Compulsory assignment 1

Lars August Melbye Olsen, Ida Sandum, Ellen Skrimstad

February 2022

Problem 1

a)

In general Mean Squared error can be decomposed like this:

$$\begin{aligned} MSE &= E\big[(y - \tilde{f}(x))^2\big] \\ &= E\big[(f(x) + \epsilon - \tilde{f}(x))^2\big] \\ &= E\big[(f(x) - \tilde{f}(x))^2\big] + E\big[\epsilon^2\big] + 2E\big[\big(f(x) - \tilde{f}(x)\big) \cdot \epsilon\big] \\ &= \big(f(x) - \tilde{f}(x)\big)^2 + var\big[\epsilon\big] + 0 \end{aligned}$$

Here using that $E[\epsilon] = 0$,

while expected test MSE can be decomposed like this:

$$\begin{aligned} MSE_{test} &= E\big[\big(y_0 - \tilde{f}(x_0)\big)^2\big] \\ &= E\big[\big(f(x_0) + \epsilon - \tilde{f}(x_0)\big)^2\big] \\ &= E\big[f(x_0)^2 + \epsilon^2 + \tilde{f}(x_0)^2 - 2f(x_0) \cdot \tilde{f}(x_0) + 2f(x_0) \cdot \epsilon - 2\tilde{f}(x_0) \cdot \epsilon\big] \\ &= f(x_0)^2 + var\big[\epsilon\big] + \big(var\big[\tilde{f}(x_0)\big] + E\big[\tilde{f}(x_0)\big]^2\big) - 2E\big[f(x_0) \cdot \tilde{f}(x_0)\big] + 0 + 0 \\ &= var\big[\epsilon\big] + var\big[\tilde{f}(x_0)\big] + f(x_0)^2 + E\big[\tilde{f}(x_0)\big]^2 - 2E\big[f(x_0) \cdot \tilde{f}(x_0)\big] \\ &= var\big[\epsilon\big] + var\big[\tilde{f}(x_0)\big] + \big(f(x_0) - E\big[\tilde{f}(x_0)\big]\big)^2 \end{aligned}$$

Where $var\big[\epsilon\big]$ is the irreducible error, $var\big[\tilde{f}(x_0)\big]$ is variance, and $\big(f(x_0) - E\big[\tilde{f}(x_0)\big]\big)^2$ is the squared bias.

b)

We can understand $var\big[\epsilon\big]$, which is the variance of the difference between y and the optimal $f(x)$ as an unavoidable mistake we will always have to do, therefore irreducible error. The variance $var\big[\tilde{f}(x_0)\big]$ is a measure of how much a resulting prediction $\tilde{f}$ varies based on the data-set. The squared bias $\big(f(x_0) - E\big[\tilde{f}(x_0)\big]\big)^2$ is a measure of how different our estimator $E\big[\tilde{f}(x_0)\big]$ is to the optimal value $f(x_0)$.

c)

TRUE, FALSE, TRUE, FALSE

d)

TRUE, FALSE, TRUE, FALSE

e)

iii).

Problem 2

a)

