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:

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

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 it's 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	ii	iii	iv
$TRUE$	$FALSE$	$TRUE$	$FALSE$

d)

i	ii	iii	iv
$TRUE$	$FALSE$	$TRUE$	$FALSE$

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 it's 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.

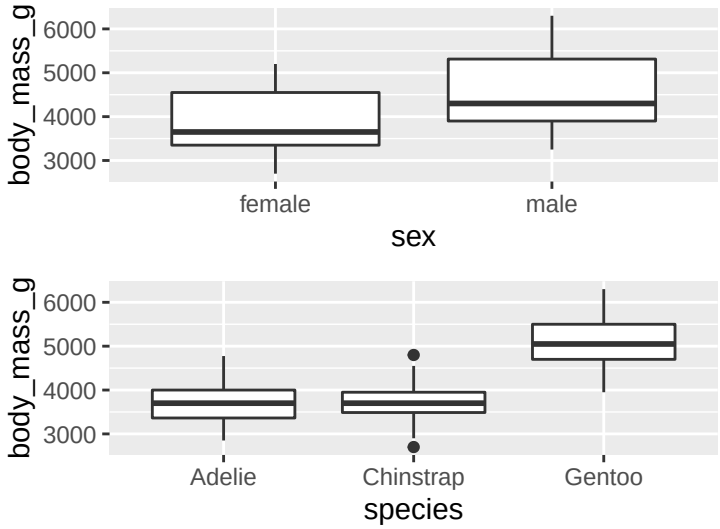
3)

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
```



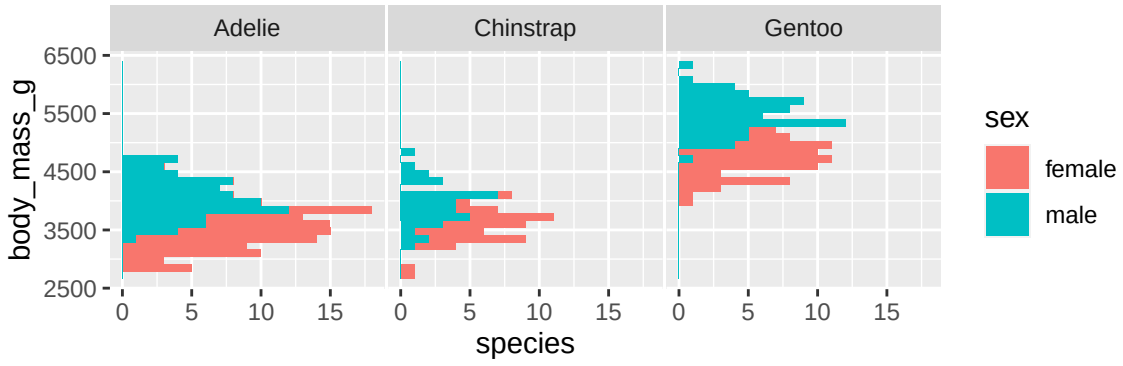
Judging from the box-plot above, the sex factor appears to be significant, with males having a larger average body mass than females. Contrary to Basel's claim, we see that Gentoo is the species with the largest average body mass, not Chinstrap. ## c)

```
library(GGally)
library(patchwork)

disc_plots <- function(){
  gpairs <- ggpairs(Penguins, mapping = aes(colour = sex))
  mass_pos <- match("body_mass_g", gpairs$xAxisLabels)
  getPlot(gpairs, i=mass_pos, j=1)
}

cont_plots <- function(colour){
  if(colour=="species"){
    gpairs <- ggpairs(Penguins, mapping = aes(colour = species))
  } else {
    gpairs <- ggpairs(Penguins, mapping = aes(colour = sex))
  }
  ind <- function(var_name){
    match(var_name, gpairs$xAxisLabels)
  }
  cont_vars <- c("bill_length_mm", "bill_depth_mm", "flipper_length_mm")
```

