MAT 3373 Winter 2023. Assignment 1 Solutions

Total Marks: 20.

Question 1 [4 marks]

Load the ***penguins*** *dataset. Select only the columns containing the species, sex, body mass, flipper length and bill length. Remove the rows with missing values in any of these columns.*

Solution: [2 marks]

```
library(palmerpenguins)

df = penguins[,c("species","sex","bill_length_mm","flipper_length_mm","body_mass_g")]
```

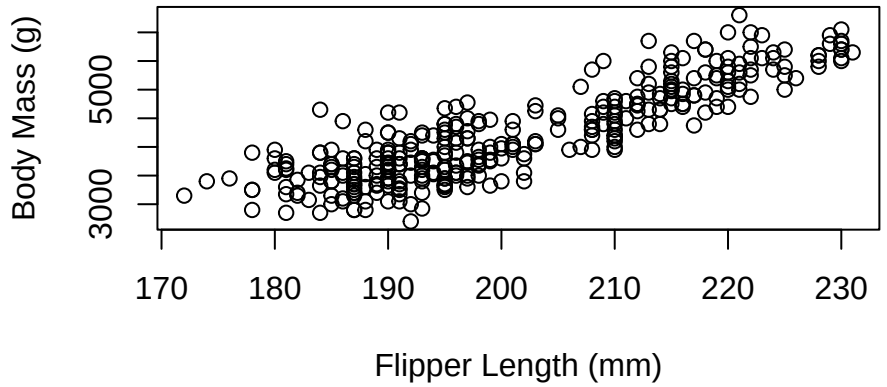
```
# or alternatively:
library(dplyr)
df = select(penguins, species, sex, body_mass_g, flipper_length_mm, bill_length_mm)
```

```
# omit all rows with any missing values
df = na.omit(df)
```

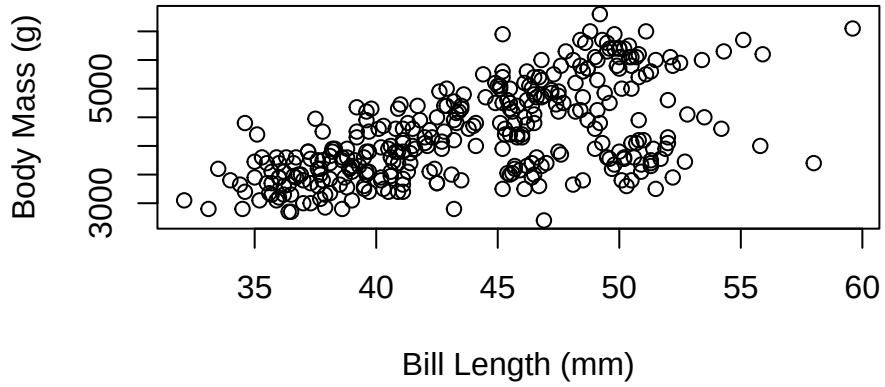
Produce plots (a maximum of 4), visualising the relationship between body mass and the other four variables. (Nothing fancy is needed here, but plots should have appropriately named axis labels, e.g. "Body Mass (g)")

Solution: [2 marks]

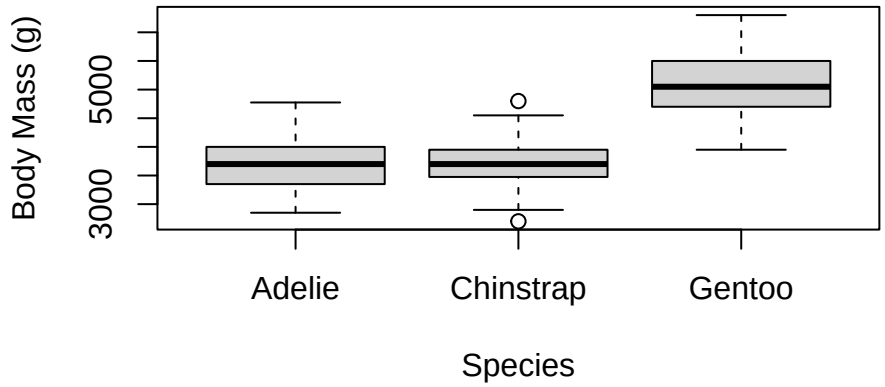
```
plot(df$flipper_length_mm, df$body_mass_g,
     xlab = "Flipper Length (mm)",
     ylab="Body Mass (g)")
```



