Compulsory exercise 1: Group 38

TMA4268 Statistical Learning V2022

Lucas Michael Cammann, Ruben Mustad and Michinori Asaka

08 august, 2022

Problem 1

a)

We are now interested in decomposing the expected test MSE into three terms, the bias, variance and irreducible error. We have that

$$E[(y_0 - \hat{f}(x_0))] = E[y_0 - f(x_0) + f(x_0) - \hat{f}(x_0)] = E[(y_0 - f(x_0))^2] + E[(f(x_0) - \hat{f}(x_0))^2] + 2E[(y_0 - f(x_0))(f(x_0) - \hat{f}(x_0))]$$

By looking at the last term, we use that f is deterministic and $E[y_0] = f(x_0)$, therefore

$$E[(y_0 - f(x_0))(f(x_0) - \hat{f}(x_0))] = 0. \tag{1}$$

We therefore obtain the expression

$$E[(y_0 - \hat{f}(x_0))] = E[(y_0 - f(x_0))^2] + E[(f(x_0) - \hat{f}(x_0))^2]$$

where the first term on the R.H.S is the irreducible term $\text{Var}(\epsilon)$ (as $y_0 = f(x_0) + \epsilon$ and $\text{Var}(\epsilon) = E[\epsilon]^2$) and the second term on the R.H.S is the reducible term. To find an expression for the variance and bias, we expand the reducible term, i.e.,

$$E[(f(x_0) - \hat{f}(x_0))^2] = E[f(x_0)^2] + E[\hat{f}(x_0)^2] - 2E[f(x_0)\hat{f}(x_0)]$$

Now, to get any further, we use that $E[X^2] = \text{Var}[X] + E[X]^2$, so we can write

$$E[(f(x_0) - \hat{f}(x_0))^2] = f(x_0)^2 + \text{Var}(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 - 2f(x_0)E[\hat{f}(x_0)],$$

where we used that $E[f(x_0)] = f(x_0)$. We therefore end up with

$$E[(y_0 - \hat{f}(x_0))] = \text{Var}(\hat{f}(x_0)) + (f(x_0) - E[\hat{f}(x_0)])^2 + \text{Var}(\epsilon),$$

where the terms on the RHS are the variance, squared bias and irreducible error, respectively.

b)

- Irreducible error: The irreducible error occurs due to the noise in the data. No matter what algorithm is used, it will be present, hence the name irreducible.
- Variance: Variance tells us how our model can adjust to new data. For example, a high variance will indicate that we have overfitted our model (to the noise in the data), and our performance on new data might not be as good.
- Bias: Bias tells us the difference between our predicted values and the actual value. A high bias will indicate that the predicted value is far away from the true value (we are underfitting our model), while a low bias indicates that we our predicted value is close to the true value.

c)

1. Decreased K corresponds to increased flexibility of the model: True
2. The variance increases with increased value of K: False
3. The blue line corresponds to the irreducible error: True
4. The squared bias decreases with increased value of K: False

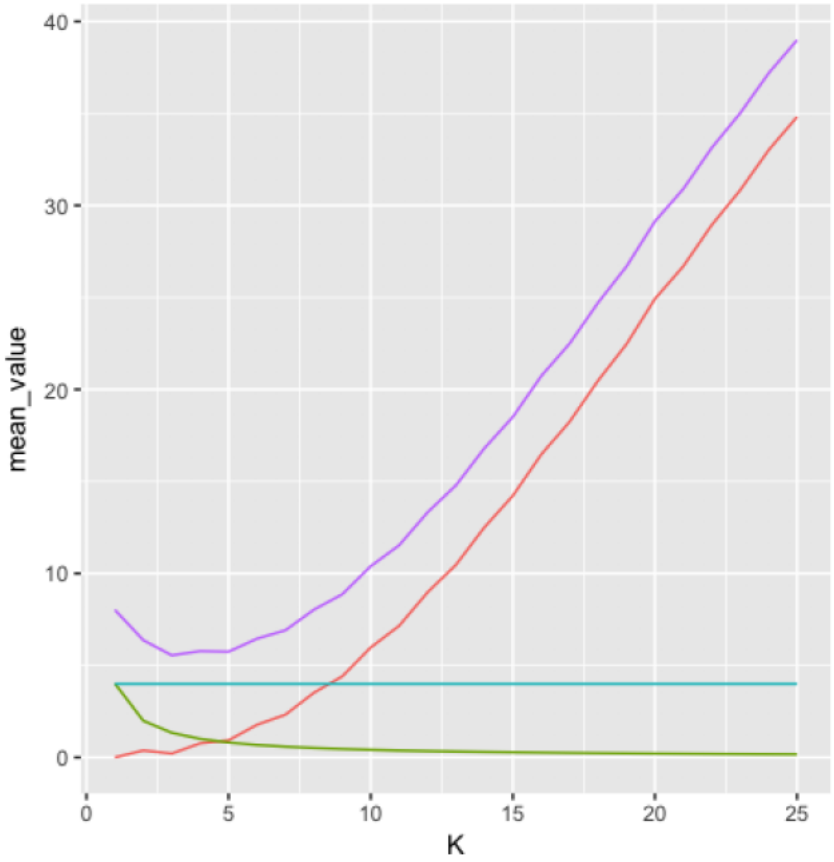


Figure 1: Image for Task 1c

d)

1. If the relationship between the predictors and response is highly non-linear, a flexible method will generally perform better than an inflexible method: False
2. If the number of predictors p is extremely large and the number of observations n is small, a flexible method will generally perform better than an inflexible method: False
3. In KNN classification, it is important to use the test set to select the value K, and not the training set, to avoid overfitting: False
4. In a linear regression setting, adding more covariates will reduce the variance of the predictor function: False

e)

Let $X = [x_1, x_2, x_3]^T$ be a 3-dimensionl random vector with covariance matrix

$$\Sigma = \begin{bmatrix} 50 & 33 & 18 \\ 33 & 38 & -10 \\ 18 & -10 & 72 \end{bmatrix} \tag{2}$$

The correlation between element x_1 and x_2 of the vector X is 0.76.

