

Compulsory exercise 1: Group 16

TMA4268 Statistical Learning V2022

Weicheng Hua, Emil Johannese Haugstvedt, Torbjørn Baadsvik

21 February, 2022

Problem 1

a)

The expected MSE on the test set is given by:

$$\begin{aligned} E[(y_0 - \hat{f}(x_0))^2] &= E[(f(x_0) - \hat{f}(x_0) + \epsilon)^2] \\ &= E[(f(x_0) - \hat{f}(x_0))^2] + 2E[\epsilon(f(x_0) - \hat{f}(x_0))] + E[\epsilon^2] \\ &= \left\{ E[f(x_0)^2 - 2f(x_0)\hat{f}(x_0)] \right\} + \left\{ E[\hat{f}(x_0)^2] \right\} + E[\epsilon^2] \\ &= \left\{ E[f(x_0)]^2 - 2E[f(x_0)\hat{f}(x_0)] + \mathbf{E}[\hat{\mathbf{f}}(\mathbf{x}_0)]^2 \right\} + \left\{ E[\hat{f}(x_0)^2] - \mathbf{E}[\hat{\mathbf{f}}(\mathbf{x}_0)]^2 \right\} + E[\epsilon^2] \\ &= \left\{ E[f(x_0) - \hat{f}(x_0)]^2 \right\} + \left\{ E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2 \right\} + E[\epsilon^2] \\ &= E[f(x_0) - \hat{f}(x_0)]^2 + Var[\hat{f}(x_0)] + Var[\epsilon] \\ &= \text{Squared bias} + \text{Variance of prediction} + \text{Irreducible error} \end{aligned}$$

b)

The squared bias term represents the expected squared deviation between the prediction of the “true” model and the prediction of the fitted model. The variance of prediction term represents the degree to which the prediction of the fitted model can vary depending on the input. Higher variance of prediction means the model can adapt its prediction to input data to a greater extent than a simpler model, implying that the model is more flexible. However, the increased “adaptability” may be unwanted if it leads to overfitting.

c)

<i>i</i>	<i>ii</i>	<i>iii</i>	<i>iv</i>
<i>TRUE</i>	<i>FALSE</i>	<i>TRUE</i>	<i>FALSE</i>

d)

<i>i</i>	<i>ii</i>	<i>iii</i>	<i>iv</i>
<i>TRUE</i>	<i>FALSE</i>	<i>TRUE</i>	<i>FALSE</i>

e)

Answer: iii) 0.76

Problem 2

```
library(palmerpenguins) # Contains the data set "penguins".
data(penguins)
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>         <dbl>         <dbl>         <int>         <int> <fct>
## 1 Adelie  Torge~          39.1           18.7           181           3750 male
## 2 Adelie  Torge~          39.5           17.4           186           3800 fema~
## 3 Adelie  Torge~          40.3           18             195           3250 fema~
## 4 Adelie  Torge~          NA             NA             NA             NA <NA>
## 5 Adelie  Torge~          36.7           19.3           193           3450 fema~
## 6 Adelie  Torge~          39.3           20.6           190           3650 male
## # ... with 1 more variable: year <int>
```

a)

1)

Basel has not visualized the data prior to fitting the model, and has instead relied on “expert knowledge” to fit the model. This has resulted in Basel dropping the bill length covariate from the model despite not having investigated its significance in the first place.

2)

Basel has not understood the meaning of p-values. He has excluded the sex covariate as he mentioned that it has the smallest p-value. However, a small p-value may indicate that the sex covariate is significant in determining the body mass of penguins. In any case, an F-test should be conducted to determine whether the sex covariate should be omitted or kept, instead of looking directly at the p-value for the sex coefficient in the full model.

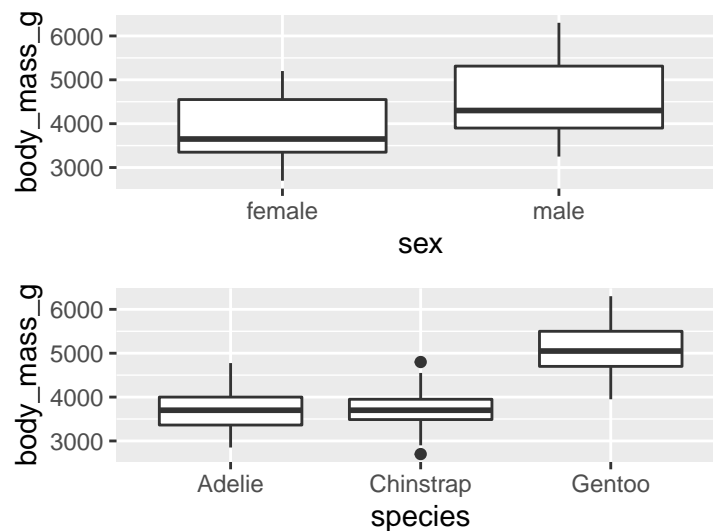
3)