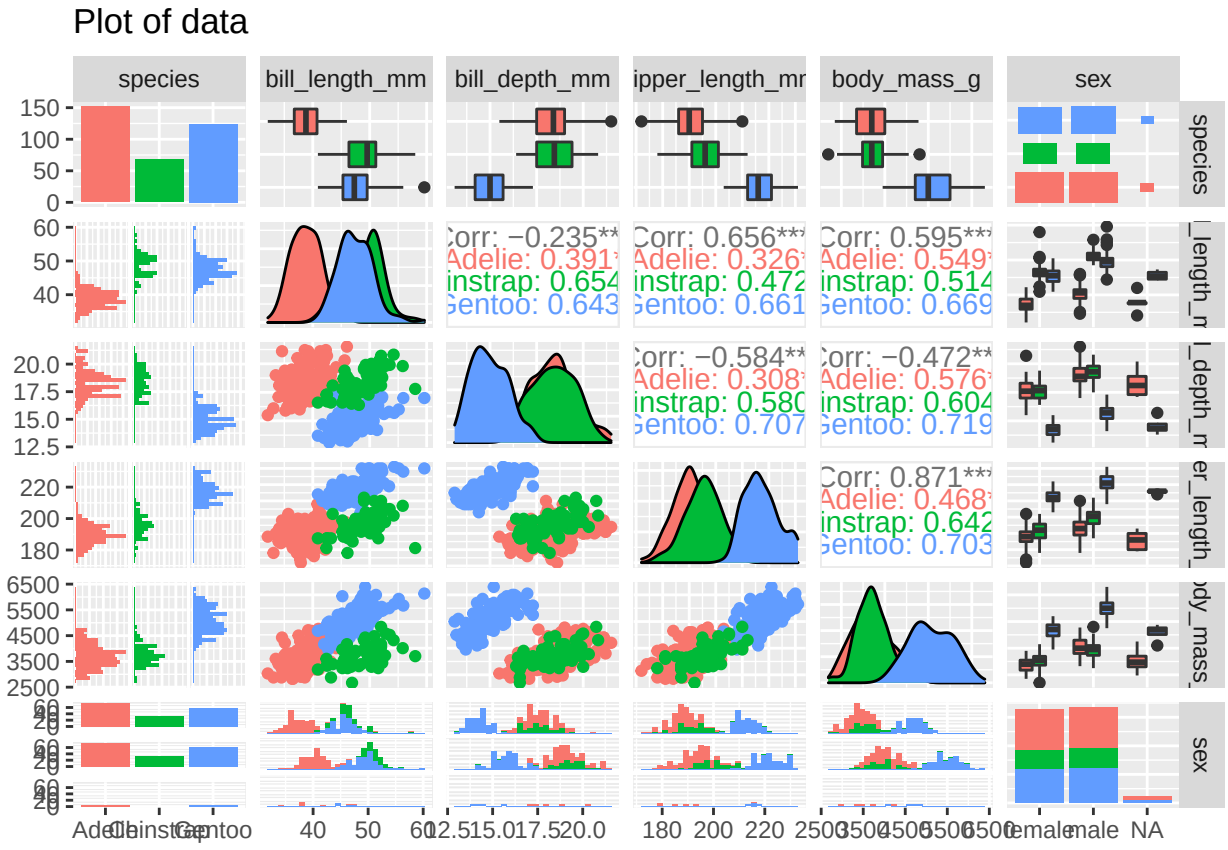
- One mistake was excluding the sex-variable because of a low p-value. A low p-value means the predictor is a meaningful addition to the model and will have an effect of the response.
- Basil should have done a F-test to see if other predictors/estimates could also have low p-values, which they do. Sex, species and flipper-length all have the same value, $< 2.2 \cdot 10^{-16}$, as can be seen in the variance table in b). Of course, he still included two of these covariates, but would have probably removed them (wrongly) if he used the same reasoning as for the sex covariate.
- We can see from the plot in b) that bill depth has a smaller correlation with body mass than bill length, and there is no apparent reason to include the former and not the latter, so Basil probably did not study his data well enough before fitting the model.

b)

As an informative plot we have chosen to use ggpairs and coloring with respect to species. We can see that one mistake was concluding that Chinstrap is the biggest species, as we can see from the plot that it is in fact Gentoo.

```
data(penguins)
# Remove island, and year variable, as we won't use those.
Penguins <- subset(penguins, select = -c(island, year))
# Fit the model as specified in advance based on expert knowledge:
penguin.model <- lm(body_mass_g ~ flipper_length_mm + sex +
                    bill_depth_mm*species, data = Penguins)

#Visualizing the data
ggpairs(Penguins, aes(colour = species)) + labs(title="Plot of data")
```



