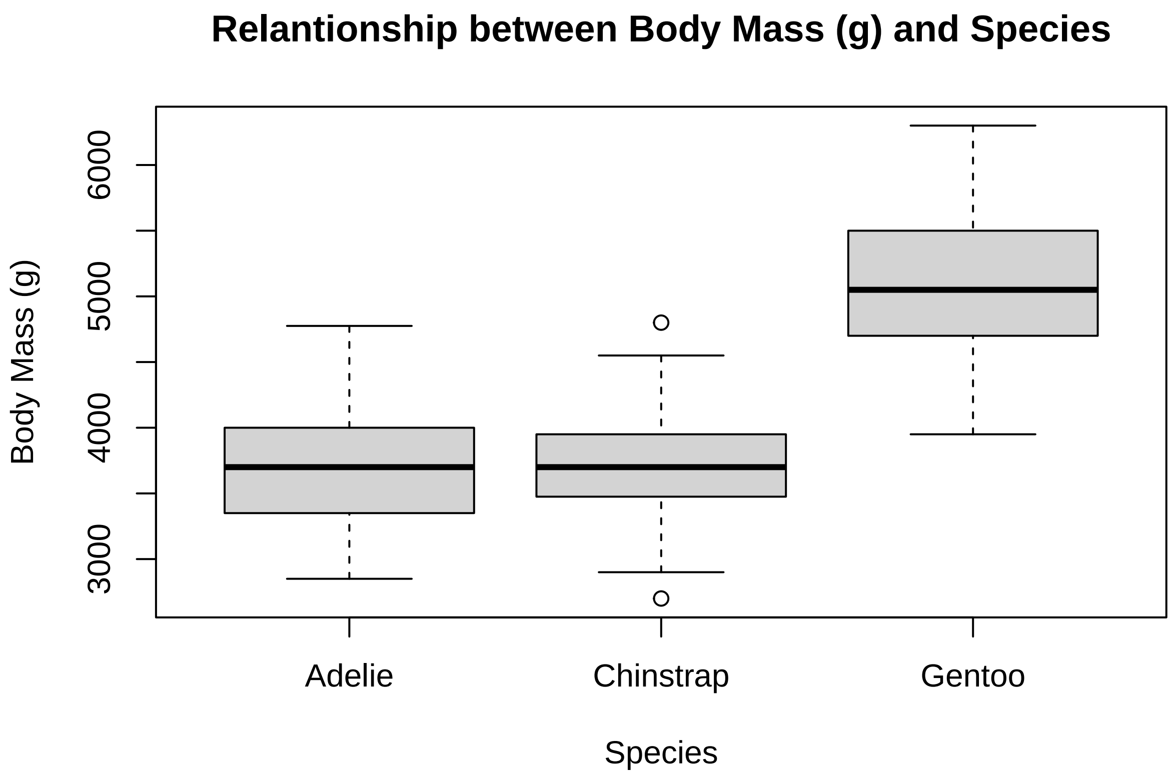
```
## Warning: package 'readr' was built under R version 4.1.3
```