Basel concludes that chinstrap penguins have the largest body mass from the fact that the coefficient for chinstrap penguins has the largest value. However, he does not consider that the only negative interaction coefficient in the model is the coefficient for the interaction between bill depth and Chinstrap. The final body mass for chinstraps can therefore be lower than the other species due to this interaction coefficient.

b)

```
library(palmerpenguins) # Contains the data set "penguins".
data(penguins)
# Remove island, and year variable, as we won't use those.
Penguins <- na.omit(subset(penguins, select = -c(island, year)))
```

```
library(patchwork)
boxplot1 <- ggplot(data=Penguins, mapping=aes(x=sex, y=body_mass_g)) + geom_boxplot()
boxplot2 <- ggplot(data=Penguins, mapping=aes(x=species, y=body_mass_g)) + geom_boxplot()
boxplot1 / boxplot2
```



Judging from the box-plot above, the sex factor appears to be significant, with males having a larger average body mass than females. Contrary to Basel's claim, we see that Gentoo is the species with the largest average body mass, not Chinstrap. ## c) Original model

```
# Fit the model as specified in advance based on "expert" knowledge:
penguin.model.initial <- lm(body_mass_g ~ flipper_length_mm + sex + bill_depth_mm * species,
                             data = Penguins)
```

Basel's model

```
penguin.model.basel <- lm(body_mass_g ~ flipper_length_mm + bill_depth_mm*species,
                             data = Penguins)
```

Improved model: We create two models: One which includes all first order terms (no interactions), and one model with interactions between bill_depth and species, as in the initial model. We then run an F-test to determine whether the more complex model improves on the simpler model:

```
# Fit improved model without interactions
penguin.model.improved0 <- lm(body_mass_g ~
                               flipper_length_mm
                               + bill_length_mm
                               + bill_depth_mm)
```

```

+ species
+ sex,
data = Penguins)

# Fit improved model wit interaction between bill depth and species
penguin.model.improved <- lm(body_mass_g ~
                             flipper_length_mm
                             + bill_length_mm
                             + bill_depth_mm
                             * species
                             + sex,
                             data = Penguins)

# compare models:
anova(penguin.model.improved0, penguin.model.improved)

## Analysis of Variance Table
##
## Model 1: body_mass_g ~ flipper_length_mm + bill_length_mm + bill_depth_mm +
##   species + sex
## Model 2: body_mass_g ~ flipper_length_mm + bill_length_mm + bill_depth_mm *
##   species + sex
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      326 26915647
## 2      324 26108473  2    807174 5.0084 0.007208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The F-test indicates that the interactions may be significant. However, since the

```

#F-test between initial model and Basel's model
anova(penguin.model.initial, penguin.model.basel)

```

```

## Analysis of Variance Table
##
## Model 1: body_mass_g ~ flipper_length_mm + sex + bill_depth_mm * species
## Model 2: body_mass_g ~ flipper_length_mm + bill_depth_mm * species
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      325 26728014
## 2      326 34464363 -1  -7736349 94.07 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We see from the p-value ($< 2.2e-16$) that the sex covariate appears to be highly significant, and that Basils model got worse (SSE increased by 7736349) by excluding it from the model.

Finally we conduct an F-test to compare our improved model with the initial model:

```

#F-Test between improved model and initial model
anova(penguin.model.improved, penguin.model.initial)

```

```

## Analysis of Variance Table
##

```

```
## Model 1: body_mass_g ~ flipper_length_mm + bill_length_mm + bill_depth_mm *
##   species + sex
## 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
```

With an F-test resulting in a p-value of 0.006, we conclude that our model is significantly better than both the initial model and Basel's model.

```
confint(penguin.model.improved)
```

```
##              2.5 %      97.5 %
## (Intercept) -3051.772630 -462.46676
## flipper_length_mm      10.177061  21.69583
## bill_length_mm         5.737942  33.76686
## bill_depth_mm         36.825559 123.85390
## speciesChinstrap      213.513770 2865.86670
## speciesGentoo        -357.924206 1756.68225
## sexmale           292.531490  478.83462
## bill_depth_mm:speciesChinstrap -170.937061 -25.31565
## bill_depth_mm:speciesGentoo    -44.709925  90.86783
```