Here we have used the anova function on the expert's model, which tells us that flipper length and bill depth also have a very low p-value.

```
## Analysis of Variance Table
##
## Response: body_mass_g
##                          Df     Sum Sq    Mean Sq   F value     Pr(>F)
## flipper_length_mm         1 164047703 164047703 1994.7424 < 2.2e-16 ***
## sex                       1   9416589   9416589  114.5013 < 2.2e-16 ***
## bill_depth_mm             1   3667377   3667377   44.5936 1.051e-10 ***
## species                   2  10670525   5335262   64.8743 < 2.2e-16 ***
## bill_depth_mm:species     2    729458    364729    4.4349   0.01258 *
## Residuals               325  26728014     82240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c)

Our model with a summary:

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + sex + bill_length_mm +
##     species * bill_depth_mm, data = Penguins)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -773.5 -174.0   -3.2  168.1  906.3
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1757.120    658.082  -2.670 0.007966 **
## flipper_length_mm                 15.936      2.928   5.444 1.03e-07 ***
## sexmale                          385.683     47.350   8.145 8.28e-15 ***
## bill_length_mm                    19.752      7.124   2.773 0.005880 **
## speciesChinstrap                1539.690    674.106   2.284 0.023015 *
## speciesGentoo                    699.379    537.435   1.301 0.194071
## bill_depth_mm                     80.340     22.119   3.632 0.000327 ***
## speciesChinstrap:bill_depth_mm   -98.126     37.010  -2.651 0.008412 **
## speciesGentoo:bill_depth_mm       23.079     34.458   0.670 0.503476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 283.9 on 324 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8757
## F-statistic: 293.4 on 8 and 324 DF,  p-value: < 2.2e-16
```

Using the anova function to compare our model to the expert's model, and a summary of the expert's model.

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ flipper_length_mm + sex + bill_length_mm + species *
##     bill_depth_mm
## Model 2: body_mass_g ~ flipper_length_mm + sex + bill_depth_mm * species
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1    324 26108473
## 2    325 26728014 -1   -619541 7.6884 0.00588 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + sex + bill_depth_mm *
##     species, data = Penguins)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -751.2 -183.8   -9.8  191.1  906.9
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1336.58     646.92  -2.066 0.039615 *
## flipper_length_mm                17.38       2.91   5.971 6.17e-09 ***
## sexmale                         432.90      44.63   9.699  < 2e-16 ***
## bill_depth_mm                    82.98      22.32   3.717 0.000237 ***
## speciesChinstrap               1460.15     680.39   2.146 0.032610 *
```