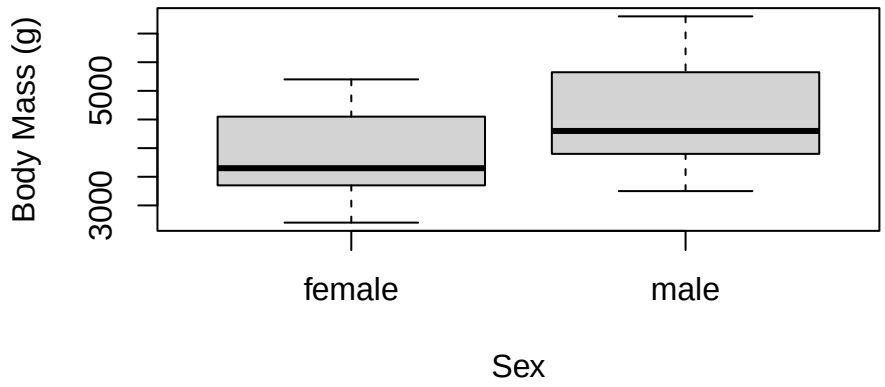

```
plot(df$bill_length_mm, df$body_mass_g,
     xlab = "Bill Length (mm)",
     ylab="Body Mass (g)")
```




```
plot(df$species, df$body_mass_g,
     xlab = "Species",
     ylab="Body Mass (g)")
```




```
plot(df$sex, df$body_mass_g,
     xlab = "Sex",
     ylab="Body Mass (g)")
```



Alternatively, see the Markdown source for `ggplot` code to produce similar plots.

Question 2 [4 marks]

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

Solution: [1 mark]

```
lm.fitted = lm(formula = body_mass_g ~ flipper_length_mm, data = df)
coefficients(lm.fitted)
```

```
##       (Intercept) flipper_length_mm
##       -5872.09268          50.15327
```

A) *State the linear model you obtained in the form*

$$\textit{body mass} = \beta_0 + \beta_1 * (\textit{flipper length})$$

where you replace "β_0" and "β_1" with the appropriate coefficients shown to 4 decimal places. Report R^2.

Solution: [1 mark]

$$\text{body mass} = -5872.0929 + 50.1533 * (\text{flipper length})$$

```
Rsquared = summary(lm.fitted)$r.squared
```

$R^2 = 0.7621$

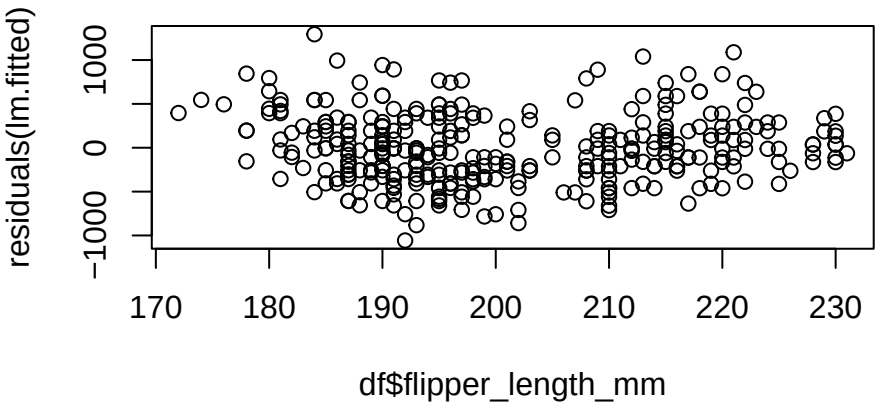
B) *Is this a good model? Answer briefly with reference to R^2 and to a plot of the residuals or some related plot.*

Solution: [2 marks]

R^2 is high (max value is 1), which indicates that flipper length is highly predictive of body mass. In fact flipper length predicts 76% of the variance of body mass. Also, this R^2 value corresponds to a correlation coefficient of $r = \sqrt{R^2} = 0.8729789$, which is a strong correlation.