```
  cont_inds <- sapply(cont_vars, function(var) ind(var))
  mass_pos <- match("body_mass_g", gpairs$xAxisLabels)
  plots <- lapply(cont_inds, function(j) getPlot(gpairs, i = mass_pos, j = j))
  ggmatrix(
      plots,
      byrow=TRUE,
      nrow = 1,
      ncol = length(cont_inds),
      xAxisLabels = c(cont_vars),
      yAxisLabels = c("body_mass_g"),
      title = "Scatter plots",
      showStrips = TRUE,
      legend=3,
      gg=theme(axis.text=element_text(size=12))
  )
}
```

Plots of body-mass distributions for the different species and sexes:

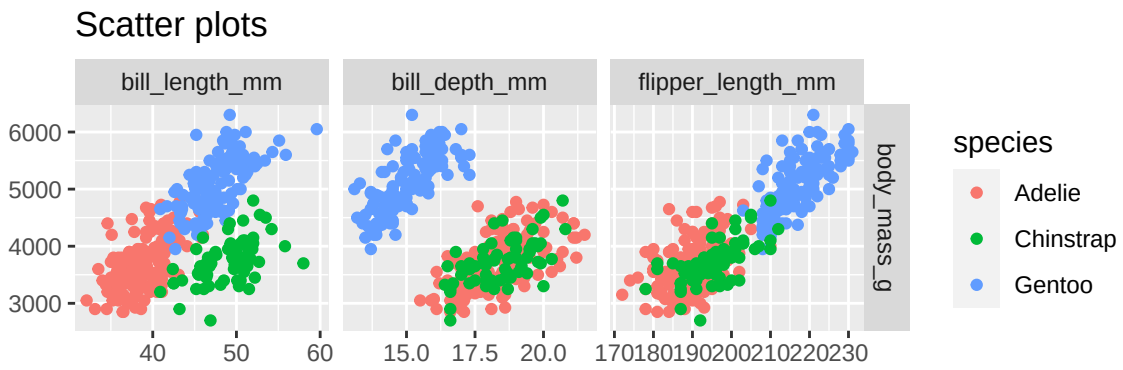
```
disc_plots()
```



From the plot above we observe that males on average have a larger body mass than females. However, the most significant difference in body mass is found between the Gentoo species and the two other species.

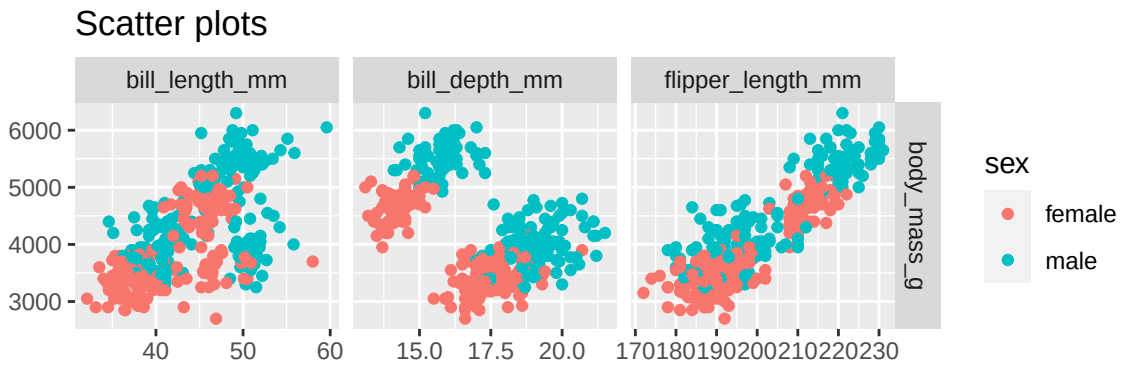
Scatter plot for each continuous variable categorized by species:

```
cont_plots("species")
```



Scatter plot for each continuous variable categorized by sex:

```
cont_plots("sex")
```



We observe from the two scatter plots that the most significant differences in attributes are found between different species, not the sexes. However, the sex scatter plot indicates that males tend to score higher in most attributes as the clusters are centered more upward and to the left than females in the plot. That being said, we see no clear evidence that the regression coefficient for any of the 3 continuous variables differ significantly among species or sexes. This motivates us to disregard interactions.

Initial 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)
```

```
#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 ($< 2.2e\text{-}16$) 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.

We begin by creating a simple model

```r
# Fit improved model without interactions
penguin.model.improved <- lm(body_mass_g ~
                               flipper_length_mm
                             + bill_length_mm
                             + bill_depth_mm
                             + species
                             + sex,
                             data = Penguins)
summary(penguin.model.improved)$coefficient
```

```
##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       -1460.99463 571.308144 -2.557280 1.100170e-02
## flipper_length_mm    15.95025   2.909612  5.481915 8.440786e-08
## bill_length_mm       18.20443   7.106258  2.561746 1.086405e-02
## bill_depth_mm        67.21763  19.741850  3.404829 7.447574e-04
## speciesChinstrap   -251.47669  81.078824 -3.101632 2.092677e-03
## speciesGentoo      1014.62666 129.560586  7.831291 6.852265e-14
## sexmale             389.89153  47.848346  8.148485 7.971622e-15
```

We see that all coefficients in the simpler model appear significant with a p-value below 0.05.