```
## speciesGentoo                     644.88     542.57   1.189 0.235481
## bill_depth_mm:speciesChinstrap   -83.53      37.01  -2.257 0.024666 *
## bill_depth_mm:speciesGentoo       36.17      34.48   1.049 0.294955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.8 on 325 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8758, Adjusted R-squared:  0.8732
## F-statistic: 327.5 on 7 and 325 DF,  p-value: < 2.2e-16
```

Prediction of penguin body mass

Our model is a linear regression model with body mass as the response, and flipper length, bill length, bill depth, species, and sex as covariates, and an interaction effect between bill depth and species.

We get the species-dependent models

$\hat{y}_{adelie} = \hat{\beta}_0 + \hat{\beta}_{fl}x_{fl} + \hat{\beta}_s x_s + \hat{\beta}_{bl}x_{bl} + \hat{\beta}_{bd}x_{bd}$

$\hat{y}_{chinstrap} = \hat{\beta}_0 + \hat{\beta}_{fl}x_{fl} + \hat{\beta}_s x_s + \hat{\beta}_{bl}x_{bl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,chinstrap})x_{bd} + \hat{\beta}_{chinstrap}$

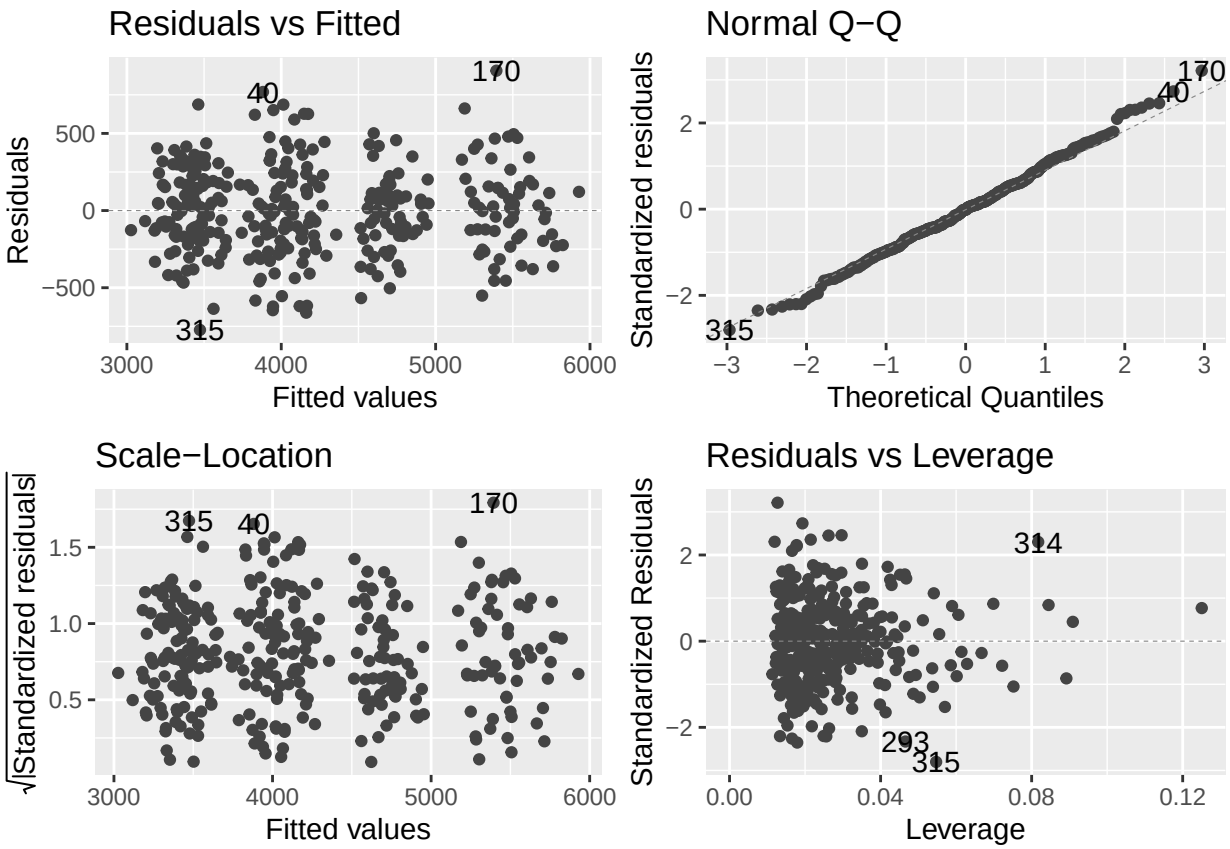
$\hat{y}_{gentoo} = \hat{\beta}_0 + \hat{\beta}_{fl}x_{fl} + \hat{\beta}_s x_s + \hat{\beta}_{bl}x_{bl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,gentoo})x_{bd} + \hat{\beta}_{gentoo}$,

where "s" stands for "sex", "fl" stands for "flipper length", "bl" stands for "bill length", and "bd" stands for "bill depth".

From the plot in b) we can see that body mass and bill length have a stronger correlation than body mass and bill depth, and bill length should therefore in theory be included if bill depth is included. Flipper length and body mass have a correlation of 0.871, and should be included. For Gentoo, the largest species, the bill is actually the smallest. Interaction between species and bill depth might therefore be a good idea, as the effect depends on the species. One could argue the same for bill length, but the trend is not as clear. Species seem to play an important part and males seem to be larger than females.

Using the anova function to compare our model to the experts model, we get a p-value of 0.00588, so it seems like our model is significantly better. Using the summary function, we see that the median of residuals of our model is -3.2, which is closer to zero than for the other model and we have an adjusted R-squared value of 0.8757, which is good.

From the coefficients in our model, we see that flipper length, sex, bill length and bill depth all have significant p-values (< 0.05), and are therefore meaningful.



The QQ-plot shows us that a linear model is a proper choice with this data. In the residual vs. fitted plot, we can see that the residuals are centered around zero, as they should be. We can see from the plot residuals vs. leverage that there are not too many points that have both high leverage and high residual value.

This plot indicates that our model fits the data well.

Problem 3

a)

All four points are in this code.