Here is a plot of the residuals as a function of the predictor (flipper length). (It would be equally valid to plot the residuals vs. the predicted values of body mass.)

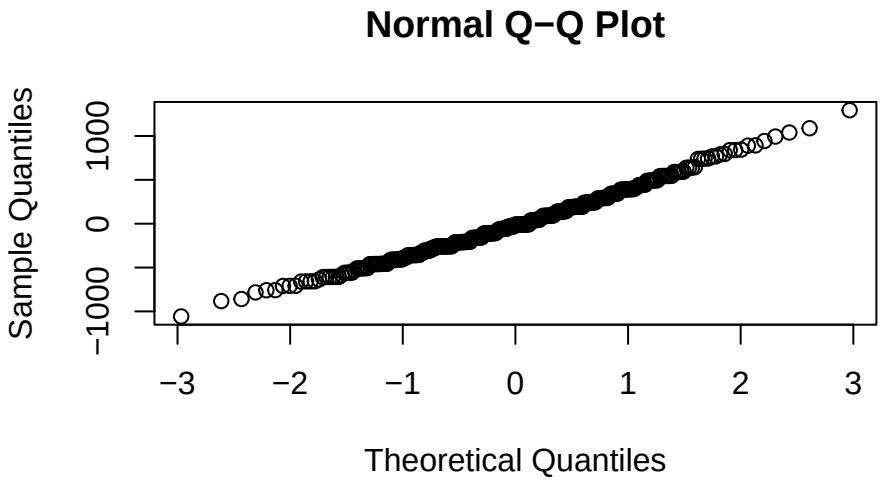
```
plot(df$flipper_length_mm, residuals(lm.fitted))
```



Notice that in this residual plot, the mean and variance of the residuals seem almost constant with respect to flipper length. (Though the mean residual is higher for the lowest flipper lengths, and variance seems lower for the very highest flipper lengths.) This is a sign that a linear model is appropriate for this data.

Also, optionally, we could examine the QQ plot:

```
qqnorm(lm.fitted$residuals)
```



The fact that this is a relatively straight line suggests that our residuals are normally distributed, which again supports our choice of a linear model.

In summary, yes this is a good model!

Question 3 [4 marks]

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 R^2 and also the residual sum of squares (RSS), which is the same thing as the sum of squared errors (SSE).*

Solution: [2 marks]

```
mlm.fitted = lm(formula = body_mass_g ~ ., data = df)
summary(mlm.fitted)
```

```
##
## Call:
## lm(formula = body_mass_g ~ ., 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
```

```
Rsquared = summary(mlm.fitted)$r.squared
```

So the model is:

$$\text{body mass} = -759 - 291 * Chin + 707 * Gen + 465 * male + 17.8 * flipper + 21.6 * bill$$

$R^2 = 0.8705$

```
ypred = predict(mlm.fitted, newdata = df)
RSS = sum((ypred - df$body_mass_g)^2)
```

$RSS = 2.7872792 \times 10^7$

B) *Which of your variables have the strongest effect on body mass? Concretely, if you wanted to predict body mass, which other variable(s) would be most useful?*

Solution: [2 marks]

Note that all predictors are significantly predictive in this model ($p < 0.01$), so we consider them all. We first standardise the variables, then recompute the model and compare the coefficients. (Note that standardization doesn't change the p values.)

```
dfz = df
dfz$flipper_length_mm = scale(dfz$flipper_length_mm)
dfz$bill_length_mm = scale(dfz$bill_length_mm)
dfz$body_mass_g = scale(dfz$body_mass_g) # optional
mlm.fitted = lm(formula = body_mass_g ~ ., data = dfz)
# summary(mlm.fitted) # to check p values
coefficients(mlm.fitted)
```

```
##       (Intercept)  speciesChinstrap     speciesGentoo           sexmale
##        -0.5313942        -0.3622763         0.8780602         0.5779755
## flipper_length_mm    bill_length_mm
##         0.3106403         0.1469185
```