Problem 2

```
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)

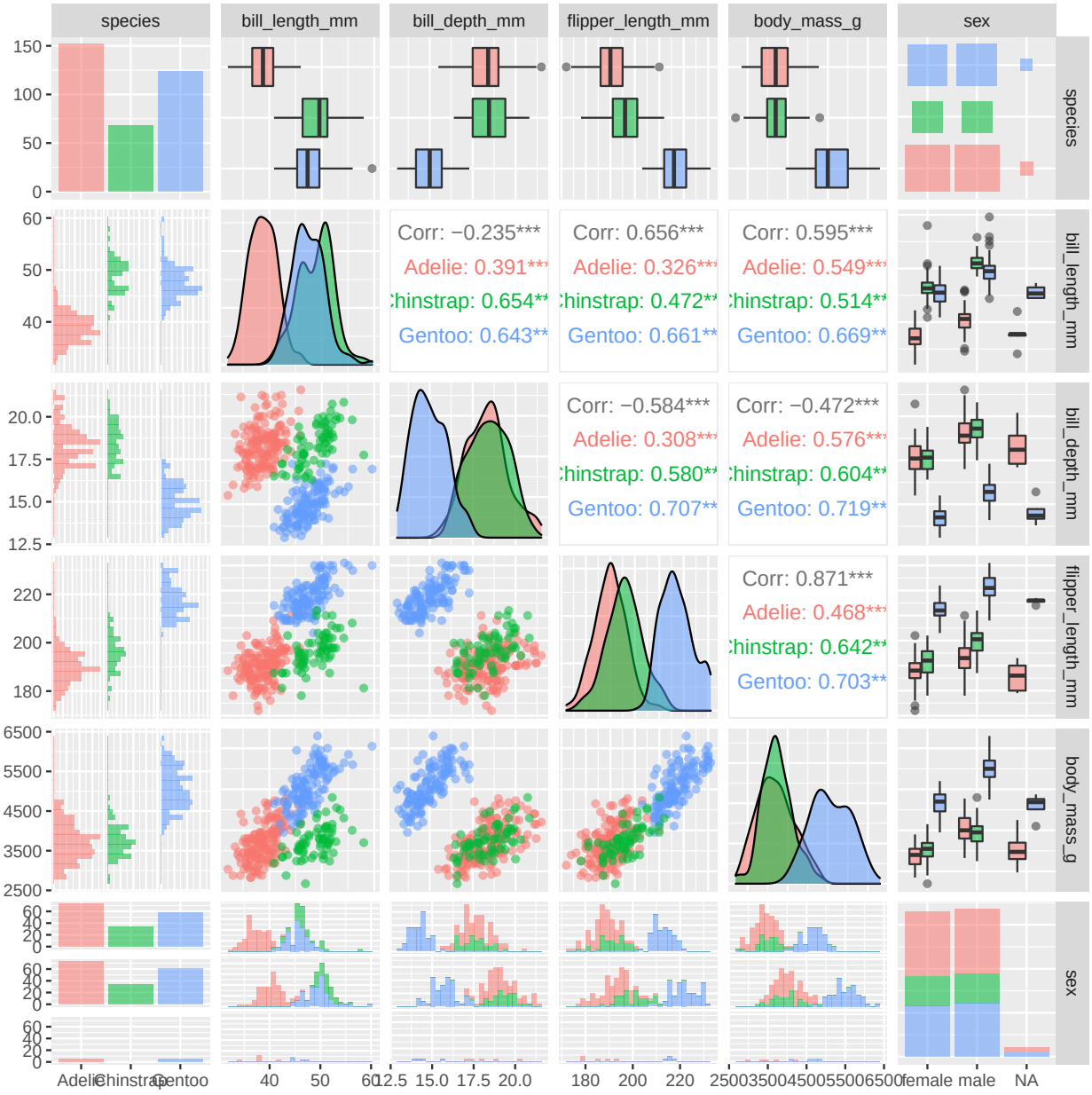
The following points are considered flaws in Basil's analysis:

1. The Sex of the penguins should have not been excluded from the analysis. On the contrary, the very low p-value found in Basil's analysis suggests that we can reject the Null-hypothesis, and that there *is* a very significant correlation between the sex and the body mass of the penguins.
2. Basil asserts that the chinstrap penguins have the largest body mass, based on the estimated intercept $\hat{\beta}_{chinstrap}$. This is however a fallacy, as the linear model requires both a slope *and* an intercept. Within Basil's model, the coefficient $\hat{\beta}_{bill_depth:chinstrap}$ is negative, meaning that overall the slope w.r.t. the bill depth is smaller than for the other species, correcting for the larger offset. Analyzing the graphical results of section b) it can be appreciated that actually Gentoo penguins have, on average, the highest body mass.
3. Basil claims that the Null-hypothesis for the species interaction cannot be rejected on basis of the p-values of the species coefficients. This is however not a permissible conclusion, as the null-hypothesis of the underlying t-test does not test whether all species coefficients are simultaneously zero (we could only use the t-test in this manner if we were dealing with only two species). Using the `anova` function Basil would have seen that the F-statistic suggests that the species is a highly significant covariate.

b)

We use the `ggpairs` function to retrieve a graphical overview of the data. We highlight both the species, as well as the sex in two distinct plots.

```
Penguins <- subset(penguins, select = -c(island, year))
ggpairs(Penguins, aes(colour = species, alpha = 0.75))
```




```
ggpairs(Penguins, aes(colour = sex, alpha = 0.75))
```

It is apparent that the Gentoo penguins have, on average, the highest body mass (as teased in a)), while the weight distribution for the other two species seems to be quite similar. Furthermore, we can appreciate from the second figure that the weight distribution for males of the same species is shifted to higher body masses, showing that sex is indeed an important predictor.

c)

Based on the knowledge from our graphical analysis that both the sex as well as the species influence the body weight, we set up a "naive" approach, in which we use both as interaction terms for all measured phenotypical covariates (i.e. bill length, bill depth and flipper length).

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

```
## Analysis of Variance Table
##
## Response: body_mass_g
##                                Df   Sum Sq  Mean Sq  F value    Pr(>F)
## bill_depth_mm                   1 47959592 47959592 623.3725 < 2.2e-16 ***
## bill_length_mm                  1 52666666 52666666 684.5545 < 2.2e-16 ***
## flipper_length_mm               1 63818497 63818497 829.5045 < 2.2e-16 ***
## species                         2 18417241  9208620 119.6924 < 2.2e-16 ***
## sex                             1  5482024  5482024  71.2546 1.234e-15 ***
## bill_depth_mm:species           2   807174   403587   5.2458  0.005749 **
## bill_length_mm:species          2   262431   131216   1.7055  0.183383
## flipper_length_mm:species       2    25651    12826   0.1667  0.846525
## bill_depth_mm:sex               1   409767   409767   5.3261  0.021669 *
## bill_length_mm:sex              1   424492   424492   5.5175  0.019457 *
## flipper_length_mm:sex           1    37238    37238   0.4840  0.487131
## species:sex                     2   489389   244694   3.1805  0.042929 *
## bill_depth_mm:species:sex       2    31154    15577   0.2025  0.816818
## bill_length_mm:species:sex      2     3730     1865   0.0242  0.976053
## flipper_length_mm:species:sex   2   651493   325747   4.2340  0.015344 *
## Residuals                     309 23773127    76936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table reveals that all phenotypical covariates as well as species and sex are highly relevant in predicting penguin weight, as the Pr(>F) is $< 10^-10$ for each. Only two double interactions terms have a Pr(>F) value of < 0.05 (`species:sex` and `flipper_length_mm:species:sex`), justifying the choice of a simpler model. The most significant interaction term is `bill_depth_mm:species`, giving us confidence in the initial expert model to which we will resort in the following. The proposed model is written as follows:

$$\hat{y}_{adelie,f} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + \hat{\beta}_{bd} x_{bd} \tag{3}$$

$$\hat{y}_{adelie,m} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + \hat{\beta}_{bd} x_{bd} + \hat{\beta}_m \tag{4}$$

$$\hat{y}_{chinstrap,f} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,chinstrap}) + \hat{\beta}_{chinstrap} \tag{5}$$

$$\hat{y}_{chinstrap,m} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,chinstrap}) + \hat{\beta}_{chinstrap} + \hat{\beta}_m \tag{6}$$

$$\hat{y}_{gentoo,f} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,gentoo}) + \hat{\beta}_{gentoo} \tag{7}$$

$$\hat{y}_{gentoo,m} = \hat{\beta}_0 + \hat{\beta}_{fl} x_{fl} + (\hat{\beta}_{bd} + \hat{\beta}_{bd,gentoo}) + \hat{\beta}_{gentoo} + \hat{\beta}_m \tag{8}$$

$$\tag{9}$$

in which the subscripts `fl, bd` and `m` stand for flipper length, bill depth and male sex, respectively. We fit the expert model and analyse the results.

```
expert.model <- lm(body_mass_g ~ flipper_length_mm + sex + bill_depth_mm * species,
    data = Penguins)
sigma <- summary(expert.model)$sigma
rsq <- summary(expert.model)$r.squared
adjrsq <- summary(expert.model)$adj.r.squared
summary(expert.model)
```

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

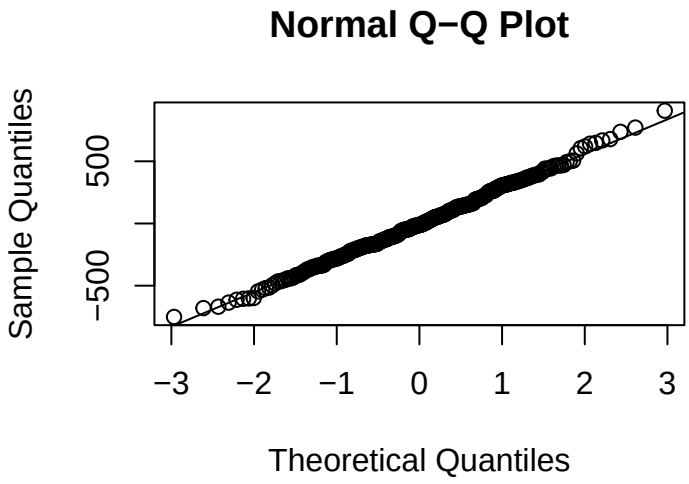
```
## Residuals:
##    Min     1Q Median     3Q    Max
## -751.2 -183.8   -9.8  191.1  906.9
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1336.58     646.92  -2.066 0.039615 *
## flipper_length_mm                17.38       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
```

```
anova(expert.model)
```

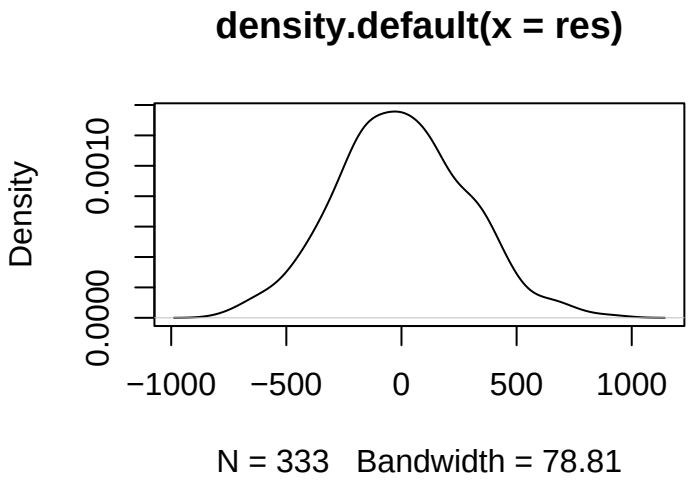
```
## Analysis of Variance Table
##
## Response: body_mass_g
##                         Df    Sum Sq   Mean Sq   F value    Pr(>F)
## flipper_length_mm        1 164047703 164047703 1994.7424 < 2.2e-16 ***
## sex                      1   9416589   9416589  114.5013 < 2.2e-16 ***
## bill_depth_mm            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
```

The expert model seems to have increased the accuracy in our predictions as compared to Basil's model, with a now higher adjusted R^2 of 0.8731593 and slightly lower residual standard error of 286.7752505‘. Based on the ANOVA table and the results of the t-test we can reject the null hypothesis that any of our covariates is statistically insignificant (i.e., that any of the corresponding coefficients is zero). As stated in subquestion a), the results of the ANOVA table are required for our hypothesis testing concerning the species, as we are dealing with three distinct groups there. In contrast to Basil, we do not attempt to infer any meaning into the absolute values of the coefficients, as this would be a slippery slope considering the various levels and interactions. Instead, we resort to assessing the model fit and whether we have introduced any systematic error in our model. To do so, we analyse the distribution of the residuals:

```
res <- resid(expert.model)
qqnorm(res)
qqline(res)
```




```
plot(density(res))
```




```
meanres <- mean(res)
meanres
```

```
## [1] -7.258583e-15
```

In the QQ-plot, the data plots seem to lie on a 45-degree straight line, indicating that our data is normally distributed. Only in the low and high theoretical quantile regions (i.e. below -2 and over 2) do some outliers stray from this line. Overall, the plot however suggests that the residuals are normally distributed, and that our model fits the data well. This conclusion is fortified by the plot of the probability density of the residuals, which looks like a bell-shape curve. Although there seems to be a slight skewness to the curve, the calculated mean of the residuals turn out to be extremely close to zero at 0. Therefore, we conclude that our residuals are normally distributed, and that our model represents the data very well.

Problem 3

a) (5P)

1) Fit a logistic regression model using the training set, and perform the classification on the test set, using a 0.5 cutoff (1P).