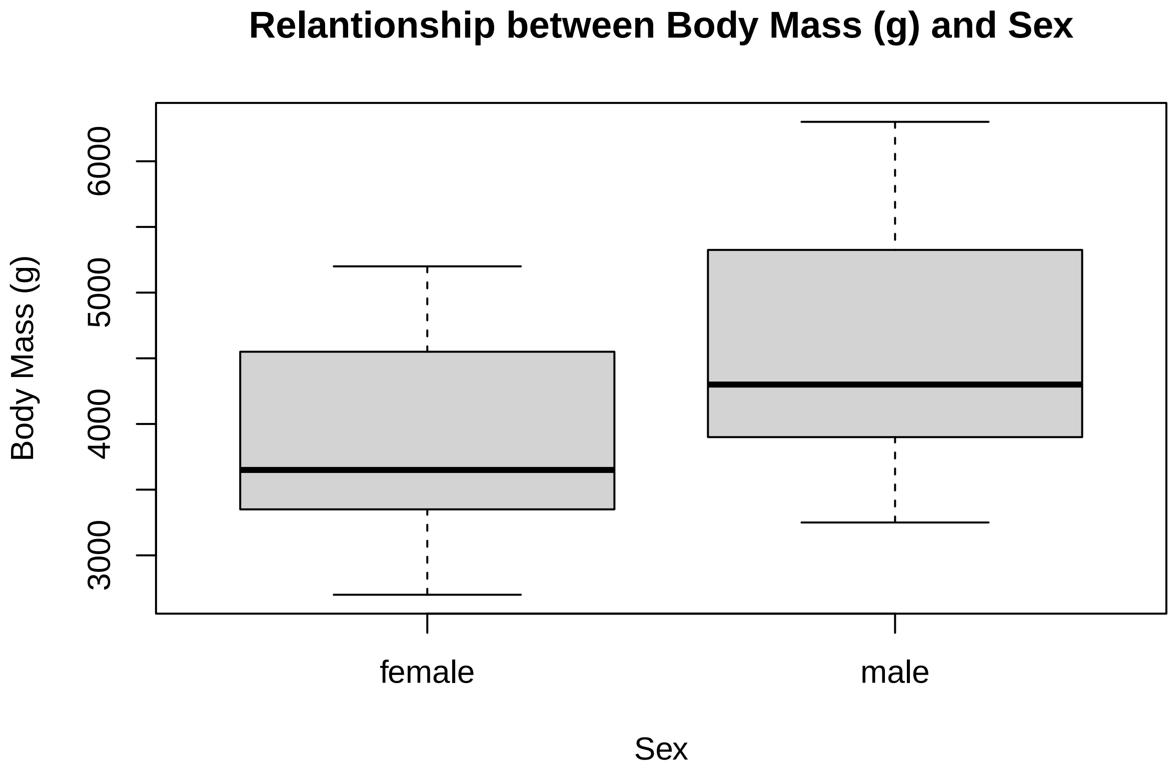
```
## Warning: package 'purrr' was built under R version 4.1.3
## Warning: package 'dplyr' was built under R version 4.1.3
```

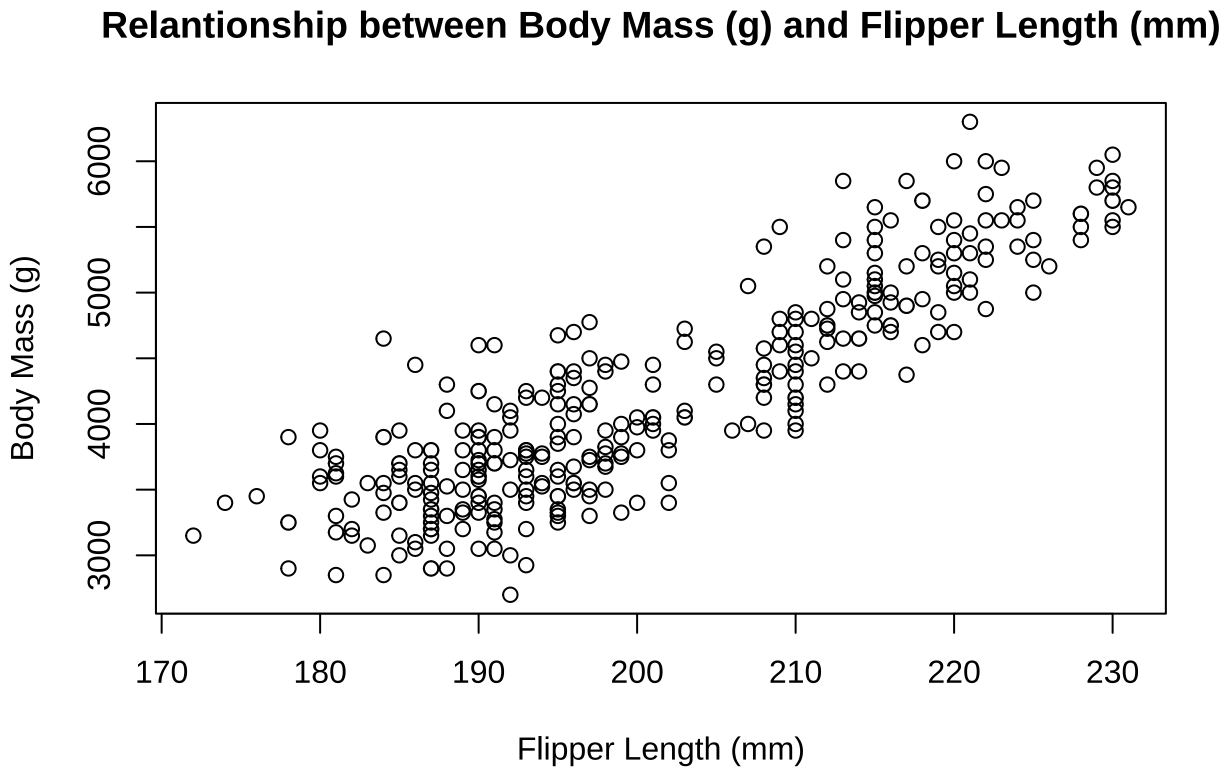
```
## Warning: package 'stringr' was built under R version 4.1.3
```