```
# Create a new boolean variable indicating whether or not the penguin is an
# Adelie penguin
Penguins$adelie <- ifelse(Penguins$species == "Adelie", 1, 0)
# Select only relevant variables and remove all rows with missing values in body
# mass, flipper length, sex or species.
Penguins_reduced <- Penguins %>% dplyr::select(body_mass_g, flipper_length_mm, adelie) %>%
mutate(body_mass_g = as.numeric(body_mass_g), flipper_length_mm = as.numeric(flipper_length_mm)) %>%
drop_na()
set.seed(4268)
# 70% of the sample size for training set
training_set_size <- floor(0.7 * nrow(Penguins_reduced))
train_ind <- sample(seq_len(nrow(Penguins_reduced)), size = training_set_size)
```

```r
train <- Penguins_reduced[train_ind, ]
test <- Penguins_reduced[-train_ind, ]

# Logistic regression
fit.glm = glm(adelie ~ body_mass_g + flipper_length_mm, family = "binomial", train)

glm.probs = predict(fit.glm, newdata = test, type ="response")
glm.preds = ifelse(glm.probs > 0.5 , 1, 0)
glm.table = table(glm.preds, test$adelie)

# QDA
fit.qda = qda(adelie ~ body_mass_g + flipper_length_mm, family = "binomial", train)
qda.preds = predict(fit.qda, newdata = test, type ="response")$class
qda.probs = predict(fit.qda,newdata=test,type="response")$posterior
qda.table = table(qda.preds, test$adelie)

# KNN
fit.knn = knn(train =train, test = test, cl = train$adelie, k=25, prob=T)
knn.probs = ifelse(fit.knn==0,1-attributes(fit.knn)$prob,attributes(fit.knn)$prob)
knn.table = table(fit.knn,test$adelie)

# Sensitivity and specificity
glm.spes= glm.table[1,1]/(glm.table[1,1]+glm.table[2,1])
glm.spes
```

```
## [1] 0.8666667
```

```r
glm.sens= glm.table[2,2]/(glm.table[2,2]+glm.table[1,2])
glm.sens
```

```
## [1] 0.9767442
```

```r
qda.spes= qda.table[1,1]/(qda.table[1,1]+qda.table[2,1])
qda.spes
```

```
## [1] 0.7666667
```

```r
qda.sens= qda.table[2,2]/(qda.table[2,2]+qda.table[1,2])
qda.sens
```

```
## [1] 0.9767442
```

```r
knn.spes= knn.table[1,1]/(knn.table[1,1]+knn.table[2,1])
knn.spes
```

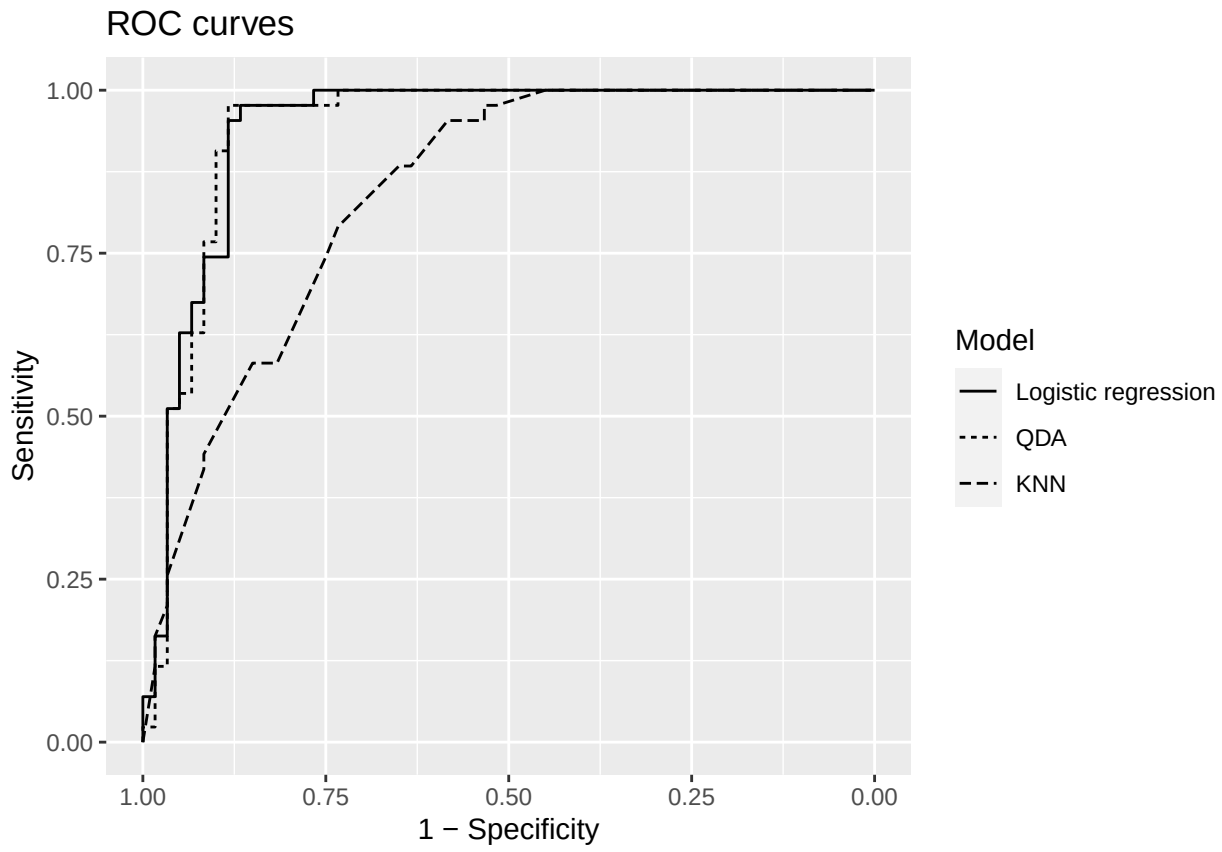
```
## [1] 0.5833333
```