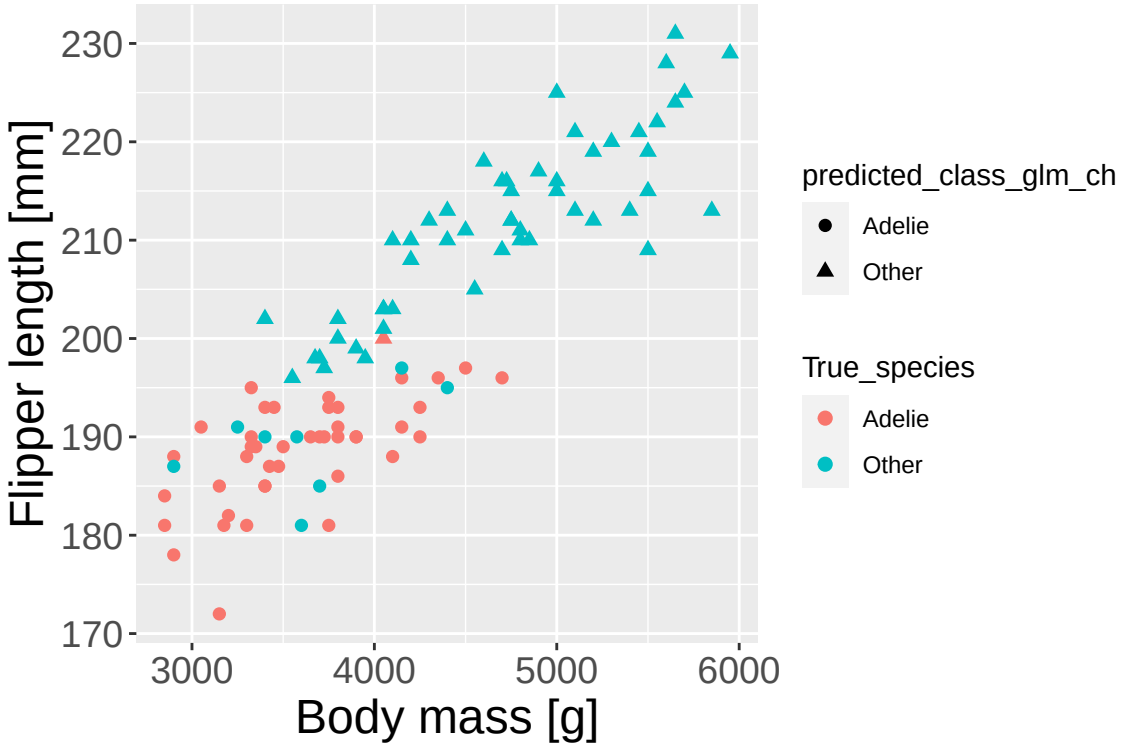
Here is code chunk:

```
# fit a logistic regression
# model-----------------------------------------------
glm_default2 = glm(adelie ~ body_mass_g + flipper_length_mm, data = train, family = "binomial")
summary(glm_default2)$coef
```

```
##                     Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)       37.7618776 5.1761640773  7.295340 2.979055e-13
## body_mass_g        0.0007120 0.0004619996  1.541127 1.232859e-01
## flipper_length_mm -0.2055804 0.0324291723 -6.339367 2.307116e-10
```

```
# calculate
# probability-----------------------------------------------------------
test$prob_glm <- 0
size = dim(test)
len = size[1]

for (i in 1:len) {
    eta <- summary(glm_default2)$coef[1, 1] + summary(glm_default2)$coef[2, 1] *
        test[i, 1] + summary(glm_default2)$coef[3, 1] * test[i, 2]
    test$prob_glm[i] <- exp(eta)/(1 + exp(eta))
}

# Perform classification using a 0.5
# cutoff-----------------------------------
cutoff <- 0.5
test$predicted_class_glm <- ifelse(test$prob_glm >= cutoff, 1, 0)
```

Here is a plot of test set in which true species are indicated by color and predicted species are indicated by shape:



2) Fit a QDA model using the training set, and perform the classification on the test set, using a 0.5 cutoff (1P).

Here is code chunk:

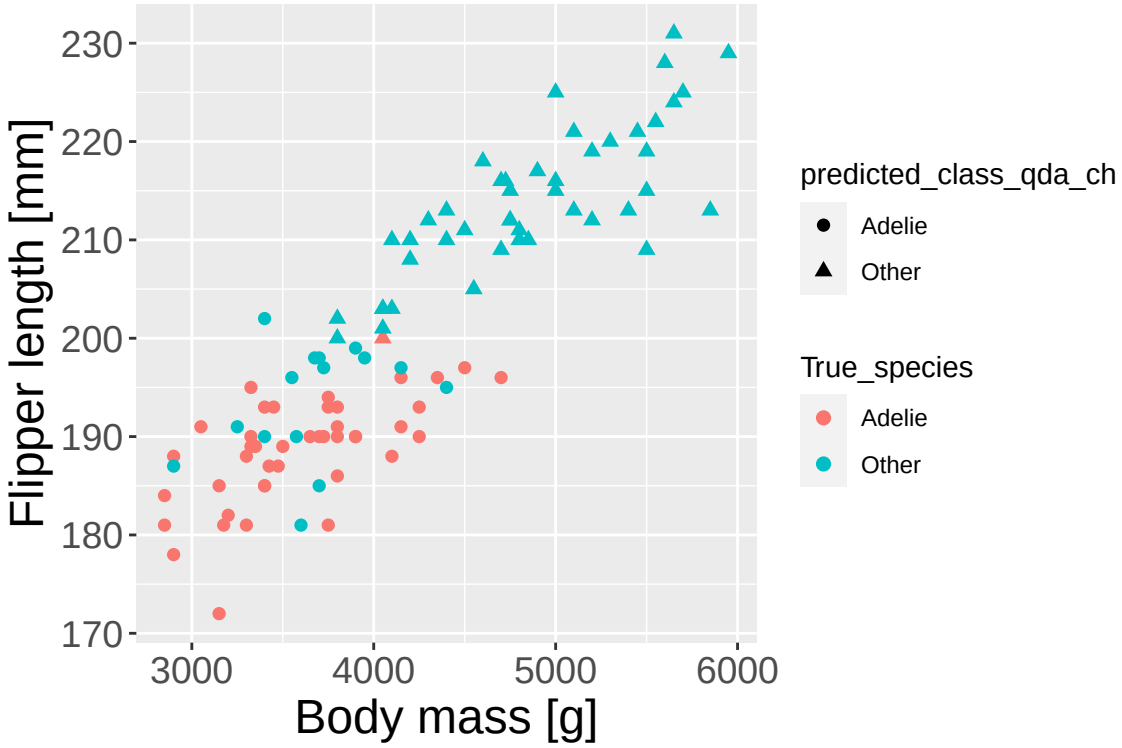
```
# fit a QDA
# model-----------------------------------------------------------------
penguins_qda = qda(True_species ~ body_mass_g + flipper_length_mm, data = train,
    prior = c(1, 1)/2)
summary(penguins_qda)
```

```
##         Length Class  Mode
## prior   2      -none- numeric
## counts  2      -none- numeric
## means   4      -none- numeric
## scaling 8      -none- numeric
## ldet    2      -none- numeric
## lev     2      -none- character
## N       1      -none- numeric
## call    4      -none- call
## terms   3      terms  call
## xlevels 0      -none- list
```

```
# calculate
# probability-----------------------------------------------------------
Posterior = predict(penguins_qda, newdata = test)$posterior
test$prob_qda <- Posterior[, 1]

# Perform the classification using a 0.5
# cutoff---------------------------------
```

```
cutoff <- 0.5
test$predicted_class_qda <- ifelse(Posterior[, 1] >= cutoff, 1, 0)
```