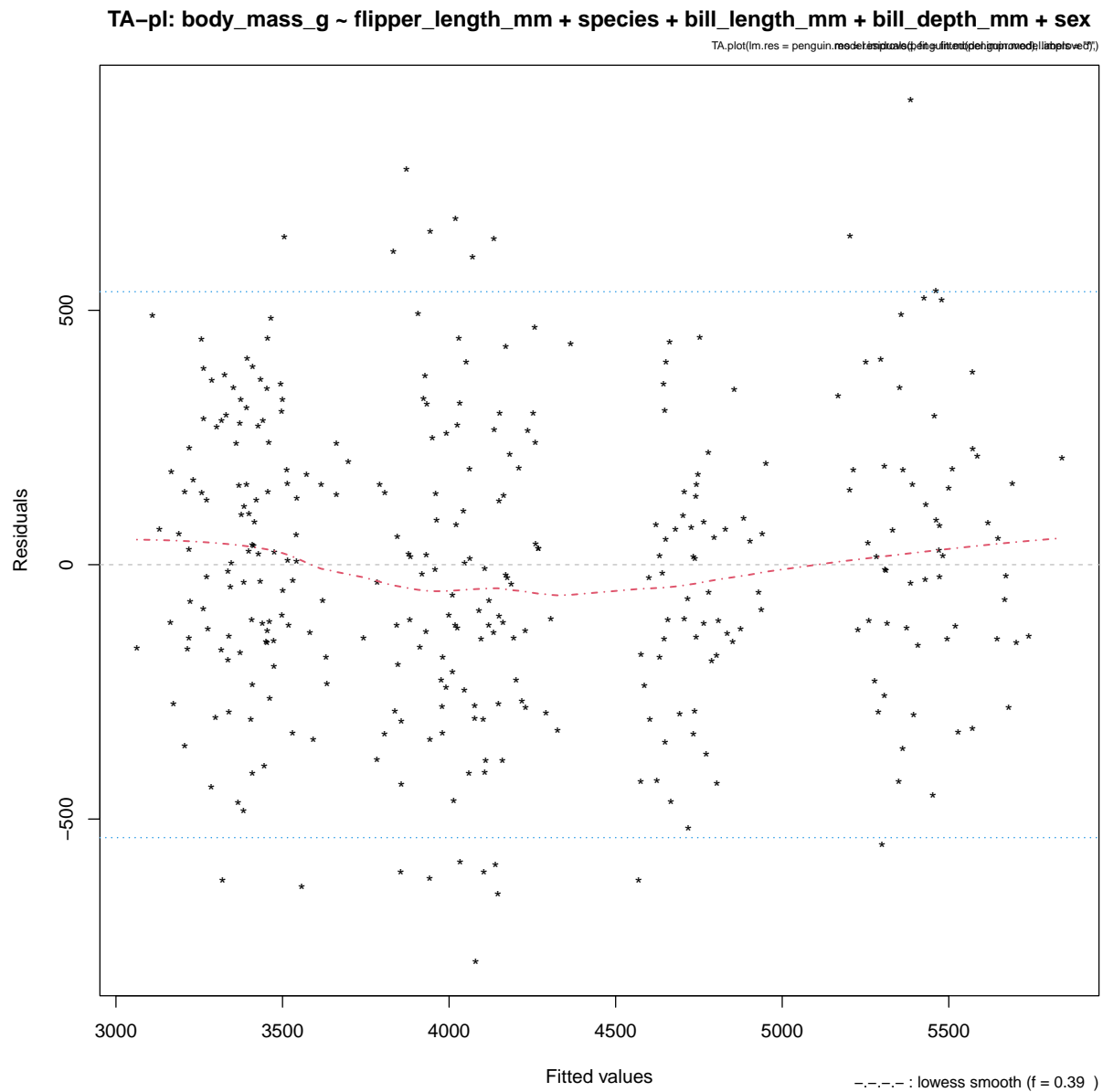
An F-test was conducted to check the importance of the interaction term between billdepth and species. This gave a F-value of 5.1625 and a $\Pr(>F)$ of 0.006193. Though this might mean that the interaction term between billdepth and species have some significance, it is the least significant and have been chosen to be omitted for simplicity. The final model without the interaction term have low p-values for all included covariates. Based on the dummy variable for Gentoo penguins being the largest (657), and no species interaction term the Gentoo penguins have in general the largest mass. However this is done assuming that the other covariates, bill_depth and flipper_length, are independent of species, which might not be the case. The final model is therefore given by the equations.

