Compulsory exercise 1: Group 5

TMA4268 Statistical Learning V2022

Aleksander Johnsen Solberg and Gjermund Oscar Lyckander

22 February, 2022

Problem 1

a)

$$\begin{aligned}
E[y_0 - \hat{f}(x_0)]^2 &= E[(f(x_0) + \varepsilon - \hat{f}(x_0))^2] \\
&= E[(f(x_0))^2] + E[\varepsilon^2] + E[\hat{f}(x_0)^2] - 2E[f(x_0)\hat{f}(x_0)] + 2E[f(x_0)\varepsilon] + 2E[\hat{f}(x_0)\varepsilon] \\
&= f(x_0)^2 + \mathrm{Var}(\varepsilon) + \mathrm{Var}(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 - 2E[f(x_0)\hat{f}(x_0)] \\
&= (f(x_0) - E[\hat{f}(x_0)])^2 + \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Var}(\varepsilon)
\end{aligned}$$

The three terms in the last line are the squared bias, variance, and irreducible error respectively.

b)

The three terms can be interpreted as the following. The bias term is the error that comes from modeling a complicated real-life problem with a simple model. The more flexible the model is, the smaller the bias will be. The variance term is how much the estimate $\hat{f}$ would change if we were using different training data. The more flexible the model is, the larger the variance will be. Lastly, the irreducible error term is simply the error that comes from the error in the data itself.

c)

(i) True
(ii) False
(iii) True
(iv) False

d)

(i) True
(ii) False
(iii) False
(iv) False

e)

(i) is true

Problem 2

Here is a code chunk:

```r
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>
```

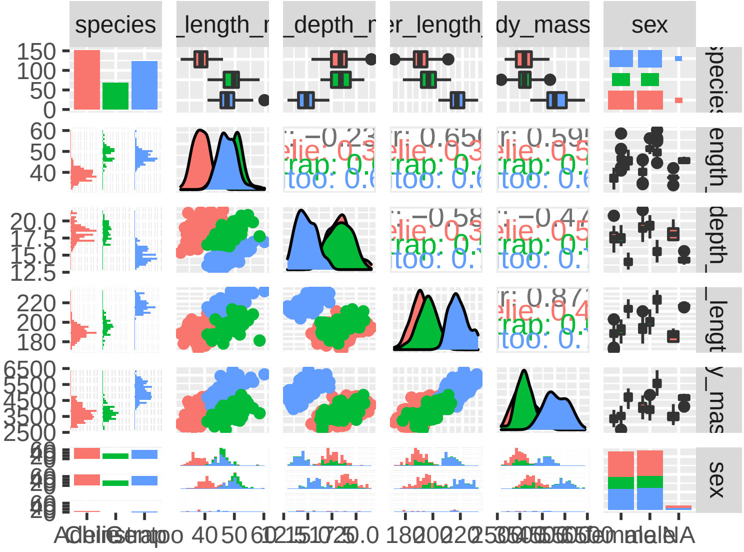
```r
Penguins <- subset(penguins, select = -c(island, year))
```

a)

- Takes the covariate ‘sex’ out of the model despite it being very segnificant. Basil clearly has clearly misunderstood what is considered a good p-value
- Leaves in covariates that are clearly not significant, such as the interaction between ‘bill_depth_mm’ and ‘species’.
- Says that the interaction term is overall significant when only this is only true for one species.
- Does not include ‘bill_length_mm’ in the model at any point, even though we suspect it might be siginificant
- Concludes that chinstrap penguins have the largest bodymass, which one can clearly see from the data is not true. There must be something wrong with the model.

b)

```r
library(GGally)
ggpairs(Penguins, aes(colour = species))
```



c)

```
penguin.model1 <- lm(body_mass_g ~ . + species*bill_depth_mm, data = Penguins)
summary(penguin.model1)
```

```
##
## Call:
## lm(formula = body_mass_g ~ . + 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 **
## speciesChinstrap                1539.690    674.106   2.284 0.023015 *
## speciesGentoo                    699.379    537.435   1.301 0.194071
## bill_length_mm                    19.752      7.124   2.773 0.005880 **
## bill_depth_mm                     80.340     22.119   3.632 0.000327 ***
## flipper_length_mm                 15.936      2.928   5.444 1.03e-07 ***
## sexmale                          385.683     47.350   8.145 8.28e-15 ***
## 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
```

```
anova(penguin.model1)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##                        Df    Sum Sq  Mean Sq  F value    Pr(>F)
## species                 2 145190219 72595110 900.8882 < 2.2e-16 ***
## bill_length_mm          1  23755815 23755815 294.8041 < 2.2e-16 ***
## bill_depth_mm           1   9791958  9791958 121.5159 < 2.2e-16 ***
## flipper_length_mm       1   4124003  4124003  51.1779 5.659e-12 ***
## sex                     1   5482024  5482024  68.0306 4.083e-15 ***
## species:bill_depth_mm   2    807174   403587   5.0084  0.007208 **
## Residuals             324  26108473    80582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not quite happy with this, as the species coavriate is not as significant as one would think from seeing the pairs plot. Try without ‘bill_length_mm’ and the interactions.