```r
# Fit improved model wit interaction between bill depth and species
penguin.model.improved2 <- lm(body_mass_g ~
                                flipper_length_mm
                              + bill_length_mm
                              + bill_depth_mm
                              * species
                              + sex,
                              data = Penguins)
summary(penguin.model.improved2)$coefficient
```

```
##                                    Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                     -1757.11969 658.081905 -2.6700623 7.966120e-03
## flipper_length_mm                  15.93645   2.927539  5.4436325 1.032568e-07
## bill_length_mm                     19.75240   7.123657  2.7727899 5.879505e-03
## bill_depth_mm                      80.33973  22.118582  3.6322278 3.265393e-04
## speciesChinstrap                 1539.69023 674.105553  2.2840492 2.301521e-02
## speciesGentoo                     699.37902 537.435249  1.3013270 1.940711e-01
## sexmale                           385.68305  47.349646  8.1454263 8.278370e-15
## bill_depth_mm:speciesChinstrap    -98.12636  37.010233 -2.6513304 8.411605e-03
## bill_depth_mm:speciesGentoo        23.07895  34.457600  0.6697783 5.034762e-01
```

The coefficient for the interaction between bill_depth and speciesGentoo appears to be insignificant with p-value > 0.5.

```r
# compare models:
anova(penguin.model.improved, penguin.model.improved2)
```

```
## 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 with a p-value of 0.0072. However due to the large p-value for interaction coefficient mentioned before, we elect to keep the simpler model. 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    326 26915647
## 2    325 26728014  1    187632 2.2815 0.1319
```

Although the F-test suggests an insignificant difference in performance for the two models, we still prefer our model since it does not include interaction terms. A summary of the improved model is given below:

```
summary(penguin.model.improved)
```

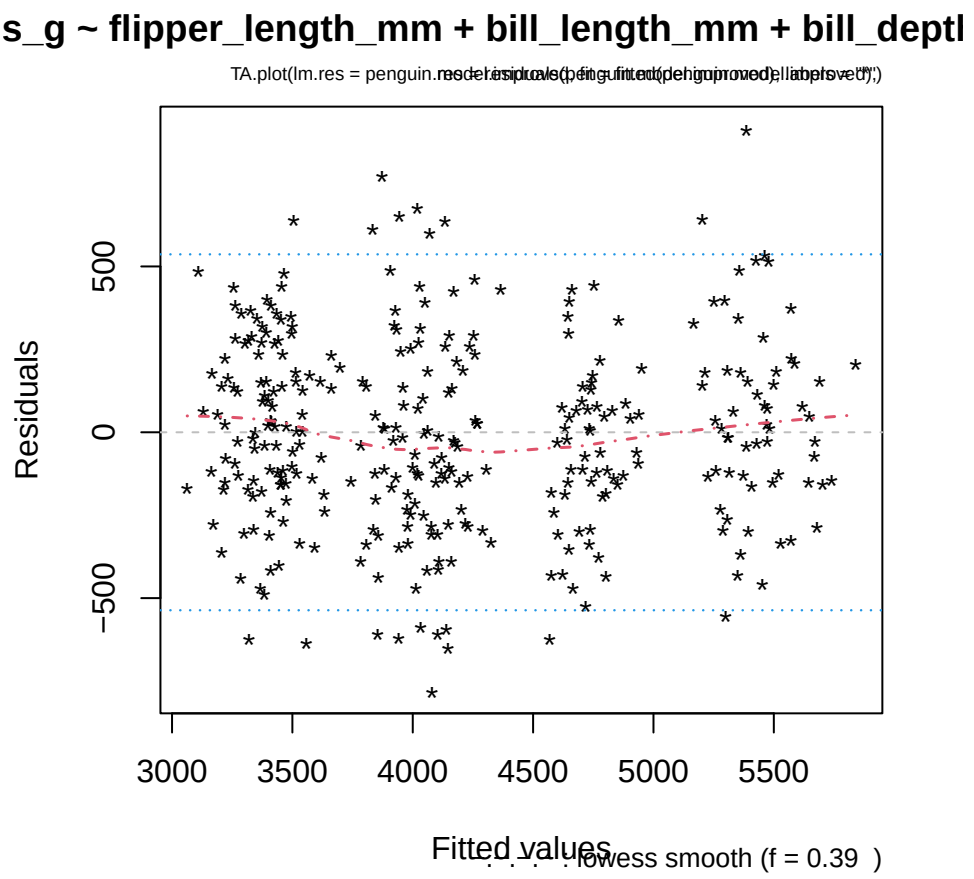
```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm +
##     bill_depth_mm + species + sex, data = Penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -779.65 -173.18   -9.05  186.61  914.11
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1460.995    571.308  -2.557 0.011002 *
## flipper_length_mm    15.950      2.910   5.482 8.44e-08 ***
## bill_length_mm       18.204      7.106   2.562 0.010864 *
## bill_depth_mm        67.218     19.742   3.405 0.000745 ***
## speciesChinstrap   -251.477     81.079  -3.102 0.002093 **
## speciesGentoo      1014.627    129.561   7.831 6.85e-14 ***
## sexmale             389.892     47.848   8.148 7.97e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 287.3 on 326 degrees of freedom
## Multiple R-squared:  0.875,  Adjusted R-squared:  0.8727
## F-statistic: 380.2 on 6 and 326 DF,  p-value: < 2.2e-16
```

The final model is given by the following equations.

$$\hat{y}_{adelie} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{bill\_length} x_{bill\_length}$$
$$\hat{y}_{chinstrap} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{bill\_length} x_{bill\_length} + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{chinstrap}$$
$$\hat{y}_{gentoo} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{bill\_length} x_{bill\_length} + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{gentoo}$$

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. The variance of residuals does not appear to be significantly different for the fitted values either, suggesting heteroskedasticity is not present.

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



Problem 3

Load and prepare data the problem 3