$$\begin{aligned}\hat{y}_{adelie} &= \hat{\beta}_0 + \hat{\beta}_{flipper_length}x_{flipper_length} + \hat{\beta}_{bill_depth}x_{bill_depth} + \hat{\beta}_{bill_length}x_{bill_length} \\ \hat{y}_{chinstrap} &= \hat{\beta}_0 + \hat{\beta}_{flipper_length}x_{flipper_length} + \hat{\beta}_{bill_length}x_{bill_length} + \hat{\beta}_{bill_depth}x_{bill_depth} + \hat{\beta}_{chinstrap} \\ \hat{y}_{gentoo} &= \hat{\beta}_0 + \hat{\beta}_{flipper_length}x_{flipper_length} + \hat{\beta}_{bill_length}x_{bill_length} + \hat{\beta}_{bill_depth}x_{bill_depth} + \hat{\beta}_{gentoo}\end{aligned}$$

The 95% confidence window for each of the covariate coefficient is also given above.

A Tukey-Anscombe plot to evaluate the fit of the model is generated below. No clear correlation can be seen between the residuals and fitted values, indicating that the expected value of the residual is 0 and that they all have the same variance.

```
library(sfsmisc)
TA.plot( penguin.model.improved, fit = fitted(penguin.model.improved), res = residuals(penguin.model.improved))
```



Problem 3

Load and prepare data the problem 3

```
# Load libraries
library(tidyverse)
library(class)
library(MASS)
library(palmerpenguins)
library(dplyr)
library(caret)
library(pROC)
```

```

# Load penguins data
Penguins <- penguins

# Add binary variable: if adelie
Penguins$adelie <- ifelse(Penguins$species == "Adelie", 1, 0)

# Extract just needed variables and remove na
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 for sample extraction
set.seed(4268)

# Find size of training set
training_set_size <- floor(0.7 * nrow(Penguins_reduced))

# Split data randomly into train and test
train_ind <- sample(seq_len(nrow(Penguins_reduced)), size = training_set_size)

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

```

a)

i)

```

# Make logistic regression model
model.LR <- glm(adelie ~ body_mass_g + flipper_length_mm,
               family="binomial",
               data=train)

# Get classification probabilities
pred.LR.prob <- model.LR %>% predict(test, type="response")

# Classify on test set using 0.5 cutoff
pred.LR <- ifelse(pred.LR.prob > 0.5, 1, 0)

```

ii)

```

# Make quadratic discriminant analysis model
model.QDA <- qda(adelie ~ body_mass_g + flipper_length_mm,
                 data=train)

# Get classification probabilities
pred.QDA.prob <- (model.QDA %>% predict(test))$posterior[0:nrow(test)]

```

```
# Classify on test set using 0.5 cutoff
pred.QDA <- ifelse(pred.QDA.prob > 0.5, 1, 0)
```

iii)

```
# K-nearest neighbor
model.KNN <- knn(train = train,
                 test = test,
                 cl = train$adelie,
                 k = 25,
                 prob = T)

# Get classification probabilities
pred.KNN.prob <- ifelse(model.KNN == 0, 1 - attributes(model.KNN)$prob, attributes(model.KNN)$prob)

# Classify on test set
pred.KNN <- model.KNN[0:nrow(test)]
```

iv)

```
# Calculate specificity and sensitivity
print("Sensitivity for logistic regression:")
```

```
## [1] "Sensitivity for logistic regression:"
```

```
sensitivity(as.factor(pred.LR), as.factor(test$adelie))
```

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

```
print("Sensitivity for QDA:")
```

```
## [1] "Sensitivity for QDA:"
```

```
sensitivity(as.factor(pred.QDA), as.factor(test$adelie))
```

```
## [1] 0.2333333
```

```
print("Sensitivity for KNN:")
```

```
## [1] "Sensitivity for KNN:"
```



```
sensitivity(as.factor(pred.KNN), as.factor(test$adelie))
```