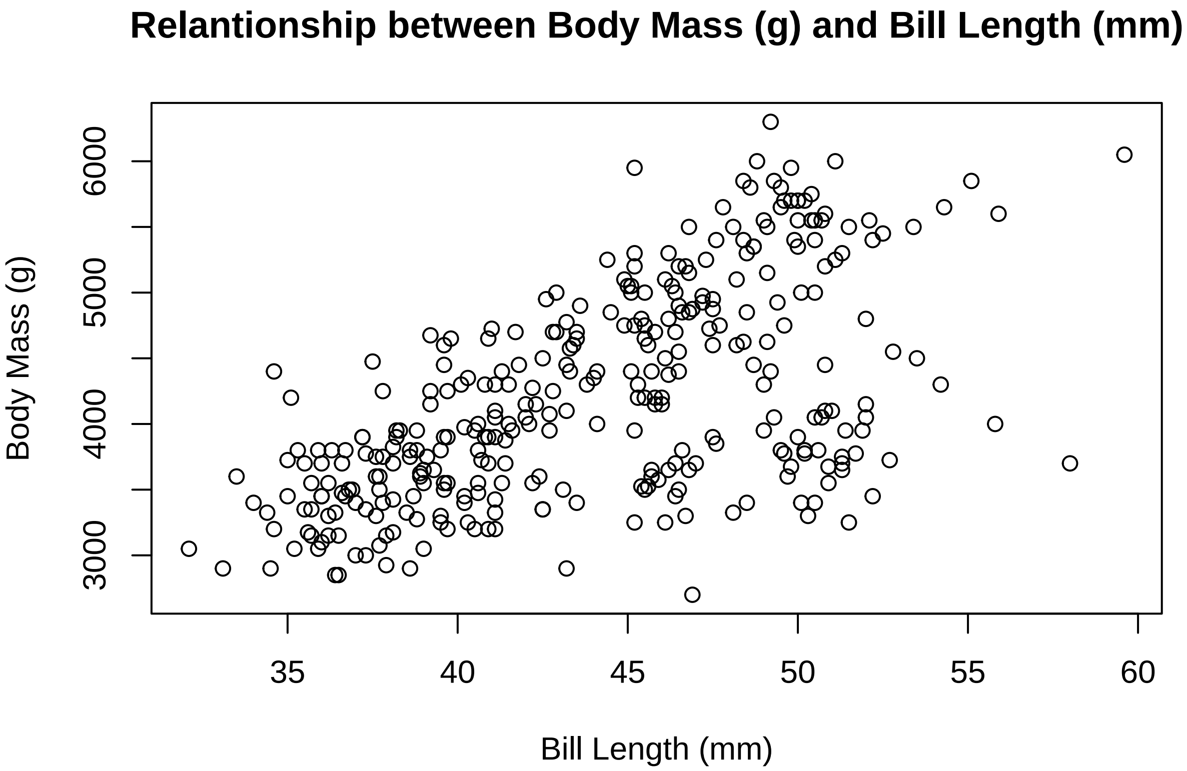
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Now we create plots to visualize the relationship between body mass and the other 4 variables:




Relantionship between Body Mass (g) and Sex


Relantionship between Body Mass (g) and Flipper Length (mm)
Body Mass (g)
3000
4000
5000
6000
170
180
190
200
210
220
230
Flipper Length (mm)