```
penguin.model2 <- lm(body_mass_g ~  bill_depth_mm + flipper_length_mm + sex + species, data = Penguins)
summary(penguin.model2)
```

```
##
## Call:
```

```
## lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_mm +
##     sex + species, data = Penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -788.93 -189.77  -18.89  196.55  914.16
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1211.534    567.716  -2.134 0.033582 *
## bill_depth_mm        74.383     19.708   3.774 0.000191 ***
## flipper_length_mm    17.544      2.866   6.121 2.66e-09 ***
## sexmale             435.433     44.800   9.720  < 2e-16 ***
## speciesChinstrap    -78.899     45.498  -1.734 0.083838 .
## speciesGentoo      1153.986    118.582   9.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 289.8 on 327 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.8705
## F-statistic: 447.3 on 5 and 327 DF,  p-value: < 2.2e-16
```

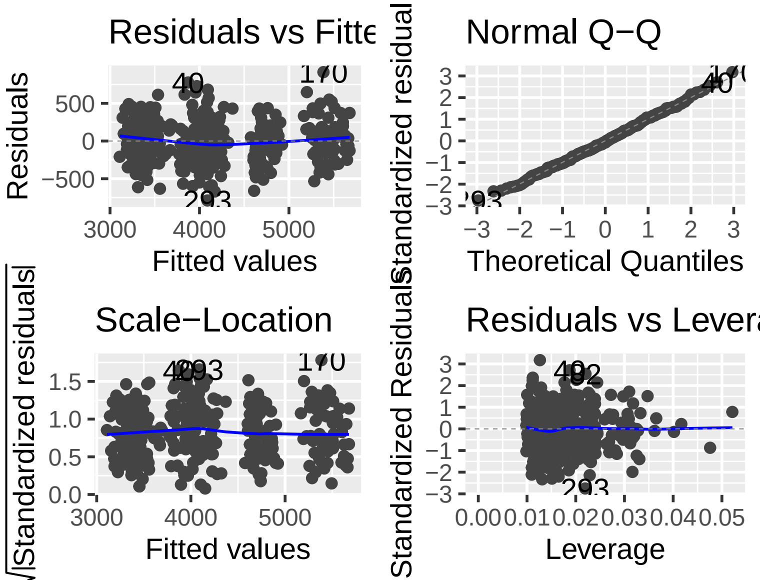
```r
anova(penguin.model2)
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##                    Df    Sum Sq   Mean Sq  F value    Pr(>F)
## bill_depth_mm       1  47959592  47959592  571.166 < 2.2e-16 ***
## flipper_length_mm   1 116426999 116426999 1386.567 < 2.2e-16 ***
## sex                 1  12745079  12745079  151.785 < 2.2e-16 ***
## species             2  10670525   5335262   63.539 < 2.2e-16 ***
## Residuals         327  27457472     83968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We begin with a model with bodey mass as the response, and species, bill length, bill depth, flipper length, sex and the interaction between species and bill depth as the covariates. We see that not all of these predictors are significant in the model, and so we try a reduced model using only bill length, flipper length, sex and species as the covariates. We now see that all the covariates have p-values that are very low, and so they should be included in the model. In the species covariate, we see that only the distinction between Gentoo and the other two species is helpful, and so we will reflect this in the model by only distinguishing between Gentoo and not Gentoo. The final model can thus be described as such:

$$
\begin{aligned}
\hat{y}_{female} &= \hat{\beta}_0 + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{flipper\_length} x_{flipper\_length} \\
\hat{y}_{male} &= \hat{\beta}_0 + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{male} \\
\hat{y}_{female\_gentoo} &= \hat{\beta}_0 + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{gentoo} \\
\hat{y}_{male\_gentoo} &= \hat{\beta}_0 + \hat{\beta}_{bill\_depth} x_{bill\_depth} + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{male} + \hat{\beta}_{gentoo}
\end{aligned}
$$

```r
library(ggfortify)
autoplot(penguin.model2)
```



From the residuals vs. fitted plot, we do not see any evidence of non-linearity. and can therefore conclude that the expected value of the residuals is zero. We do however see some structure in that the points are grouped together into four groups. This might come from the fact that we have a model that is split into four given by the sex and the species, however, the significance of this structure is unknown.