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

```
print("Specificity for logistic regression:")
```

```
## [1] "Specificity for logistic regression:"
```

```
specificity(as.factor(pred.LR), as.factor(test$adelie))
```

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

```
print("Specificity for QDA:")
```

```
## [1] "Specificity for QDA:"
```

```
specificity(as.factor(pred.QDA), as.factor(test$adelie))
```

```
## [1] 0.02325581
```

```
print("Specificity for KNN:")
```

```
## [1] "Specificity for KNN:"
```

```
specificity(as.factor(pred.KNN), as.factor(test$adelie))
```

```
## [1] 0.9534884
```

b)

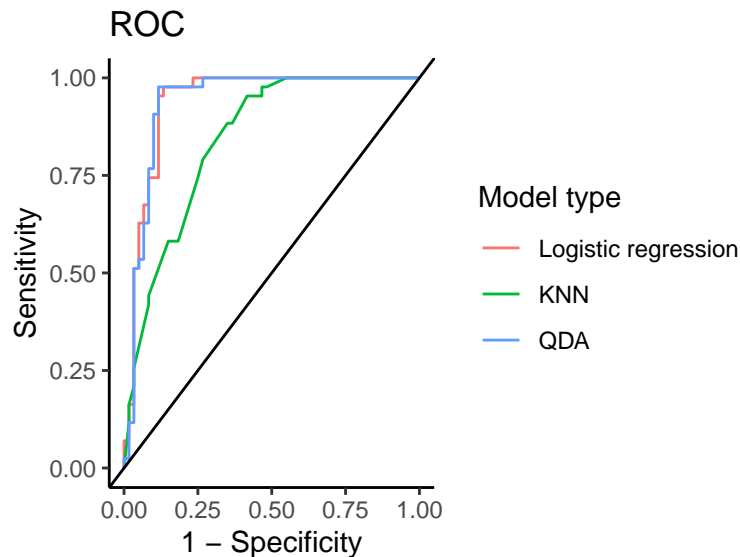
i)

```
# ROC for different classifiers
roc.LR <- roc(test$adelie, pred.LR.prob)
roc.KNN <- roc(test$adelie, pred.KNN.prob)
roc.QDA <- roc(test$adelie, pred.QDA.prob)

roc.list <- list(
  "Logistic regression" = roc.LR,
  "KNN" = roc.KNN,
  "QDA" = roc.QDA)

# Plot ROC
ggroc(roc.list, aes = "col", legacy.axes = TRUE) +
  geom_abline() +
```

```
theme_classic() +
ggtitle("ROC") +
labs(x = "1 - Specificity",
     y = "Sensitivity",
     col = "Model type")
```



```
# AUC for the classifiers
roc.LR$auc
```

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

```
roc.KNN$auc
```

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

```
roc.QDA$auc
```

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

ii)

The ROC curve shows how the binary classifiers are performing with a varying threshold. The plot shows the true positive rate, or sensitivity, against the true negative rate, or 1 - specificity. If a classifier is on the black diagonal line it performs just as good as guessing randomly and if it lays above the diagonal line it performs better. This means that a classifier is better the further up in the left corner its ROC curve are. In this case the logistic regression and the QDA performs better than the KNN classifier.

The AUC, or area under the curve, is just the area under the ROC curve and provides an aggregate measure of all classifiers for all thresholds. As expected, the AUC for the logistic regression and the QDA are higher than the one for the KNN indicating better performance.

iii)

We would say the logistic regression model is the most interpretable model. This model provides easy to understand betas telling about the relationship between the body mass and flipper length and if the penguin is Adelie or not. The two other models does not provide the same information about the data.

c)

```
model.LR$coefficients
```

```
##      (Intercept)      body_mass_g flipper_length_mm
##      37.7618776      0.0007120      -0.2055804
```

Increasing the body mass of penguin with 1000g will result in the following change in the odds

$$\exp(0.0007120 * 1000) = 2.038$$

The odds increases by a factor of 2.038. Alternative iii.