Question 2

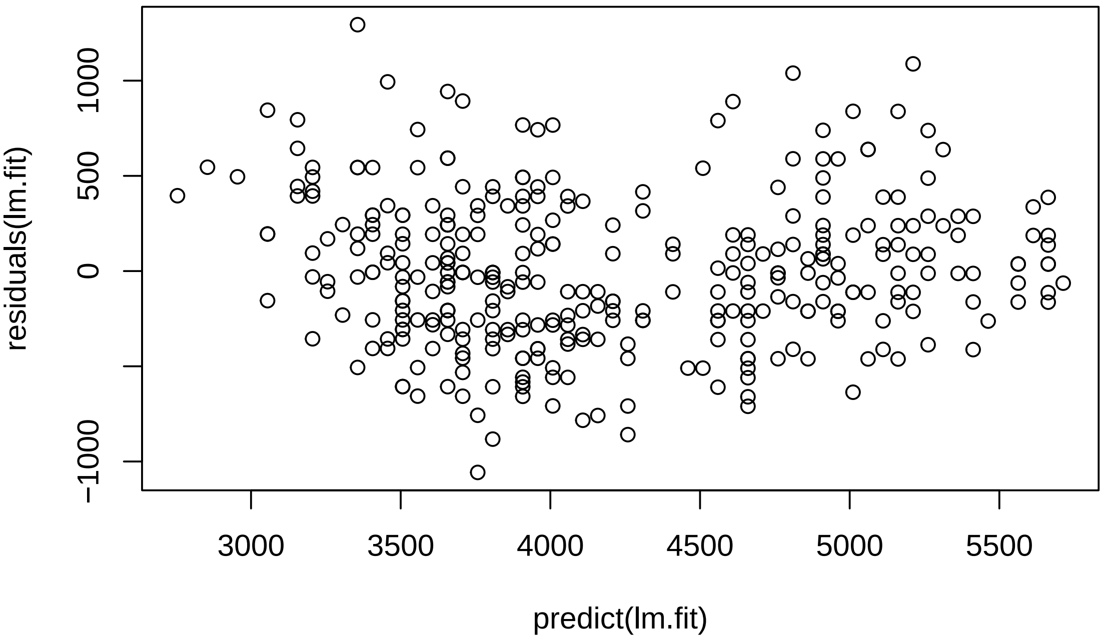
Performing a simple OLS linear regression of body mass (the response variable) on flipper length (the predictor), using all rows from the data.

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1057.33  -259.79   -12.24   242.97  1293.89
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
## flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 393.3 on 331 degrees of freedom
## Multiple R-squared:  0.7621, Adjusted R-squared:  0.7614
## F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

```
## [1] -5872.093
```

A) Linear model obtained from OLS: body mass = -5872.0927 + 50.1533 * (flipper length).

The value for R^2 is $R^2 = 0.7620922$.



B)

This is a good model, residuals seem normally distributed around zero, with a mean and variance that is independent of the predicted value. Looking at the R^2 value it is a pretty high value.

Question 3

A)Do an OLS multiple linear regression of body mass on the four predictors: species, sex, flipper length and bill length. You may use any sensible encoding of your qualitative predictors. State the resulting linear model in the same form as in Question 2(A). Report R2 and also the residual sum of squares (RSS), which is the same thing as the sum of squared errors (SSE).

```
##
## Call:
## lm(formula = body_mass_g ~ species + sex + flipper_length_mm +
##     bill_length_mm, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -718.50 -201.60  -12.75  198.45  878.24
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -759.064    541.377  -1.402 0.161834
## speciesChinstrap  -291.711     81.502  -3.579 0.000397 ***
## speciesGentoo      707.028     94.359   7.493 6.35e-13 ***
## sexmale            465.395     43.081  10.803  < 2e-16 ***
## flipper_length_mm   17.847      2.902   6.150 2.25e-09 ***
## bill_length_mm      21.633      7.148   3.027 0.002670 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292 on 327 degrees of freedom
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8685
## F-statistic: 439.7 on 5 and 327 DF,  p-value: < 2.2e-16
```

```
## Warning: package 'fastDummies' was built under R version 4.1.3
```

```
## # A tibble: 333 x 8
##    bill_length_mm flipper_leng~1 body_~2 speci~3 speci~4 speci~5 sex_f~6 sex_m~7
##             <dbl>          <int>   <int>   <int>   <int>   <int>   <int>   <int>
##  1           39.1            181    3750       1       0       0       0       1
##  2           39.5            186    3800       1       0       0       1       0
##  3           40.3            195    3250       1       0       0       1       0
##  4           36.7            193    3450       1       0       0       1       0
##  5           39.3            190    3650       1       0       0       0       1
##  6           38.9            181    3625       1       0       0       1       0
##  7           39.2            195    4675       1       0       0       0       1
##  8           41.1            182    3200       1       0       0       1       0
##  9           38.6            191    3800       1       0       0       0       1
## 10           34.6            198    4400       1       0       0       0       1
## # ... with 323 more rows, and abbreviated variable names 1: flipper_length_mm,
## #   2: body_mass_g, 3: species_Adelie, 4: species_Chinstrap, 5: species_Gentoo,
## #   6: sex_female, 7: sex_male
```