```r
knn.sens= knn.table[2,2]/(knn.table[2,2]+glm.table[1,2])
knn.sens
```

```
## [1] 0.9761905
```

b)

i)



AUC for glm, qda and knn ,respectively.

```
## Area under the curve: 0.9391
```

```
## Area under the curve: 0.938
```

```
## Area under the curve: 0.8417
```

ii)

The ROC curve for the KNN model is unsatisfactory compared to the curve for the other two models. We want the curve to go as near the upper left corner as possible, so this model performs worst. The ROC curves for the logistic regression model and the quadratic discriminant analysis model are very similar. The areas under the curves are 0.9391473 and 0.9379845, respectively. Ideally we would want both to be as near 1 as possible, which means that the logistic regression model performs slightly better than the QDA method.

iii)

We would choose KNN if the task is to create an interpretable model, because it is easy to understand the concept behind it. If you want to predict which class a point belongs to, you look at some number K of the nearest points, and choose the class for which most neighbouring points belong to.

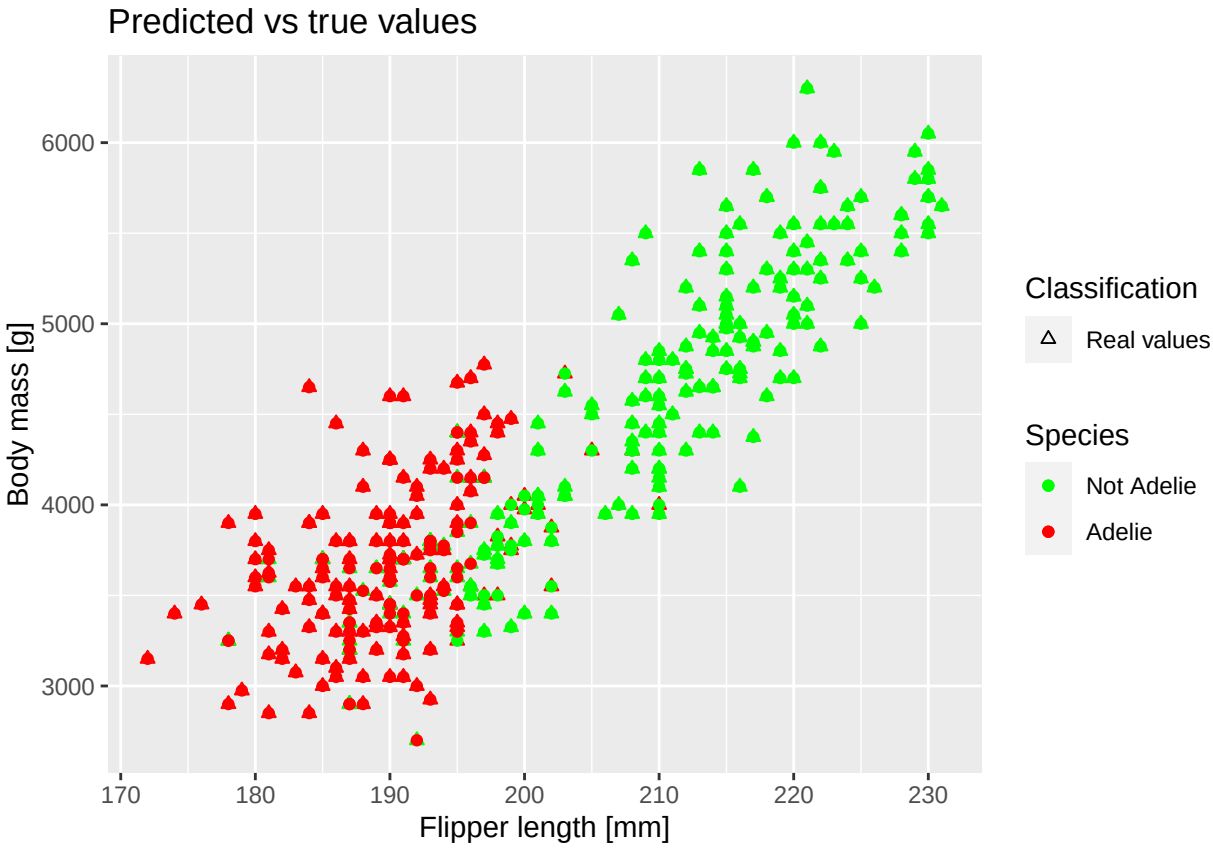
c)

iii).

d)