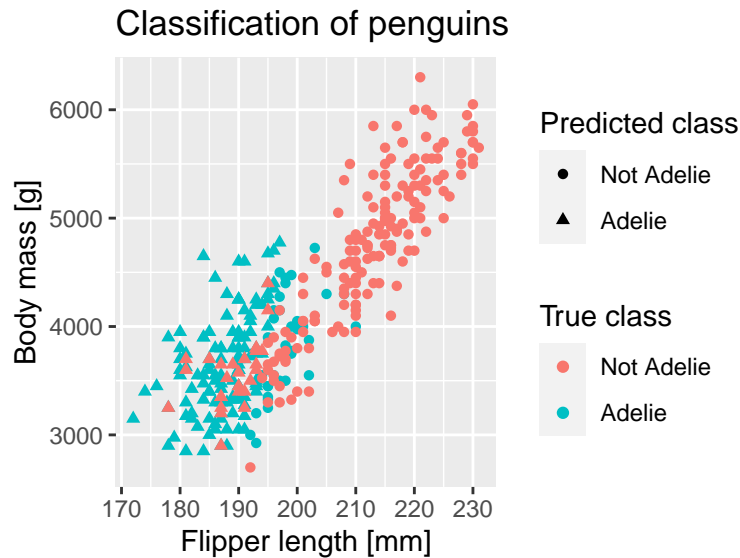
d)

```
# Get all data, from both test and training
penguins.pred.all <- Penguins_reduced %>% dplyr::select(-adelie)

# Classify on both test and training data using LR model
pred.all <- ifelse(model.LR %>% predict(penguins.pred.all) > 0.5, 1, 0)

penguins.pred.all <- Penguins_reduced %>% dplyr::mutate(pred = pred.all)

# Plot result
ggplot(penguins.pred.all, aes(x=flipper_length_mm, y=body_mass_g)) +
  ggtitle("Classification of penguins") +
  geom_point(aes(col = as.factor(adelie), shape = as.factor(pred))) +
  labs(x = "Flipper length [mm]",
       y = "Body mass [g]",
       col = "True class",
       shape = "Predicted class") +
  scale_colour_discrete(labels=c("Not Adelie", "Adelie")) +
  scale_shape_discrete(labels=c("Not Adelie", "Adelie"))
```



Problem 4

a)

- (i) True. The process for validation set approach is only necessary to be repeated once but 10 times in 10-fold CV. (ii) False. LOOCV has the highest variance as the datasets used between training are highly correlated and differ only by one observation which can lead to high variance between completely new datasets. (iii) False. In the validation set-approach the data is split randomly into 2 equal sets with 1 set being the training set and the other being the validation set. In the 2-fold CV the data is again randomly split into 2 equal parts but each data set take turn being the training and validation set.
- (ii) False. LOOCV is the most computationally expensive way to do cross-validation.

b)

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

```
m <- glm(family="binomial", formula = chd ~ sbp + sex + smoking, data=d.chd)
d.chd.new <- data.frame(
  sex=as.integer(c(1)),
  sbp = as.numeric(c(150)),
  smoking=as.integer(c(0)))
pred <- predict(m, d.chd.new, type="response")
pred
```

```
##          1
## 0.10096
```

The probability of chd for a non-smoking male with sbp=150 is 10%.

c)

1)

```
set.seed(4268)
library(boot)

prob <- function(df, index){
  m <- glm(family="binomial", subset=index, formula = chd ~ sbp + sex + smoking, data=d.chd)
  return(predict(m, d.chd.new, type="response"))
}
```

2)

```
B <- 1000
boot.result <- boot(d.chd, prob, B)
boot.result

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d.chd, statistic = prob, R = B)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1*  0.10096 0.003454964   0.0442774
```

From the bootstrapping method we observe an estimated standard error of 0.044.

3)

```
boot.ci(boot.result, 0.95)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.result, conf = 0.95)
##
## Intervals :
## Level      Normal          Basic
## 95%  ( 0.0107, 0.1843 )  (-0.0040, 0.1626 )
##
## Level      Percentile      BCa
## 95%  ( 0.0393, 0.2059 )  ( 0.0425, 0.2210 )
## Calculations and Intervals on Original Scale
```

From the bootstrapping method we obtain $[0.0107, 0.1843]$ as the first order normal approximation of the 95 % CI.

4)

Since the 95 % CI is wide (nearly approaching zero in the left-hand limit) we conclude based on the bootstrapping computations that the conditional probability of chd in a non-smoking male with sbp=150 is quite uncertain. Plausible values for conditional chd probability thus lie in the interval $[0.0107, 0.1843]$. The upper limit of .1843 is perhaps most useful as it provides an upper 97.5 % confidence bound on estimated conditional chd risk.

d)

<i>i</i>	<i>ii</i>	<i>iii</i>	<i>iv</i>
<i>FALSE</i>	<i>FALSE</i>	<i>TRUE</i>	<i>TRUE</i>