Here is a plot of test set in which true species are indicated by color and predicted species are indicated by shape:



3) Finally, do the same as 1) and 2) using KNN with k = 25 (1P).

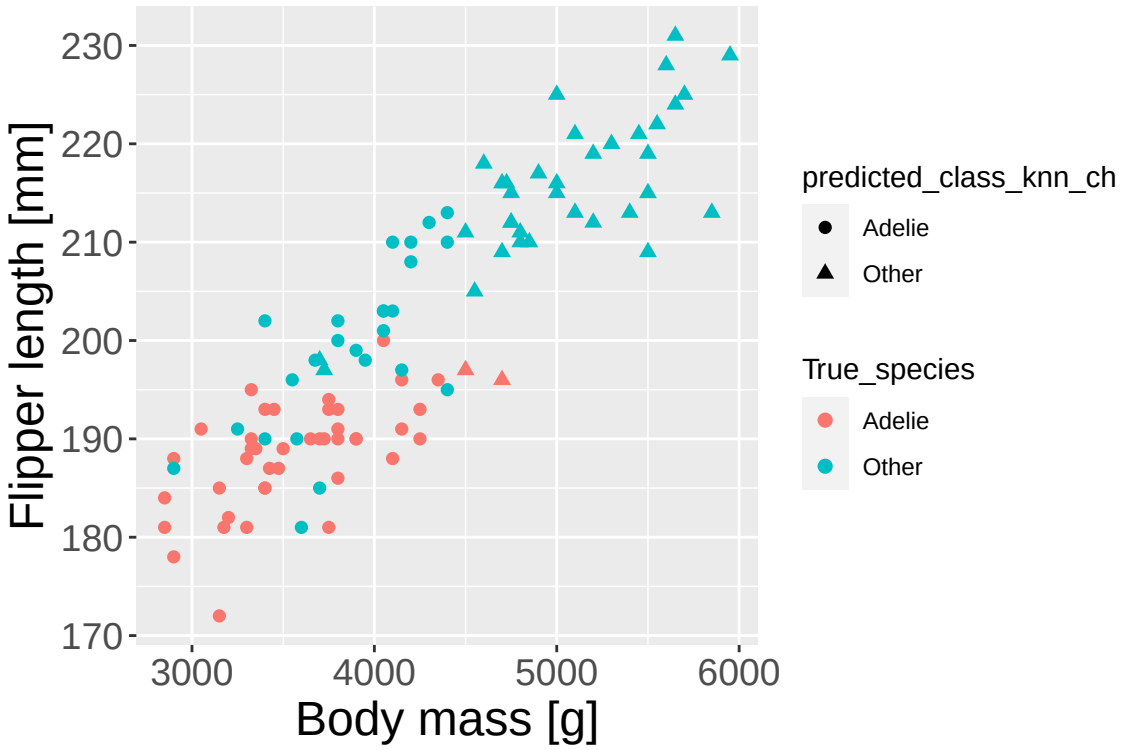
Here is code chunk:

```
# prepare data set containing only
# covariates-----------------------------------
new_train <- subset(train, select = c(body_mass_g, flipper_length_mm))
new_test <- subset(test, select = c(body_mass_g, flipper_length_mm))

# fit a KNN
# model----------------------------------------------------------------
knnMod = knn(new_train, new_test, cl = train$adelie, k = 25, prob = T)
summary(knnMod)
```

```
##  0  1
## 37 66
```

```
# precidted class
test$predicted_class_knn <- knnMod

# probability of being adelie
test$prob_knn <- ifelse(knnMod == 0, 1 - attributes(knnMod)$prob, attributes(knnMod)$prob)
```

Here is a plot of test set in which true species are indicated by color and predicted species are indicated by shape:



4) calculate the sensitivity and specificity for the three predictions performed on the test set (2P).

Here is code chunk to calculate the sensitivity and specificity:

```
table_glm <- table(predicted = test$predicted_class_glm, true = test$adelie)
table_qda <- table(predicted = test$predicted_class_qda, true = test$adelie)
table_knn <- table(predicted = test$predicted_class_knn, true = test$adelie)

Sensitivity_glm = table_glm[2, 2]/(table_glm[1, 2] + table_glm[2, 2])
Sensitivity_qda = table_qda[2, 2]/(table_qda[1, 2] + table_qda[2, 2])
Sensitivity_knn = table_knn[2, 2]/(table_knn[1, 2] + table_knn[2, 2])

Specificity_glm = table_glm[1, 1]/(table_glm[1, 1] + table_glm[2, 1])
Specificity_qda = table_qda[1, 1]/(table_qda[1, 1] + table_qda[2, 1])
Specificity_knn = table_knn[1, 1]/(table_knn[1, 1] + table_knn[2, 1])
```

Sensitivity for logistic regression:

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

Sensitivity for QDA:

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

Sensitivity for KNN:

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

Specificity for logistic regression:

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

Specificity for QDA:

```
## [1] 0.75
```

Specificity for KNN:

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

b) (5P)

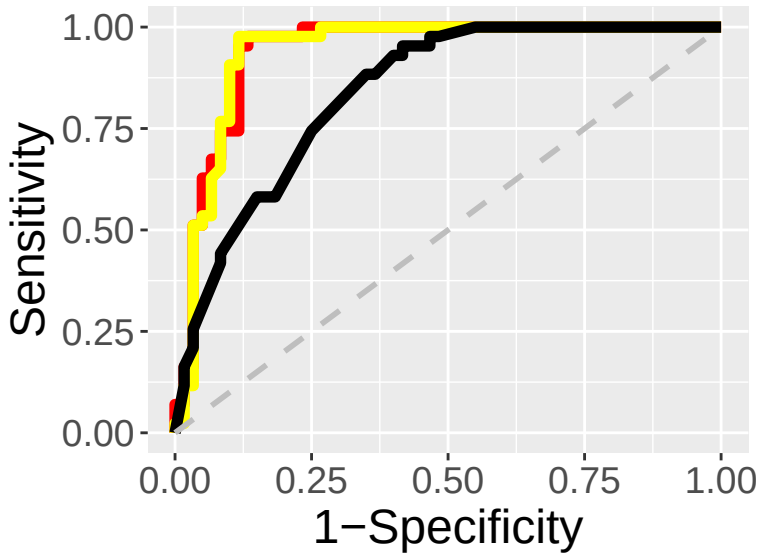
1) Present a plot of ROC curves and calculate the area under the curve (AUC) for each of the classifiers in a) (1P for each model).