Because of the scaling, the coefficients for the two continuous predictors are now comparable, and we see that flipper length has a bigger effect on body mass than bill length. (*Note that we would have gotten a different answer using the unstandardized coefficients!*)

The reasoning here is: in the context of the present set of predictor variables, if all of the other variables are held constant, the *effect* of increasing standardized flipper length by 1 is 0.31. Recall that standardized flipper length has standard deviation 1, and one way to interpret that is that 1 is a typical difference of this variable from 0, so we could say that observing this variable typically changes our estimate of it by 1, which then changes our estimate of standardized body mass by 0.31.

The binary variable `sex` could also be scaled using `scale()`, if we first changed its type to `numeric`. In fact, since the penguins include approximately equal numbers of males and females, we know the standard deviation of the current `sex` variable is approximately 0.5, so we conclude that we would have to scale `sexmale` by a factor of 2 (to give standard deviation 1), which would halve its coefficient from 0.58 to 0.29.

The `species` variable is trickier because it is coded as two highly-dependent binary variables. But as a rough approximation, we could take the average of the absolute values of the two coefficients, and then divide the result by 2, as for the `sex` variable, giving: $((.36 + .88)/2)\ /2 = 0.31$. But see below!

So far we can conclude that **flipper length and sex, and perhaps species, have a higher effect on body mass than bill length**, *in this model*. But: (1) the coefficients for species are more difficult to interpret (see below); and (2) the practical question of "which other variable(s) would be most useful?" brings in the question of variable selection, and the fact that the effect of a predictor variable depends on the context

of which other predictors are being used (recall the discussion of `cylinders` and `displacement` in Lecture 5). Since we didn't consider (1) in the lectures, and (2) was touched on only briefly, this part of the question was marked leniently, with any sensible answer getting full marks.

Here's another way to think about the categorical variables. If we don't know the sex of a penguin, then we could give a value of 0.5 to the dummy variables `sexmale`. So observing the sex of a penguin changes our estimate of this variable by ± 0.5, which then changes our estimate of standardized body mass by $(\pm 0.5) * 0.58 = \pm 0.29$.

If we apply the same analysis to the species variable, assuming that the species are equally probable to start with, then if we don't know the species, we would assign the value 0.33 to both `speciesChinstrap` and `speciesGentoo`. Then observing the species of a penguin has the following effects on our estimate of standardized body mass:

- If we find it's a Chinstrap, that changes the `speciesChinstrap` variable from 0.33 to 1, i.e. increases it by $2/3 = .67$, and also decreases the `speciesGentoo` variable from 0.33 to 0, i.e. decreases it by 0.33, so the overall change change in our estimate of standardized body mass is $0.67 * (-0.36) + (-0.33) * 0.89 = -0.53$.
- If we find it's a Gentoo, this changes the `speciesChinstrap` variable from 0.33 to 0 and the `speciesGentoo` variable from 0.33 to 1, so the overall change in our estimate of standardized body mass $(-0.33) * (-0.36) + (0.67) * 0.89 = 0.72$.
- If we find it's an Adelie, our estimate doesn't change.

So the mean absolute value of the change in our estimate is $(0.53 + 0.72 + 0)/3 = 0.42$. Note that this is also the mean of the three regression coefficients for species, if we say that the coefficient for Adelie is 0!

In conclusion, our predictors with highest effects are: **Species (0.42), Flipper Length (0.31) and Sex (0.29)**. So these are the ones I would most want to know if estimating body mass *using this model*.

That does not mean that, if we were to consider other subsets of these variables, and build a linear model for each subset, that the best one would correspond to the three variables listed above! That would be true if the variables were all independent, but they aren't here, and they rarely are in practice. This is why **variable selection** is so difficult and important; see Chapter 6 of ISLR2.

Question 4 [5 marks]

In this question we take a more "machine learning" approach: a model is good if it predicts well on a held-out test set.