```
# Making predictions on both the training set and test set
glm.probs2 = predict(fit.glm, type ="response", newdata = Penguins_reduced)
glm.preds2 = ifelse(glm.probs2>0.5,1,0)

ggplot()+
  # real data
  geom_point(data = Penguins_reduced, aes(x=flipper_length_mm, y=body_mass_g, color= factor(adelie), sha
  # predicted data
  geom_point(data = Penguins_reduced, aes(x=flipper_length_mm, y=body_mass_g, color = factor(glm.preds2)
  # plot specifications
  labs(color = c("Species"))+scale_shape_identity(guide="legend", labels= c("Real values"), name="Classi
```



The true values are plotted as triangles and the predicted values as circles. The color red represents the species Adelie. Incorrect predictions are either green dots with red triangles around or red dots with green triangles around. We can see that the predictions are not always correct, which is to be expected from previous calculations.

Problem 4

a)

TRUE, FALSE, FALSE, FALSE

b)

```
id <- "1chRpybM5cJn4Eow3-_xwDKPKyddL9M2N" # google file ID
d.chd <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",id))


#logistic regression
glm.fit = glm(chd ~ sbp + sex + smoking, data = d.chd, family = "binomial")

# x_1, x_2, x_3
```

```r
sbp = 150
sex = 1
smoke = 0

# eta = beta_0 + beta_1*x_1 + beta_2*x_2 + beta_3*x_3
eta = summary(glm.fit)$coef[,1]%*%c(1,sbp,sex,smoke)
probability_chd_glm = (exp(eta)/(1+exp(eta)))
```

The probability of coronary heart disease (chd) for a non-smoking male with sbp=150 is 0.10096.

c)

```r
#parameters
B = 1000
n = 500 #how many observations in the data set.

#Here I will make a regression out of some data and find the probability from the regression
prob = function(data_set){
  glm.fit = glm(chd ~ sbp + sex + smoking, data = data_set, family = "binomial") #regression of given d
  eta = summary(glm.fit)$coef[,1]%*%c(1,150,1,0)
  answer = (exp(eta)/(1+exp(eta))) # calculating the probability
  return(answer)
}

#vector for all the probabilities
vec_of_prob = c(1:B)

#Using B = 1000 bootstrap samples
set.seed(1)  #to compare the result with others
for (i in 1:B){

 #finding "new" data set from the original data set.
  new = d.chd[sample(n,n,replace=TRUE),]

  #adding the probability of this new data set to the vector
  vec_of_prob[i] = prob(new)
}
```

Standard deviation and 96% quantile interval

```
## [1] 0.04427338
```

```
##       2.5%      97.5%
## 0.03976772 0.20609863
```

The expected probability is 0.107722 and the plausible values are between approximately 4% and 21%.

We see that a non-smoking man with a sbp of 150 has the probability of about 10% of getting coronary heart disease. You can with bigger certainty see that the probability lies between 4% and 21%, which is a pretty big interval.

d)

FALSE, FALSE, TRUE, TRUE