Here is code chunk to calculate sensitivity and specificity for various cutoff values:

```
# calculate sensitivity and specificity for various cutoff
# values---------------
Sens_glm <- 0
Sens_qda <- 0
Sens_knn <- 0
Spec_glm <- 0
Spec_qda <- 0
Spec_knn <- 0

i <- 1
for (cutoff in 1000:0) {

    cutoff <- cutoff/1000
    test$predicted_class_glm <- ifelse(test$prob_glm >= cutoff, 1, 0)
    test$predicted_class_qda <- ifelse(test$prob_qda >= cutoff, 1, 0)
    test$predicted_class_knn <- ifelse(test$prob_knn >= cutoff, 1, 0)

    TN_glm <- ifelse(test$adelie == 0, ifelse(test$predicted_class_glm == 0, 1, 0),
        0)
    TP_glm <- ifelse(test$adelie == 1, ifelse(test$predicted_class_glm == 1, 1, 0),
        0)
    TN_qda <- ifelse(test$adelie == 0, ifelse(test$predicted_class_qda == 0, 1, 0),
        0)
    TP_qda <- ifelse(test$adelie == 1, ifelse(test$predicted_class_qda == 1, 1, 0),
        0)
    TN_knn <- ifelse(test$adelie == 0, ifelse(test$predicted_class_knn == 0, 1, 0),
        0)
    TP_knn <- ifelse(test$adelie == 1, ifelse(test$predicted_class_knn == 1, 1, 0),
        0)

    FN_glm <- ifelse(test$adelie == 1, ifelse(test$predicted_class_glm == 0, 1, 0),
        0)
    FP_glm <- ifelse(test$adelie == 0, ifelse(test$predicted_class_glm == 1, 1, 0),
        0)
    FN_qda <- ifelse(test$adelie == 1, ifelse(test$predicted_class_qda == 0, 1, 0),
        0)
    FP_qda <- ifelse(test$adelie == 0, ifelse(test$predicted_class_qda == 1, 1, 0),
        0)
    FN_knn <- ifelse(test$adelie == 1, ifelse(test$predicted_class_knn == 0, 1, 0),
        0)
    FP_knn <- ifelse(test$adelie == 0, ifelse(test$predicted_class_knn == 1, 1, 0),
        0)
```

```
    Sens_glm[i] <- sum(TP_glm)/(sum(TP_glm) + sum(FN_glm))
    Sens_qda[i] <- sum(TP_qda)/(sum(TP_qda) + sum(FN_qda))
    Sens_knn[i] <- sum(TP_knn)/(sum(TP_knn) + sum(FN_knn))

    Spec_glm[i] <- sum(TN_glm)/(sum(TN_glm) + sum(FP_glm))
    Spec_qda[i] <- sum(TN_qda)/(sum(TN_qda) + sum(FP_qda))
    Spec_knn[i] <- sum(TN_knn)/(sum(TN_knn) + sum(FP_knn))

    i <- i + 1
}


# Preapare plot
# data--------------------------------------------------------------------
plotdata <- tibble(Sens_glm = Sens_glm, Sens_qda = Sens_qda, Sens_knn = Sens_knn,
    Spec_glm = 1 - Spec_glm, Spec_qda = 1 - Spec_qda, Spec_knn = 1 - Spec_knn, )
```

Here is a plot for ROC curves (Red: Logistic regression, Yellow: QDA, Black: KNN):



Here is code chunk to calculate AUC:

```
# AUC
# calculation-------------------------------------------------------------
AUC_glm <- 0
AUC_qda <- 0
AUC_knn <- 0

for (i in 2:length(Spec_glm)) {
    if (plotdata$Spec_glm[i] > plotdata$Spec_glm[i - 1]) {
        AUC_glm <- AUC_glm + (plotdata$Spec_glm[i] - plotdata$Spec_glm[i - 1]) *
            (plotdata$Sens_glm[i] + plotdata$Sens_glm[i - 1])/2
    }
    if (plotdata$Spec_qda[i] > plotdata$Spec_qda[i - 1]) {
        AUC_qda <- AUC_qda + (plotdata$Spec_qda[i] - plotdata$Spec_qda[i - 1]) *
            (plotdata$Sens_qda[i] + plotdata$Sens_qda[i - 1])/2
    }
```

```
    if (plotdata$Spec_knn[i] > plotdata$Spec_knn[i - 1]) {
        AUC_knn <- AUC_knn + (plotdata$Spec_knn[i] - plotdata$Spec_knn[i - 1]) *
            (plotdata$Sens_knn[i] + plotdata$Sens_knn[i - 1])/2
    }
}
```

AUC for logistic regression:

```
## [1] 0.9391473
```

AUC for QDA:

```
## [1] 0.9381783
```

AUC for knn:

```
## [1] 0.8403101
```

2) Briefly discuss the ROC curves and the AUC. Which model performs best and worst (1P)?

ROC curves of Logistic regression and QDA are similar and closer to the top left corner than that of KNN. AUCs of Logistic regression and QDA are similar and larger than that of KNN. Logistic regression and QDA therefore perform similarly and better than the KNN.

3) If the task is to create an interpretable model, which model would you choose (1P)?

As mentioned above, logistic regression and QDA perform similarly and better than the KNN, however, logistic regression makes no assumptions about the covariates and preferred for interpretability.

c) (1P) Single choice

we are again looking at the logistic regression model that you fitted to the training data in a). According to this model, how would the odds that an observed animal is from the Adelie species change if the body mass increases by 1000g?