```r
# 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)

```r
# 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))
```

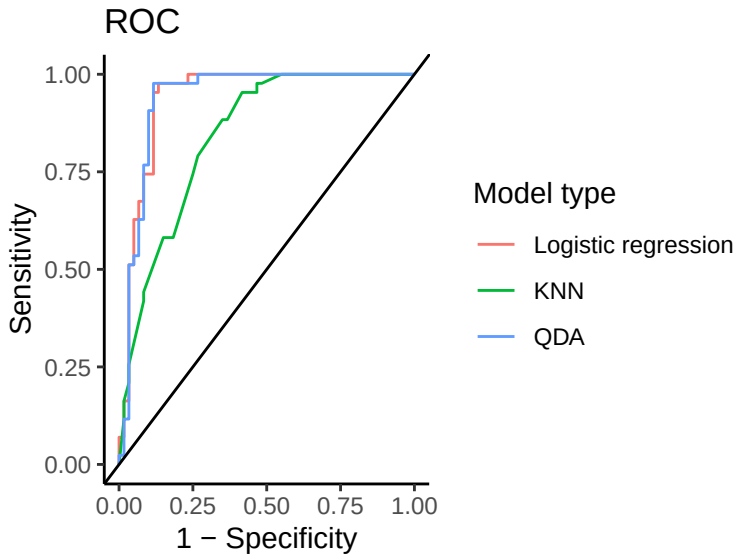
```
## [1] 0.9534884
```

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 realtionship 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
```

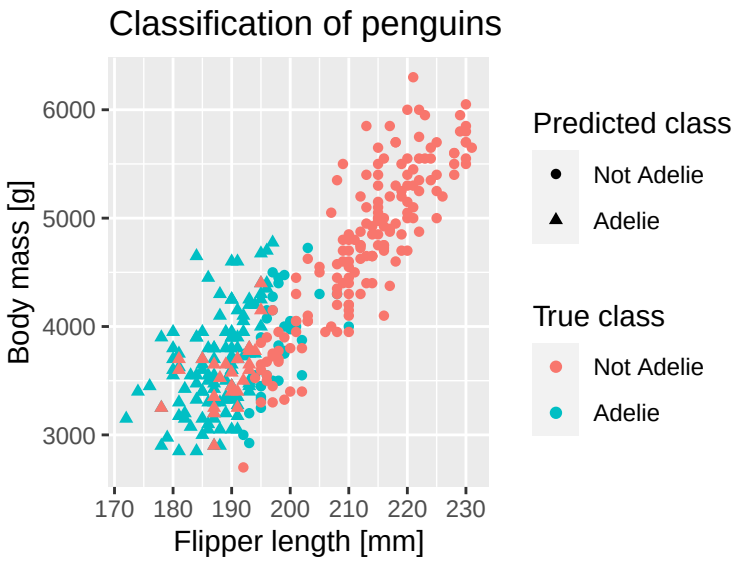
Increasing the body mass of penguin with 1000g will result in the following change in the odds

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

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	*ii*	*iii*	*iv*
FALSE	*FALSE*	*TRUE*	*TRUE*