In the QQ-plot, we can see the points follow the straight line very well, and we can say that the residuals are normally distributed.

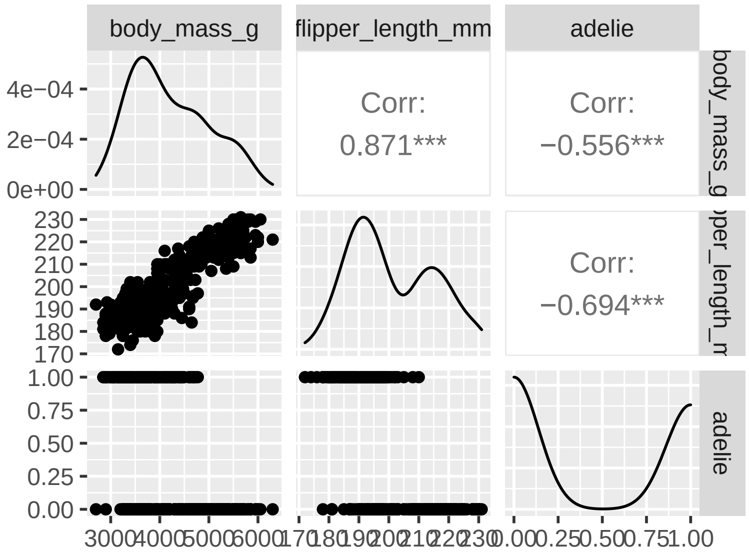
Problem 3

```r
library(tidyverse)
library(GGally)
# 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(0.70 * nrow(Penguins_reduced))
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
ggpairs(Penguins_reduced)
```



```r
log.model <- glm(adelie ~ ., data = train, family = binomial)
summary(log.model)
```

```
##
## Call:
## glm(formula = adelie ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6506  -0.4133  -0.1161   0.6550   2.2962
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       37.761878   5.176164   7.295 2.98e-13 ***
## body_mass_g        0.000712   0.000462   1.541    0.123
## flipper_length_mm -0.205580   0.032429  -6.339 2.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.11  on 238  degrees of freedom
## Residual deviance: 184.21  on 236  degrees of freedom
## AIC: 190.21
##
## Number of Fisher Scoring iterations: 6
```

```r
log.probabilities <- predict(log.model, newdata = test, type = 'response')
log.predicted.classes <- ifelse(log.probabilities > 0.5, 1, 0)
mean(log.predicted.classes == test$adelie)
```

```
## [1] 0.9126214
```

(ii)

```r
library(MASS)
qda.model <- qda(adelie ~ ., data = train)
summary(log.model)
```

```
##
## Call:
## glm(formula = adelie ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6506  -0.4133  -0.1161   0.6550   2.2962
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       37.761878   5.176164   7.295 2.98e-13 ***
## body_mass_g        0.000712   0.000462   1.541    0.123
## flipper_length_mm -0.205580   0.032429  -6.339 2.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.11  on 238  degrees of freedom
## Residual deviance: 184.21  on 236  degrees of freedom
## AIC: 190.21
##
## Number of Fisher Scoring iterations: 6
```

```
qda.probabilities <- predict(qda.model, newdata = test, type = 'response')$posterior
qda.predicted.classes <- predict(qda.model, newdata = test, type = 'response')$class
mean(qda.predicted.classes == test$adelie)
```

```
## [1] 0.8543689
```

(iii)

```
library(class)
knn.model <- knn(train = train, test = test, cl = train$adelie, k=25, prob = T)
table(knn.model, test$adelie)
```

```
##
## knn.model  0  1
##         0 35  2
##         1 25 41
```

```
mean(knn.model == test$adelie)
```

```
## [1] 0.7378641
```

(iv)

```
library(caret)
sensitivity(table(log.predicted.classes, test$adelie))
```

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

```
specificity(table(log.predicted.classes, test$adelie))
```

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

```
sensitivity(table(qda.predicted.classes, test$adelie))
```

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

```
specificity(table(qda.predicted.classes, test$adelie))
```

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

```
sensitivity(table(knn.model, test$adelie))
```

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

```
specificity(table(knn.model, test$adelie))
```