The coefficient for the body mass is 0.000712:

```
##                     Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)       37.7618776 5.1761640773  7.295340 2.979055e-13
## body_mass_g        0.0007120 0.0004619996  1.541127 1.232859e-01
## flipper_length_mm -0.2055804 0.0324291723 -6.339367 2.307116e-10
```

The odds therefore increase by exp(0.000712*1000) =

```
## [1] 2.038063
```

The correct value is 2.038.

d) (2P)

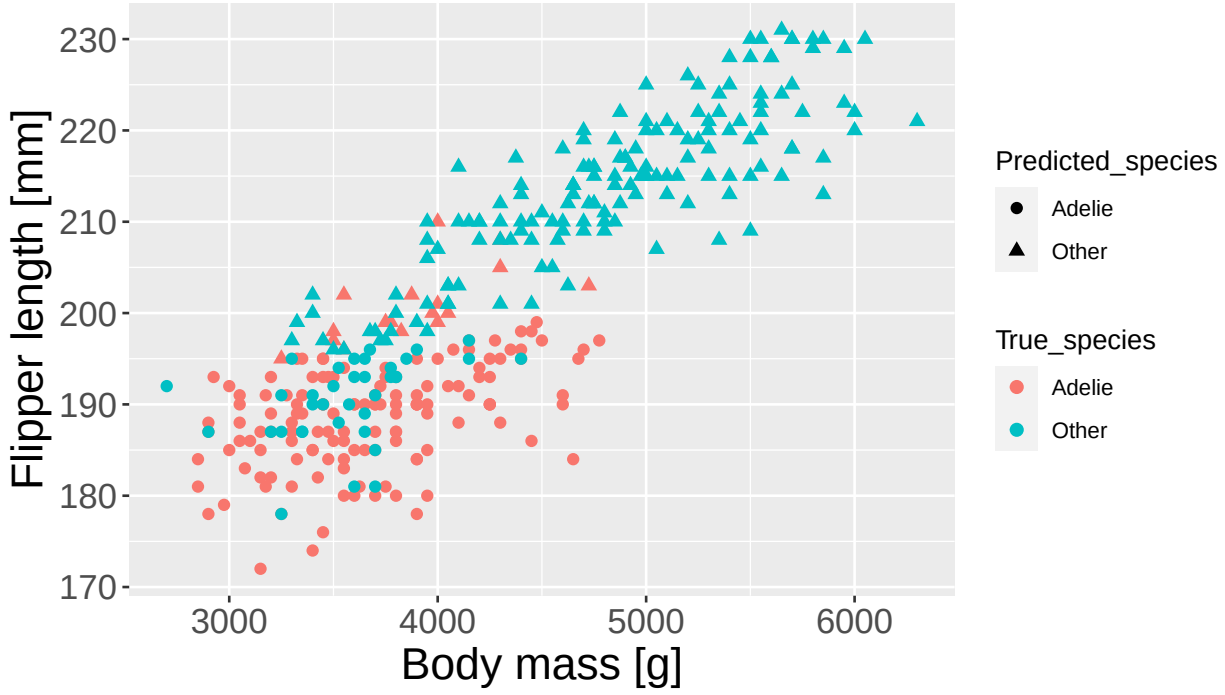
Plot the full data with the two covariates as the x- and y-axis, and use color and some other attribute of your choice (e.g. shape of highlight) to visualize the true species as well as the predicted species from the best model in b).

Here is code chunk to calculate the probability and perform the classification for full data:

```
# calculate probability for full
# data----------------------------------------
size = dim(Penguins_reduced)
len = size[1]
```

```r
for (i in 1:len) {
    eta <- summary(glm_default2)$coef[1, 1] + summary(glm_default2)$coef[2, 1] *
        Penguins_reduced[i, 1] + summary(glm_default2)$coef[3, 1] * Penguins_reduced[i,
        2]
    Penguins_reduced$prob_glm[i] <- exp(eta)/(1 + exp(eta))
}

# perform the
# classification-----------------------------------------------------
cutoff <- 0.5
Penguins_reduced$predicted_class_glm <- ifelse(Penguins_reduced$prob_glm >= cutoff,
    1, 0)
Penguins_reduced$Predicted_species <- ifelse(Penguins_reduced$predicted_class_glm ==
    1, "Adelie", "Other")
```

Here is a plot:



Problem 4

a)

1. The validation set-approach is computationally cheaper than 10-fold CV: True
2. 5-fold CV will generally lead to less bias, but more variance than LOOCV in the estimated prediction error: True
3. The validation set-approach is the same as 2-fold CV: True
4. LOOCV is always the cheapest way to do cross-validation: True

b)

We load the dataset and use logistic regression as our model.

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

summary(d.chd)  # Get an overview of the dataset
```

```
##       chd              sex              sbp            smoking
##  Min.   :0.000   Min.   :0.0   Min.   :111.3   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.0   1st Qu.:125.0   1st Qu.:0.000
##  Median :0.000   Median :0.5   Median :129.8   Median :0.000
##  Mean   :0.122   Mean   :0.5   Mean   :130.2   Mean   :0.222
##  3rd Qu.:0.000   3rd Qu.:1.0   3rd Qu.:135.7   3rd Qu.:0.000
##  Max.   :1.000   Max.   :1.0   Max.   :157.1   Max.   :1.000
```

```
method = "binomial"  # logistic regression in GLM
test_sbp = 150  # this sbp is larger than the mean (130.2), and close to the max (157.1)
test_sex = 1  # male
test_smoking = 0  # non-smoking

# Logistic regression
model_4b <- glm(chd ~ sbp + sex + smoking, data = d.chd, family = method)


# Make a test subject
test_subject = data.frame(sbp = test_sbp, sex = test_sex, smoking = test_smoking)

# Make prediction on test subject
probability = predict(model_4b, test_subject, type = "response")

print(c("The probability that a male of...", probability))
```

```
##                                                                      1
## "The probability that a male of..."                 "0.100959971824563"
```

```
sprintf("The probability of a non-smoking male with %i having chd is %#.3f", test_sbp,
    probability)
```

```
## [1] "The probability of a non-smoking male with 150 having chd is 0.101"
```

```
summary(model_4b)  # to compare with task 4d
```

```
##
## Call:
## glm(formula = chd ~ sbp + sex + smoking, family = method, data = d.chd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0184  -0.5950  -0.3790  -0.2954   2.5570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.65884    2.36740  -2.813  0.00491 **
## sbp          0.03877    0.01794   2.162  0.03066 *
```

```
## sex         -1.34351    0.32148  -4.179 2.93e-05 ***
## smoking      0.41031    0.31014   1.323  0.18584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.89  on 499  degrees of freedom
## Residual deviance: 342.91  on 496  degrees of freedom
## AIC: 350.91
##
## Number of Fisher Scoring iterations: 5
```

We see that the mean sbp is 130.2 and the max is 157.1. We want to predict if a non-smoking male with sbp = 150 has chd. We find from the given dataset that a non-smoking male with sbp = 150 will only have a 10% of having chd.

c)