A) *Write a function `SSE` that takes two inputs that are vectors of the same size, and returns the "sum of squared differences". In particular, if we evaluate `SSE(ypred, ytrue)`, where `ypred` and `ytrue` are predicted and true response values, we should get the Sum of Squared Errors.*

Solution [1 mark]:

```
SSE = function(a, b){
  return(sum((a-b)^2))
}
```

B) *Set your random seed by running `set.seed(1)`. Randomly split your dataset into Training and Test subsets of equal (or approximately equal) sizes. Fit an OLS **simple** linear regression model to predict body mass from flipper length (only), using the Training data only. Then apply that model to predict body mass for penguins in the Training and Test sets. Calculate $SSE_{Training}$ (also called RSS) and SSE_{Test}.*

Solution [2 marks]:

```
set.seed(1)

train_size = floor(0.5 * nrow(df))
```

```
train_ind = sample(seq_len(nrow(df)), size = train_size)
train = df[train_ind, ]
test = df[-train_ind, ]

lm.fitted2 = lm(formula = body_mass_g ~ flipper_length_mm, data = train)

ypred.train = predict(lm.fitted2, newdata = train)
ypred.test = predict(lm.fitted2, newdata = test)
SSE.train = SSE(ypred.train, train$body_mass_g)
SSE.test = SSE(ypred.test, test$body_mass_g)
SSE.train
```

```
## [1] 24398752
```

```
SSE.test
```

```
## [1] 26891724
```

Note that the Test SSE is higher than the Train SSE, even though the two data subsets are the same size (approximately). This is typical.

C) *Fit an OLS* ***multiple*** *linear regression model to predict body mass from the other 4 variables, using the Training data only. Then apply that model to predict body mass for penguins in the Training and Test sets. Calculate* $SSE_{Training}$ *(also called RSS) and* SSE_{Test}*.*

Solution [1 mark]:

Note: only one line changes from previous part. The pretty printing of the SSE's is optional.

```
lm.fitted2 = lm(formula = body_mass_g ~ ., data = train)

ypred.train2 = predict(lm.fitted2, newdata = train)
ypred.test2 = predict(lm.fitted2, newdata = test)

SSE.train2 = SSE(ypred.train2, train$body_mass_g)
SSE.test2 = SSE(ypred.train2, test$body_mass_g)
```

```
## Warning in a - b: longer object length is not a multiple of shorter object
## length
```

```
print(paste('Training SSE (i.e. RSS) is: ', round(SSE.train2,0)))
```

```
## [1] "Training SSE (i.e. RSS) is:  12169172"
```

```
print(paste('Test SSE is: ', round(SSE.test2,0)))
```

```
## [1] "Test SSE is:  190586382"
```

D) *Fill in the following table to summarize your results, or make a similar table with a different method. Each entry should use an R variable, so you don't have to type the values by hand. If filling in the table below (which may only show up in the PDF), you can use inline R code.*

Solution [1 mark]

Method	Training SSE	Test SSE
Simple Linear (flipper only)	2.4398752×10^7	2.6891724×10^7
Multiple Linear	1.2169172×10^7	1.9058638×10^8

Note: (1) Test SSE (i.e. Test Error) is higher than Train SSE (i.e. Training Error) for both models. (This is typical.) (2) Using more predictors lowers the Training Error (which is typical), and in this case also lowers

the Testing Error, which is ideal but doesn't always happen (due to possible overfitting).

The final 3 marks for this assignment were awarded for following these instructions: Submit to Brightspace a zip file named **MAT3373_A1_LastName_FirstName_StudentID.zip** (with "LastName" replaced by your last name etc.). The zip file should contain the following files only:

- A single R project file named **A1.Rproj**.
- A single R Markdown file named **A1.Rmd** containing your answers to all of the questions. Your R Markdown file should be well-structured and easy to read, with some helpful comments.
- A single PDF file named **A1.pdf** obtained by rendering A1.Rmd. Note that in order make this work, you will have to have a version of TeX installed, such as (recommended) TinyTeX.