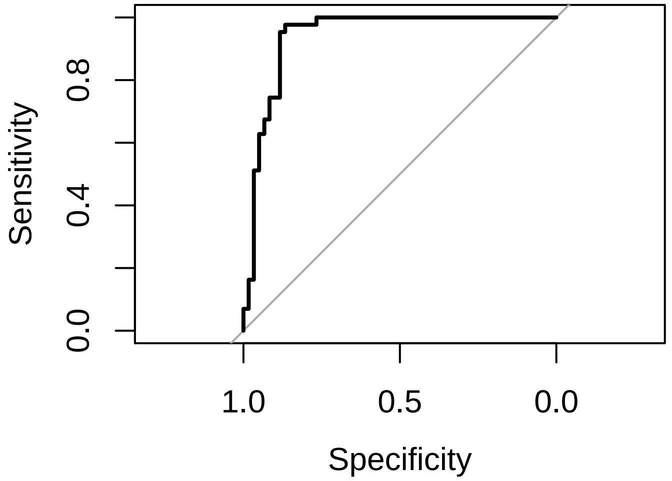
```
## [1] 0.9534884
```

The logistic regression model has a sensitivity of 87%, and a specificity of 98%. The QDA model has a sensitivity of 77%, and a specificity of 98%. The KNN model has a sensitivity of 58%, and a specificity of 95%.

b)

(i)

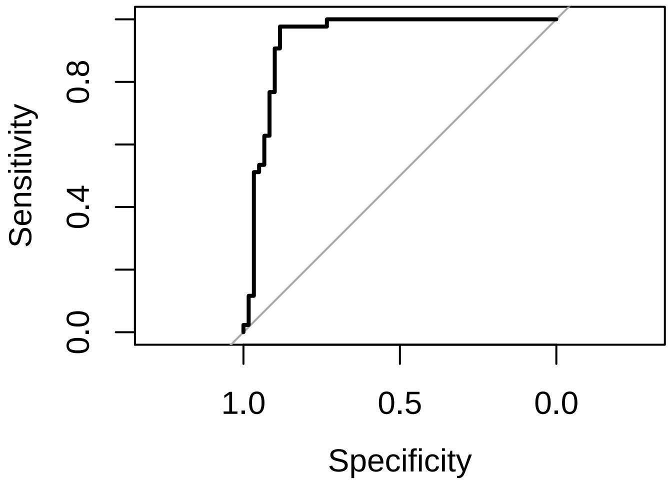
```
library(pROC)
log.roc <- roc(test$adelie, log.probabilities, direction = '<', lwd=3)
plot(log.roc)
```



```
auc(log.roc)
```

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

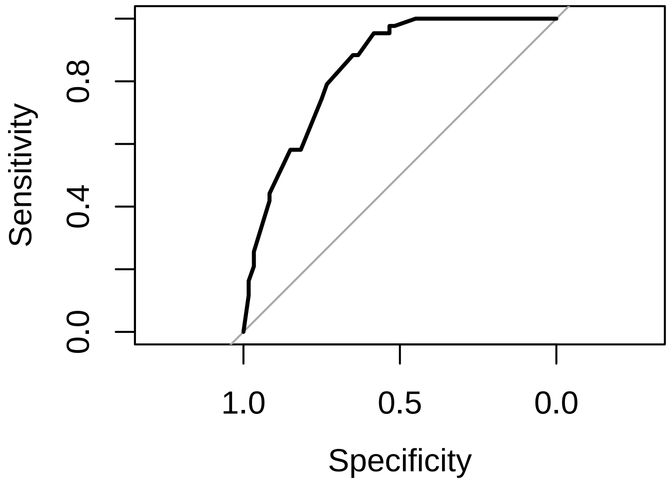
```
qda.roc = roc(test$adelie, qda.probabilities[,2], direction = '<', lwd=3)
plot(qda.roc)
```



```
auc(qda.roc)
```

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

```
probKNN = ifelse(knn.model == 0, 1 - attributes(knn.model)$prob, attributes(knn.model)$prob)
knn.roc <- roc(test$adelie, probKNN, direction = '<', lwd=3)
plot(knn.roc)
```



```
auc(knn.roc)
```

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

(ii) We see that the logistic regression model and the QDA model both perform very well in this instance, while the KNN model does a decent job, but far worse than the other two.