```
set.seed(1)

observations = dim(d.chd)[1]

b_iter = 1000  # number of bootstrap iterations

estimation = c()

for (i in 1:b_iter) {
    model_4c <- glm(chd ~ sbp + sex + smoking, data = d.chd, family = method, subset = sample(observati
        size = observations, replace = TRUE))

    estimated = predict(model_4c, test_subject, type = "response")

    # store the estimated probability on test subject
    estimation[i] = estimated
}
```

c) Standard error

To calculate the standard error, we use the formula (from ch. 5 in the book ISLR)

$$\mathrm{SE}_B = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\text{estimation}_i - \frac{1}{B}\sum_{j=1}^{B}\text{estimation}_j\right)^2} \tag{10}$$

```
mean = mean(estimation)

SE_B = sqrt(1/(b_iter - 1) * sum((estimation - mean)^2))


sprintf("The standard error calculated is %#.3f", SE_B)
```

```
## [1] "The standard error calculated is 0.044"
```

c) Confidence interval and mean

For a confidencen interval of 95%:

```
z = 1.96
CI_interval = c()
CI_interval[1] = mean - z * SE_B
CI_interval[2] = mean + z * SE_B

sprintf("The confidence interval is [%#.3f,%#.3f] with mean %#.3f", CI_interval[1],
    CI_interval[2], mean)
```

```
## [1] "The confidence interval is [0.021,0.194] with mean 0.108"
```

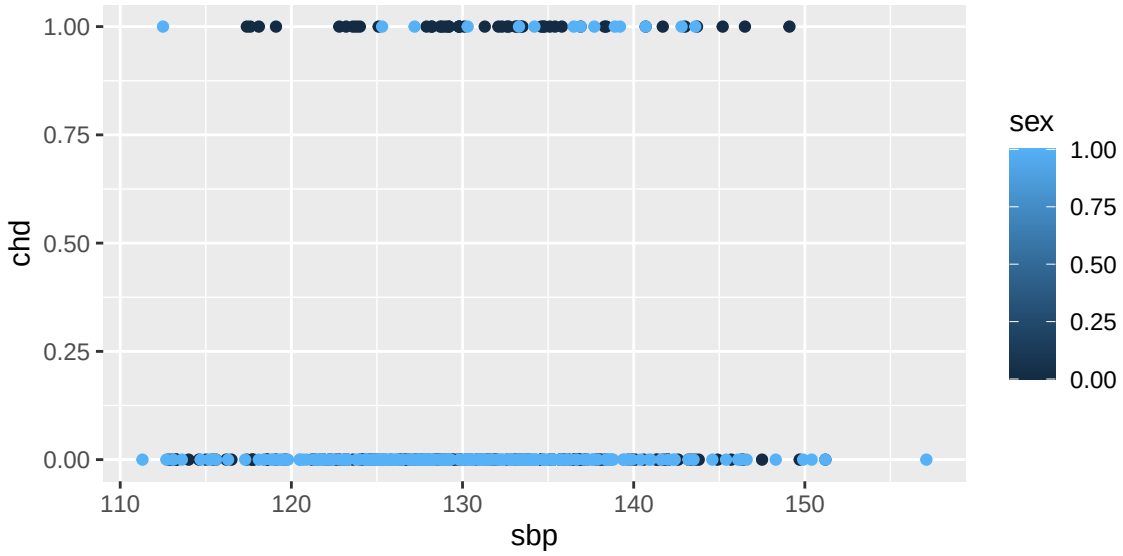
c) Comment on results

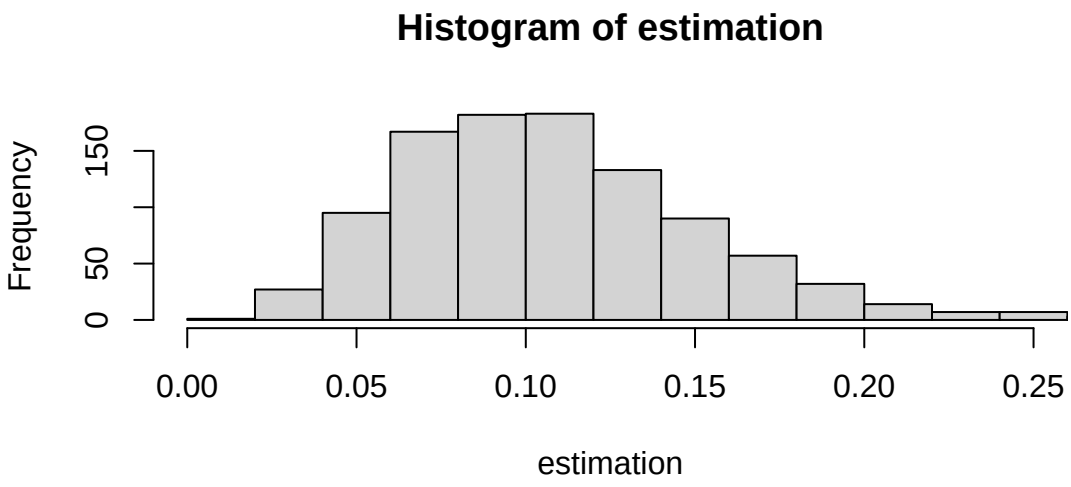
The confidence interval of 95% (CI95) tells us that the true mean, for 95% of the population of non-smoking males with sbp=150, will lie in the interval $[0.021, 0.194]$.

We observe that we get quite a spread in our CI95, and the values seems a bit low for such a high sbp. To understand why, we should analyze the data. By looking at the data, we find that there are 47 cases of females with chd, while only 14 cases of male with chd. By plotting the sbp vs. chd, separated by sex, we also observe that the values are quite spread for both genders. These reasons may explain why we get such a low probability of male having chd with such a high sbp. If we run the code for a non-smoking female with sbp=150, we get a mean probability of 0.311, and the CI95 interval $[0.151, 0.471]$, which seems more plausible. To get a more accurate model, one would probably have to include more variables, such as weight, age and eating and exercise habits.

```
## [1] "Males with chd in dataset is 14"
```

```
## [1] "Females with chd in the dataset is 47"
```





d)

1. The bootstrap relies on random sampling the same data without replacement: False
2. The estimated standard errors from the glm() function are smaller than those estimated from the bootstrap, which indicates a problem with the bootstrap: False
3. In general, differences between the estimated standard errors from the bootstrap and those from glm() may indicate a problem with the assumptions taken in logistic regression.: True
4. The p-values from the glm() output are probably slightly too small: True