Linear model obtained from OLS multiple linear regression: body mass = -759.0644 + -291.7106 * (species) + 707.028 * (sex) + 465.395 * (flipper length) + 17.8465 * (bill length).

The value for R^2 is $R^2 = 0.8705155$.

The value for the Residual Sum of Squares(RSS) is RSS $= 2.7872792 \times 10^7$.

B)

```
##
## Call:
## lm(formula = body_mass_g ~ species_Chinstrap + species_Gentoo +
##     sex_male + flipper_length_mm + bill_length_mm, data = penguins_dummy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -718.50 -201.60  -12.75  198.45  878.24
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -759.064    541.377  -1.402 0.161834
## species_Chinstrap -291.711     81.502  -3.579 0.000397 ***
## species_Gentoo     707.028     94.359   7.493 6.35e-13 ***
## sex_male           465.395     43.081  10.803  < 2e-16 ***
## flipper_length_mm   17.847      2.902   6.150 2.25e-09 ***
## bill_length_mm      21.633      7.148   3.027 0.002670 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 292 on 327 degrees of freedom
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8685
## F-statistic: 439.7 on 5 and 327 DF,  p-value: < 2.2e-16
```

```
dfz = as.data.frame(scale(penguins_dummy))
lm4.fit = lm(body_mass_g ~ species_Chinstrap + species_Gentoo + sex_male + flipper_length_mm + bill_leng
summary(lm4.fit)
```

```
##
## Call:
## lm(formula = body_mass_g ~ species_Chinstrap + species_Gentoo +
##     sex_male + flipper_length_mm + bill_length_mm, data = dfz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89231 -0.25037 -0.01584  0.24646  1.09069
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.946e-16  1.987e-02   0.000 1.000000
## species_Chinstrap -1.463e-01  4.086e-02  -3.579 0.000397 ***
## species_Gentoo     4.214e-01  5.624e-02   7.493 6.35e-13 ***
## sex_male           2.894e-01  2.679e-02  10.803  < 2e-16 ***
## flipper_length_mm  3.106e-01  5.051e-02   6.150 2.25e-09 ***
## bill_length_mm     1.469e-01  4.854e-02   3.027 0.002670 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3626 on 327 degrees of freedom
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8685
## F-statistic: 439.7 on 5 and 327 DF,  p-value: < 2.2e-16
```

After standardizing the variables, we can see the coefficients of speciesGentoo has the largest absolute value(4.214e_01) so it has the strongest value on the prediction of the variable of body mass.

Question 4

A)

```
SSE <- function(ypred,ytrue)
{
  sum=0
  for(x in 1:length(ypred)){
     ss = (ypred[x] - ytrue[x])**2
     sum = sum + ss
  }
    return(sum)
  }
```

B)

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -885.93 -247.61  -32.89  265.04 1056.83
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5667.743    458.518  -12.36   <2e-16 ***
## flipper_length_mm    49.112      2.295   21.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 385.7 on 164 degrees of freedom
## Multiple R-squared:  0.7363, Adjusted R-squared:  0.7347
## F-statistic:   458 on 1 and 164 DF,  p-value: < 2.2e-16
```

The values for SSE_Training and SSE_Test are : SEE_Training $= 2.4398752 \times 10^7$ and SEE_Test $= 2.6891724 \times 10^7$

C)

The values for SSE_Training and SSE_Test are : SEE_Training $= 1.2169172 \times 10^7$ and SEE_Test $= 1.5841883 \times 10^7$

D)

Method	Training SSE	Test SSE
Simple Linear (flipper only)	2.4398752×10^7	2.6891724×10^7
Multiple Linear	1.2169172×10^7	1.5841883×10^7