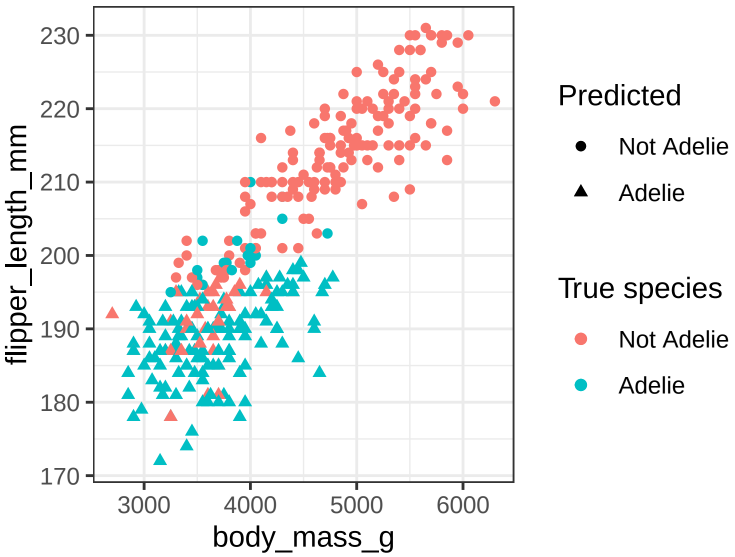
(iii) In order to get an interpretable model, I would choose the logistic regression model, as you can easily interpret the effect each covariate has on the prediction from the coefficients the model gives.

c)

(iii) is True

d)

```r
library(ggplot2)
log.train.probabilities <- predict(log.model, newdata = train, type = 'response')
log.train.predicted.classes <- ifelse(log.train.probabilities > 0.5, 1, 0)
ggplot(test, aes(x = body_mass_g, y = flipper_length_mm, group=adelie)) +
      geom_point(aes(colour = factor(adelie, labels = c('Not Adelie', 'Adelie')), shape = factor(log.pr
      geom_point(data=train, aes(colour = factor(adelie, labels = c('Not Adelie', 'Adelie')), shape = fa
      labs(color = 'True species', shape = 'Predicted') +
      theme_bw()
```



Problem 4

a)

(i) True
(ii) False
(iii) False
(iv) False

b)

```r
id <- "1chRpybM5cJn4Eow3-_xwDKPKyddL9M2N" # google file ID
d.chd <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
d.chd$sex <- as.factor(d.chd$sex)
d.chd$smoking <- as.factor(d.chd$smoking)
glm.fit <- glm(chd ~ sbp + sex + smoking, data = d.chd, family = "binomial")
summary(glm.fit)$coef
```

```
##                Estimate Std. Error  z value     Pr(>|z|)
## (Intercept) -6.65883685 2.36740155 -2.812720 4.912446e-03
## sbp          0.03877165 0.01793731  2.161508 3.065610e-02
## sex1        -1.34351384 0.32148322 -4.179110 2.926516e-05
## smoking1     0.41031080 0.31014166  1.322979 1.858425e-01
```

```
glm.predict <- predict(glm.fit, data.frame(sbp = 150, sex = as.factor(1), smoking = as.factor(0)), type
glm.predict
```

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

The probability of chd for a non-smoking male with a sbp=150 is 0.101.

c)

```
B <- 1000
n <- 101
estimate <- rep(NA, B)
for (b in 1:B){
  set.seed(b)
  thisboot <- d.chd[sample(nrow(d.chd), n), ]
  boot.fit <- glm(chd ~ sbp + sex + smoking, data = thisboot, family = "binomial")
  boot.predict <- predict(boot.fit, data.frame(sbp = 150, sex = as.factor(1), smoking = as.factor(0)),
  estimate [b] <- boot.predict
}

std.err <- function(x) sd(x)/sqrt(length(x))
estimate.stderr <- std.err(estimate)

estimate.mean <- mean(estimate)

alpha <- 0.05
degrees.freedom <- length(estimate)-3
t.score = qt(p=alpha/2, df = degrees.freedom, lower.tail = F)

margin.error <- t.score*estimate.stderr
lower.bound <- estimate.mean - margin.error
upper.bound <- estimate.mean + margin.error


estimate.stderr
```

```
## [1] 0.003224361
```

```
estimate.mean
```

```
## [1] 0.1202212
```

```
lower.bound
```

```
## [1] 0.1138939
```

```
upper.bound
```

```
## [1] 0.1265485
```

The standard error is 0.0032244.

The 95% quantile interval is (0.1138939, 0.1265485).

From the results of the bootstrap we see that a non-smoking male with a sbp=150 has an expected probability of coronary heart disease of 0.1202212. There is a probability of 95% that the expected probability of chd for a non-smoking male with a sbp=150 lies between 0.1138939 and 0.1265485.

d)

(i) False

(ii) True

(iii)

(iv)